A Model to Predict Driver Task Performance When Interacting with In-Vehicle Speech Interfaces for Destination Entry and Music Selection

by

Ei-Wen Lo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environmental Health Sciences)
in the University of Michigan
2013

Doctoral Committee:

Professor Alfred Franzblau, Co-Chair
Research Professor Paul A. Green, Co-Chair
Associate Professor Steven P. Abney
Professor Thomas J. Armstrong
Professor Stanley Berent
Professor Andrew D. Maynard

For my grandfather…

Acknowledgments

I would like to thank my advisors, Dr. Alfred Franzblau and Dr. Paul Green, for all of their guidance and encouragements during all phases of my dissertation. I would also like to thank the other members of my dissertation committee, Dr. Thomas Armstrong, Dr. Steven Abney, Dr. Stanley Berent, and Dr. Andrew Maynard, for their advice and input to this work.

I would like to thank the entire support staff within the University of Michigan Transportation Research Driver Interface Group for their supports and contributions. In addition, Dr. Brian Tsang-Wei Lin has done much to make this work possible.

I would like to thank my many friends who have encouraged and supported me over the years. Their supports have always helped me overcome the many stumbling blocks along the way. I am fortunate to have met so many wonderful people during my years in Michigan and I am grateful for our friendships.

My parents have always expected the best from me and provided financial supports unconditionally through these years. Most importantly, I thank my wife, Pei-yin Wen, for all she has done to keep me on my feet and moving forward during this long process. I thank her for keeping cheery attitude even during the nights when our life seemed overwhelming because of this dissertation. Through me, she has experienced all the hardships and struggles that accompany an undertaking like this. None of this could have been possible without her love and encouragement. Thanks, my dear.

Table of Contents

List of Figures

List of Tables

Abstract

Motor vehicle crashes were estimated to be the eleventh leading cause of death in United States in 2009. The percentage of fatal crashes in which driver distraction was a causal factor increased from 10% in 2005 to 16% in 2009, and this was particularly likely for systems with visual-manual driver interfaces, such as infotainment systems. Using a speech interface to operate infotainment systems while driving can potentially reduce driver distraction. Unfortunately, evaluations of driver interfaces are typically conducted after the hardware and software are developed, which is often too late to make changes. An alternative approach is to model driver task performance when using speech interfaces and to use the model to predict system performance early in design when changes are easier to make. The purposes of this research are to understand how drivers interact with current in-vehicle speech interfaces and based on that knowledge, develop and validate a simulation model of how drivers interact with speech interfaces to aid speech-interface development. Specifically, this model will predict user task performance (task completion times and errors) when drivers interact with in-vehicle speech-controlled interfaces to complete destination entry and music selection tasks.

To develop the simulation model, a preliminary survey and a driving simulator experiment were conducted to identify how these tasks are carried out and the values for the process parameters. First, using a survey, frequency data for tasks and methods, (e.g., how often destinations are selected using street addresses vs. point of interest), and the

content in user-generated databases (e.g., prior destination lists, play lists, etc.) were collected to assure that real tasks and constraints are considered in the simulation model. Next, a driving simulator experiment involving 48 subjects interacting with an existing in-vehicle speech interface was conducted to understand how drivers perform destination entry and music selection and to determine the time drivers need to construct utterances (and their distributions), the types of errors drivers make (and their probability), and the probability of various correction strategies are used for each type of error. Half of these data were used to create the simulation model structure and provide the model parameters for entering destinations and selecting music using speech. Finally, the simulation model was validated for these two tasks using the second half of the data from the previous experiment.

This research provides a model (both structure and content) for use with existing simulation software packages to predict user task performance with speech interfaces in motor vehicles. Use of this model supports the design of safer and easier to use speech interfaces in vehicles that can minimize eyes-off-road time and should reduce crash risk, and thereby protect public health. This model can be exercised to examine alternative speech interface configurations months before a physical interfaces is available for user testing when changes are easier to make, which saves time, reduces cost, and improves the quality of the interface produced.

CHAPTER 1

Introduction

1.1 Research Background

In recent years, MP3 players, navigation systems, and other systems have been introduced into motor vehicles. The number and types of these systems, as well as their use, is expected to grow considerably. Most of those systems have visual-manual interfaces that require drivers to look away from the road to operate them, which could lead to distraction-related crashes. One potential way to reduce distraction and the associated crash risk is to use speech interfaces instead of visual-manual interfaces. The major weakness of the state of art is that the usability of speech interfaces is assessed experimentally by having real users try them, and collecting task time and error data. Often that process occurs late in design, when the design of the interfaces is complete, so such tests may have little impact on what is produced. An alternative developmental approach to testing is to model driver performance, in the case of using a speech interface, and to use the model to predict system performance early in design. Accordingly, the major goal of this research is to develop a simulation model of driver use of two in-vehicle applications – destination entry and music selection.

1.2 Rationale

Motor vehicle crashes were estimated to be the eleventh leading cause of death in the United States in 2009 [1]. According to the Fatality Analysis Reporting System, there were 30,797 motor vehicle crashes resulting in 33,808 deaths in 2009 [2]. Fatal crashes in which driver distraction was a major cause increased from 10% in 2005 to 16% in 2009 [2]. Distracting tasks such as *adjusting radio, cassette, CD*, *using/dialing cellular phone*, and *adjusting vehicle/climate controls* accounted for 19% of distraction-related crashes [3]. Over time, frequent distracting tasks have changed from CD and cassette use to MP3 player use and various tasks associated with navigation systems, as well as other recently introduced systems. The use of all of these newer systems is expected to grow considerably. To operate systems with visual-manual interfaces, drivers must look away from the road, which could lead to distraction-related crashes. Research has shown that the chance of crashes increases when the eyes-off-road duration is greater than 2 seconds [4]. To reduce driver distraction, the National Highway Traffic Safety Administration (NHTSA) of the U.S. Department of Transportation (DOT) posted proposed visual-manual driver-distraction guidelines for in-vehicle electronic devices for public comment on February 15, 2012 to encourage manufactures to develop "less distracting" in-vehicle electronic devices (http://www.nhtsa.gov/About+NHTSA/Press+Releases/2012/U.S.+Department+of+Transportation+Proposes+'Distraction'+Guidelines+for+Automakers, retrieved May 15, 2012). Revised guidelines were released on April 23, 2013 (http://www.nhtsa.gov/staticfiles/nti/distracted_driving/pdf/distracted_guidelines-FR_04232013.pdf. Retrieved April 23, 2013).

Using a speech interface instead of a visual-manual interface for in-vehicle tasks is one potential way to reduce crash risk. Research has shown that using speech interfaces resulted in better driving performance, fewer and briefer off-road glances, and less workload than the visual-manual interfaces [5-13]. Although using a speech interface can reduce eyes-off-road time, speech interfaces still impose a cognitive demand, which can also interfere with the primary driving task. Lee et al showed that drivers' reaction time increased by 180 ms while a complex speech-controlled e-mail system was used when compared with a simple system[14]. They also reported that subjective workload ratings and probe questions also introduced a significant cognitive load, which was highest for the complex e-mail system. To assess usability of speech interfaces experimentally involves having real users try them, and collecting task time and error data, evaluations that often occur late in design [5-13, 15, 16]. In fact, such evaluations often occur when the design of the interfaces is complete, so such tests may have little impact on what is produced, a major weakness of the state of art.

An alternative approach is to model driver task performance while using a speech interface, and to use the model to predict system performance early in design. Because task times and probabilities are not strictly deterministic, but vary from instance to instance and driver to driver, it is appropriate to model system performance using Monte-Carlo simulation methods.

A Monte-Carlo simulation can be represented as a flow diagram or a task network, often with tasks looping back (as when a system asks a person to repeat what they said, over and over). Tasks have duration distributions (normal, exponential, etc.), and there are rules or probabilities to determine which task is next. A flow chart for a hypothetical simulation model is shown in Figure 1-1. Beginning at the first task in the model, values are randomly selected from the time distribution and following the rules for that task which determine the time so far and the next task. The process is repeated until an ending task is reached, which provides the total duration and path for a single pass. This, in turn, is repeated to develop a time distribution for the task network. Because speech interfaces often have task sequences that involve looping back, deterministic computations of total task time are very difficult, if not impossible to compute. For example, a user could say something that the system may or may not understand. In some instances, the user may need to repeat the utterance several times before what they say is understood, and they may change what they say and how they say it to aid recognition. (Machine: “Please say a state.” User: “Michigan.” M: “Do not understand. Please say a state.” U: “Michigan.” M: “Do not understand. Please say a state.” U: “Mee-she-gan.” M: “Please say a city.” …)

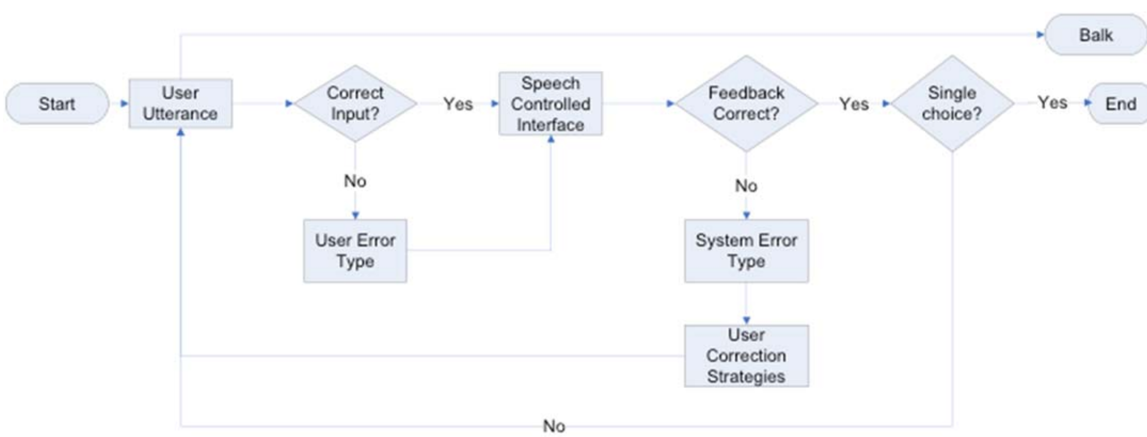


Figure 1-1. A Hypothetic Flow Chart for Driver and Speech Interface Interaction

1.3 Research Objectives

The goal of this research is to understand the interaction of drivers and an existing speech interface and to provide a model structure (the tasks users perform and their sequence for various error contingencies) and data (either distributions of task time or predictions of them from various task variables, as well as estimates of error probabilities) for use with existing discrete-event simulation software packages to predict user task performance with speech interfaces in motor vehicles. This supports the design of safer and easier to use speech interfaces in vehicles that can minimize how long and how often drivers look away from the road. That, in turn, should reduce crash risk and thereby protect public health.

This effort was undertaken in three distinct phases:

1. Conduct a preliminary survey to determine task frequencies, and the information in user-generated databases (e.g., prior destination lists, play lists, etc.) to assure that real tasks and constraints are considered in the model.
2. Conduct a driving simulator experiment examining existing in-vehicle speech interfaces to understand how drivers perform destination entry and music selection and to determine the time drivers need to construct utterances (and their distributions), the types of errors drivers make and their probabilities, and the probability of various correction strategies are used for each error. These data will help determine the model structure and parameter values.
3. Develop and validate a simulation model to predict the user performance with a speech interface for destination entry and music selection tasks.

1.4 Thesis Outline

Based on the above research objectives, this thesis is comprised of five subsequent chapters. The outline is summarized as follows:

Chapter 2 - Review of Related Studies: This chapter reviews previous human factors research, and linguistics terms and principles related to this thesis research.

Chapter 3 – Survey of Navigation Systems and MP3 Players Use by Typical Drivers and Auto Experts: This chapter provides data on how and for what purposes drivers use navigation devices and MP3 players. These data are used to identify test conditions for the next experiment, as well as relevant variables and parameters for the simulation model to be built.

Chapter 4 – Destination Entry and Music Selection Using Speech in a Driving Simulator: This chapter provides data on the time drivers need to construct utterances (and their distributions), the types of errors that drivers make and their probability, and the probability of various correction strategies used, data used to construct a simulation model of driver use of speech interfaces.

Chapter 5 – Development of a Simulation Model to Predict Driver Task Performance Using Speech While Driving for Destination Entry and Music Selection Tasks: This chapter describes the construction and validation of a simulation model to predict the driver task performance with a speech interface for destination entry and music selection.

Chapter 6 – Discussion, Conclusions and Future Work: The final chapter summarizes the impact of this research and discusses future research directions.

CHAPTER 2

Review of Related Studies

In this chapter, previous studies, and linguistics terms and principles related to this thesis research are reviewed. The studies give insight into the motivations and applications of this research, as well as reviewing contemporary technologies. After reviewing these studies, the author summarizes and describes how to address these problems based on this thesis work. The major content of this chapter is reproduced from the author's previously published paper: "Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and Related Literature," published in the *International Journal of Vehicular Technology*, 2013 [17].

2.1 Examples of Automotive Speech

In the U.S., current speech interfaces include Ford SYNC®, Chrysler UConnect®, GM MyLink®, Hyundai Genesis®, Toyota navigation with Entune™, and others. The commonly supported applications are navigation (e.g., entering destination, following route guidance, and receiving traffic information) and music (selecting, playing, and pausing songs on MP3 players, AM/FM/XM radios), as well as those related to cellular phones (answering and placing calls, searching contact lists, and various tasks associated with text messages).

These systems were developed based on ideas from a number of predecessor systems (Tables 2-1 and 2-2). Notice that the core functions were navigation, music selection, and cellular phone support, and that many of them started out as either university or collaborative research projects involving several partners. In several cases, the result was either a product or ideas that later led to products. Of them, probably SYNC® has received the most attention.

The best-known nonautomotive natural speech interface is Siri, released by Apple in October 2011. Siri can help users make a phone call, find a business and get directions, schedule reminders and meetings, search the web, and perform other tasks supported by built-in apps on the Apple iPhone 4S and iPhone 5.

Similarly, Google's Voice Actions supports voice search on Android phones (http://www.google.com/mobile/voice-actions/, retrieved May 14, 2012). This application supports sending text messages and email, writing notes, calling businesses and contacts, listening to music, getting directions, viewing a map, viewing websites, and searching webpages. Both Siri and Voice Actions require off-board processing, which is not the case for most in-vehicle speech interfaces.

Table 2-1. Examples of Well-Known Automotive Speech Interfaces and Applications They Support

System	*Research/ Product*	*Navigation*	*Restaurant finder*	*Music selection*	*Audio + CD*	*Car info*	*Traffic info*	*Cellular phone*	*Text message*
CHAT [18], [19]	Research	X	X	X					
CU Move [20]	Research	X					X		
Ford Model U [21]	Research	X		X		X		X	
Linguatronic [22]	Product	X		X		X		X	
SENECA [9]	Research	X			X			X	
SYNC® [23]	Product	X		X		X		X	X
VOIC [24]	Research	X		X		X	X		
Volkswagen [16]	Product	X						X	

Table 2-2. Origins of Some Well-Known Automotive Speech Applications

System	*Full Name*	*Developed by*
CHAT	Conversational Helper for Automotive Tasks	Center for the Study of Language and Information at Stanford University, Research and Technology Center at Bosch, Electronics Research Lab at Volkswagen of America, and Speech Technology and Research Lab at SRI International. [18], [19]
CU Move	Colorado University Move	University of Colorado speech group in 1999. [20]
Ford Model U		Ford. [21]
Linguatronic		DaimlerChrysler Research and Technology in Ulm, Germany and TEMIC in 1996. [22]
SENECA	Speech control modules for Entertainment, Navigation, and communication Equipment in CArs	EU-project involving DaimlerChrysler, TEMIC Research and Department of Information Technology, University of Ulm. [9]
SYNC		Ford in collaboration with Microsoft and Nuance. [23]
VOIC	Virtual Intelligent Co-Driver	European project funded by five different partners: Robert Bosch GmbH, DaimlerChrysler AG, ITCirst, the University of Southern Denmark, and Phonetic Topographics N.V. [24]
Volkswagen		Volkswagen. [16]

2.2 Speech Interface Use

Real-world data on the use of speech applications in motor vehicles is extremely limited. One could assume that anyone who drives is a candidate user, but one could speculate that the most technically savvy are the most likely users.

How often these interfaces are used for various tasks is largely unknown. The authors do not know of any published studies on the frequency of use of automotive speech interfaces by average drivers, though they probably exist.

The most relevant information available is a study by Lo et al. [25] concerning navigation-system use, which primarily concerned visual-manual interfaces. In this study, 30 ordinary drivers and 11 auto experts (mostly engineers employed by Nissan) completed a survey and allowed the authors to download data from their personal navigation systems. Data was collected regarding the purpose of trips (business was most common) and the driver's familiarity with the destination. Interestingly, navigation systems were used to drive to familiar destinations. Within these two groups, use of speech interfaces was quite limited, with only two of the ordinary drivers and two of the auto experts using speech interfaces. The article also contains considerable detail on the method of address entry (street address being used about half of the time followed by point of interest (POI)) and other information useful in developing evaluations of navigation systems.

Also relevant is the Winter et al. [26] data on typical utterance patterns for speech interfaces, what drivers would naturally say if unconstrained. Included in that paper is information on the number and types of words in utterances, the frequency of specific words, and other information needed to recognize driver utterances for radio tuning, music selection, phone dialing, and POI and street-address entry. Takeda et al. [27] presents related research on in-vehicle corpora, which may be a useful resource to address on who, when, and how often drivers used the speech interfaces.

2.3 Key Research Findings on the User Performance Comparisons of Using Speech Interfaces vs. Visual-Manual Interfaces.

There have been a number of studies on this topic. Readers interested in the research should read Barón and Green [28] and then read more recent studies.

Using Barón and Green [28] as a starting point, studies of the effects of speech interfaces on driving are summarized in four tables. Table 2-3 summarizes bench-top studies of various in-vehicle speech interfaces. Notice that the value of the statistics varied quite widely between speech interfaces, mainly because the tasks examined were quite different. As an example for CU-Communicator [29], the task required the subject to reserve a one-way or round-trip flight within or outside the United States with a phone call. Performing this task involved many turns between users and machines (total 38 turns) and the task took almost 4.5 minutes to complete. Within speech interfaces, task-completion time varied from task to task depending upon the task complexity [24], [16].

Table 2-4, which concerns driving performance, shows that the use of speech interfaces as opposed to visual-manual interfaces led to better lane keeping (e.g., small standard deviation of lane position).

Table 2-3. Speech Interface Performance Statistics from Selected Bench-Top Studies

System	*CHAT [19]*	*CHAT [18]*	*CU Communicator [29]*	*CU Move [20]*	*SENECA [9]*	*VOIC [24]*	*Volkswagen [16]*
Tasks	1. NAV 2. Restaurant Finder (RF)	1. MP3 2. Restaurant Finder (RF)	1. Phone for travel plan	1. NAV	1. NAV 2. Phone Dialing 3. Address Book	1. NAV 2. Current time 3.Tourism 4. Fuel 5. Car Manual 6. Hotel Reservation 7. Traffic Information	1. NAV 2. Map Control 3. Phone
Completion Time (s)			260.3		Mean: 63	1. 73 2. 15 3. 146 4. 78 5. 57 6. 180 7. 53	
Completion Rate	98%	MP3: 98% RF: 94%	73.6%		Mean: 79%		
Turns[1]	2.3[1]	RF: 4.1[1]	User: 19 Machine: 19				
Word Recognition Accuracy[2] (%)	NAV: 85.5% RF: 85%	MP3: 90% RF: 85%					
Word Error Rate (%)			26%	30-65%			

User Satisfaction Rating[3]	1.98	MP3: 2.24 RF: 2.04

Notes: 1. *Turn* is defined as one user utterance to the system during a dialog exchange between the user and the system while attempting to perform the task.

2. Word Recognition Accuracy (WA)

$WA = 100 (1 – (Ws + Wi + Wd) / W)$ %

W: Total number of words in reference

Ws: number of reference words that were substituted in output

Wi: Number of reference words that were inserted in output

Wd: Number of reference words that were deleted in output

3. User Satisfaction Rating: 1 = strong agreement; 5 =strong disagreement

Table 2-4. Driving Performance Statistics from Selected Studies (S: speech; M: manual, K: Keyboard)

Study	*Method*	*Lane Keeping*	*Brake Reaction Time*	*Peripheral Detection Time*	*Following Distance*
Carter & Graham [5]	Simulator	S < M	S < M		
Forlines et al. [6]	Simulator	S < M	No diff.		
Garay-Vega et al. [30]	Simulator	No diff.			
Gärtner et al. [31]	On road	S < M			
Itoh et al. [7]	Simulator	S < M	No diff.		
Maciej & Vollrath [8]	Simulator	S < M			
McCallum et al. [13]	Simulator	No diff.	No diff.		
Minker et al. [9]	On road	S < M			
Ranney et al. [10]	On road			S < M (0.8 vs. 0.87 s)	
Shutko et al. [11]	Simulator	S < M		S < M (Except incoming call)	
Tsimhoni et al. [32]	Simulator	S < K			S < K (88 vs. 167m)
Villing & Larsson [33]	On road				

Table 2-5 shows that task completion times for speech interfaces were sometimes less than for visual-manual interfaces and sometimes greater, even though people speak faster than they can key-in responses. This difference is due to the inability of the speech interface to correctly recognize what the driver says, requiring utterances to be repeated.

Speech recognition accuracy has been an important factor that affects the task performance. Kun et al. [34] reported that low recognition accuracy (44%) can lead to greater steering angle variance. Gellatly et al. [35] reported that driving performance (peak lateral acceleration, peak longitudinal acceleration) was not statistically affected until the 60% recognition accuracy level was reached. Gellatly et al. [35] also showed that the task completion time was also affected when the speech recognition accuracy was lower than 90%. Although speech recognition accuracy was found to affect driving and task performance, no research has been reported on drivers' responses to errors, how long drivers need to take to correct errors, or what strategies drivers use to correct errors. Understanding how users interact with the spoken dialogue systems can help designers improve system performance and make drivers feel more comfortable using speech interfaces.

Table 2-6 shows that when using speech interfaces while driving, as opposed to visual-manual interfaces, subjective workload was less, fewer glances were required, and glance durations were shorter.

In general, driving performance while using speech interfaces is generally better than when using visual-manual interfaces. That is, speech interfaces are less distracting.

Table 2-5. Task Performance Statistics from Selected Studies (S: speech; M: manual)

Study	*Task Completion Time*	*Speech Recognizer Rate*	*Task Completion Rate*
Carter & Graham [5]	S > M	92.7%	
Forlines et al. [6]	S < M (18.0 vs. 25.2 s)		
Garay-Vega et al. [30]	S (dialog-based) > M S (query-based) < M		
Gärtner et al. [31]	S > M Simple: 24.6 vs. 12.8 s Complex: 74.4 vs. 58.7 s	79.4% (recognition error rate: 20.6%)	
Minker et al. [9]	S < M (63 vs. 84 sec)		S < M (79 vs. 90 %)
Ranney et al. [10]	No difference		
Shutko et al. [11]	S < M (Except dialing phone)		
Villing & Larsson [33]	S > M		

Table 2-6. Subjective Rating and Driving Behavior Statistics from Selected Studies (S: speech; M: manual)

Study	*Subjective Rating*		*Driving Behavior - Glances*	
	Workload	*Preference*	*Glance Duration*	*Number of Glances*
Carter & Graham [5]	S < M			
Faerber et al. [36]		S is preferred	Radio or CD control: M: 1 s; S: 1/3 s Using phone: M: 1 s	Radio or CD control: M: 3; S: 1.1 Using phone: M: 12; S: 0
Garay-Vega et al. [30]	S (query-based) < M		S < M	
Gärtner et al. [31]	No difference			Simple task: no difference. Complex tasks: S < M
Itoh et al. [7]	S < M		S < M	
Maciej & Vollrath [8]			S < M	
McCallum et al. [13]	S < M			
Minker et al. [9]		S is preferred		
Shutko et al. [11]			S < M (Except receiving an incoming call)	

2.4 Key Design Standards and References, Design Principles, and Results from Research

2.4.1 Relevant Design and Evaluation Standards

For speech interfaces, the classic design guidelines are Schumacher et al. [37], and the one set that is not very well known, but extremely useful, is the Intuity Guidelines [38]. Najjar et al. [39] described user-interface design guidelines for speech recognition applications. Hua and Ng [40] also proposed guidelines on in-vehicle speech interfaces based on a case study.

Several technical standards address the topic of the evaluation of speech system performance. These standards, such as ISO 9921: 2003 (Ergonomics – Assessment of speech communication) [41], ISO 19358: 2002 (Ergonomics – Construction and application of tests for speech technology) [42], ISO/IEC 2382-29: 1999 (Artificial intelligence – Speech recognition and synthesis) [43], and ISO 8253-3: 2012 (Acoustics – Audiometric tests methods –Part 3: Speech Audiometry) [44], focus on the evaluation of the whole system and its components. However, no usability standards related to speech interfaces have emerged other than ISO/TR 16982:2002 (Ergonomics of human-system interaction – Usability methods supporting human-centered design) [45].

From its title (Road vehicles — Ergonomic aspects of transport information and control systems — Specifications for in-vehicle auditory presentation), one would think that ISO 15006:2011 [46] is relevant. In fact, ISO 15006 concerns nonspoken warnings.

There are standards in development. SAE J2988, Voice User Interface Principles and Guidelines [47], contains 19 high-level principles (e.g., principle 17: "Audible lists should be limited in length and content so as not to overwhelm the user's short-term memory."). Unfortunately, no quantitative specifications are provided. The draft mixes definitions and guidance in multiple sections making the document difficult to use, does not support guidance with references, and in fact, has no references.

The National Highway Traffic Safety Administration (NHTSA) of the U.S. Department of Transportation posted proposed visual-manual driver-distraction guidelines for in-vehicle electronic devices for public comment on February 15, 2012 (http://www.nhtsa.gov/About+NHTSA/Press+Releases/2012/U.S.+Department+of+Transportation+Proposes+'Distraction'+Guidelines+for+Automakers, retrieved May 15, 2012), and "final" guidelines appeared this week (http://www.nhtsa.gov/staticfiles/nti/distracted_driving/pdf/distracted_guidelines-FR_04232013.pdf, retrieved April 23, 2012). NHTSA has plans to issue guidelines for speech interfaces.

The distraction focus group of the International Telecommunication Union (FG-Distraction - ITU) is interested in speech interfaces and may eventually issue documents on this topic, but what and when is unknown. In addition, various ITU documents that concern speech-quality assessment may be relevant, though they were intended for telephone applications. ITU-P.800 (Methods for subjective determination of transmission quality) and related documents are of particular interest. See http://www.itu.int/rec/T-REC-P/e.

2.4.2 Key Linguistic Principles

The linguistic literature provides a framework for describing the interaction, the kinds of errors that occur, and how they could be corrected. Four topics are touched upon here.

Turn and Turn-taking. When can the user speak? When does the user expect the system to speak? Taking a turn refers to an uninterrupted speech sequence. Thus, the back and forth dialog between a person and a device is turn-taking, and the number of turns is a key measure of an interface's usability, with fewer turns indicating a better interface. In general, overlapping turns, where both parties speak at the same time, account for less than 5% of the turns that occur while talking [48]. The amount of time between turns is quite small, generally less than a few hundred milliseconds. Given the time required to plan an utterance, planning starts before the previous speaker finishes the utterance.

One of the important differences between human-human and human-machine interactions is that humans often provide nonverbal feedback that indicates whether they understand what is said (e.g., head nodding), which facilitates interaction and control of turn taking. Most speech interfaces do not have the ability to process or provide this type of feedback.

A related point is that most human-human interactions accept interruptions (also known as barge-in), which makes interactions more efficient and alters turn taking. Many speech interfaces support barge-in, which sometimes requires the users to press the voice-activation button. However, less than 10% of subjects (unpublished data from the author) knew and used this function.

Utterance Types (Speech Acts). Speech acts refer to the kinds of utterances made and their effect [49]. According to Akmajian et al. [50], there are four categories of speech acts:

- Utterance acts include uttering sounds, syllables, words, phrases, and sentences from a language including filler words ("umm").
- Illocutionary acts include asking, promising, answering, and reporting. Most of what is said in a typical conversation is this type of act.
- Perlocutionary acts are utterances that produce an effect on the listener, such as inspiration, persuasion, etc.
- Propositional acts are acts in which the speaker refers to or predicts something.

Searle [51] classifies speech acts into five categories:

- Assertives commit the speaker to address something (suggesting, swearing, and concluding).
- Directives get the listener to do something (asking, ordering, inviting).
- Commissives commit the speaker to some future course of action (promising, planning).
- Expressives express the psychological state of the speaker (thanking, apologizing, welcoming).

Declarations bring a different state to either speaker or listener (such as "You are fired.")

Intent and Common Understanding (Conversational Implicatures and Grounding) Sometimes speakers can communicate more than is uttered. Grice [52] proposed that conversations are governed by the cooperative principle, which means that speakers make conversational contributions at each turn to achieve the purpose or direction of a conversation. He proposed four high levels conversational maxims that may be thought of as usability principles (Table 2-7).

Table 2-7. Grice's Conversational Maxims with Examples

Maxim	*Example*	*Guidance*
Maxim of Quantity: be informative.	a. **M**achine: Please say the street name. **U**ser: 2901 Baxter Road (2901 is the house number)	b. Make your contribution of information as is required, i.e., for the current purpose of the conversation. c. Do not make your contribution more informative than is required.
Maxim of Quality: make your contribution one that is true:	d. **U**: "Toledo Zoo, Michigan" (but Toledo is in Ohio)	e. Do not say what you believe to be false. f. Do not say that for which you lack evidence.
Maxim of Relevance: be relevant.	a. **U**: "I want to go to 'Best Buy'" and the system responds with all Best Buy stores, including those hundreds of miles away, not just the local ones.	
Maxim of Manner: be perspicuous	**g.** **M**: Please say set as destination, dial or back. **U**: Dial. O no Don't dial, Back (User wants to say "back") **h.** **M**: Please say the POI category. **U**: Let's see. Recreation	i. Avoid obscurity of expression j. Avoid unnecessary ambiguity k. Be brief (avoid unnecessary prolixity) l. Be orderly.

Errors Skantze [53] provides one of the better-known schemes for classifying errors (Table 2-8). Notice that Skantze does so from the perspective of a device presenting an utterance and then processing a response from a user.

Table 2-8. Examples of Errors in Different Modules of Speech-Controlled Interfaces (Adapted from Sakntze [53])

Modules	*Possible sources of errors*
Speech detection	Truncated utterances, artifacts such as noise and side talk; Barge-in problems
Speech recognition	Insertions, deletions, substitutions
Language processing/parsing	Concept failure, Speech act tagging
Dialogue manager	Error in reference resolution, error in plan recognition
Response generation	Ambiguous references, too much information presented at once, TTS quality, audio quality

Véronis [54] presents a more detailed error-classification scheme that considers device and user errors, as well as the linguistic level (lexical, syntactic, semantic). Table 9 is an enhanced version of that scheme. Competence, one of the characteristics in his scheme, is the knowledge the user has of his or her language, whereas performance is the actual use of the language in real-life situations [55]. Competence errors result from the failure to abide by linguistic rules or from a lack of knowledge of those rules ("the information from users is not in the database"), whereas performance errors are made despite knowledge of rules ("the interface does not hear users' input correctly").

Table 2-9. Enhanced Version of Véronis [54] Error-Classification Scheme

	System		*User*	
	Performance	*Competence*	*Performance*	*Competence*
Lexical level (word)	Letter substitution Letter insertion Letter deletion	Word missing in dictionary Missing inflection rule	Letter substitution Letter insertion Letter deletion Letter transposition Syllabic error Slips of tongue	Non-word or completely garbled word
Syntactic level (sentence structure)		Missing rule	Word substitution Word insertion Word deletion Word transposition	Construction error
Semantic level (meaning)		Incomplete or contradictory knowledge representation Unexpected situation		<u>Conceptual error</u> include: Incomplete or contradictory knowledge representation <u>Pragmatic error</u> Dialogue law violation

As an example, a POI category requested by the user that was not in the database would be a semantic competence error. Problems in spelling a word would be a lexical performance error. Inserting an extra word in a sequence ("iPod iPod play …") would be a lexical performance error.

A well-designed speech interface should help avoid errors, and when they occur, facilitate correction. Strategies to correct errors include repeating and rephrasing the utterances, spelling out words, contradicting a system response, correcting using a different modality (e.g., manual entry instead of speech), and restarting, among others [56] [57] [58] [59].

Knowing how often these strategies occur suggests what needs to be supported by the interface. The SENECA project ([9], [31]) revealed that the most frequent errors for destination entry tasks were spelling problems of various types, entering or choosing the wrong street, and using wrong commands. For phone dialing tasks, the most frequent errors were stops within digit sequences. In general, most of the user errors were vocabulary errors (partly spelling errors), dialogue flow errors, and PTA (push to active) errors, i.e. missing or inappropriate PTA activation.

Lo, Walls, and Green [60] reported that construction and relationship errors were 16% and 37%, respectively. Construction errors occur when subjects repeat words, forget to say command words (a violation of grounding), or forget to say any other words that were given. Relationship errors occur when subjects make incorrect matches between the given words and song title, album name and/or artist name. Relationship errors were common because subjects were not familiar with the given songs/albums/artists.

2.5 Performance Assessment and Measurement of Speech Interfaces

2.5.1 Methods Used to Measure User Performance

Given the lack of models to predict user performance with speech interfaces, the evaluation of the safety and usability (usability testing) of those interfaces has become even more important. Evaluations may either be performed only with the system itself (on a bench top) or with the system integrated into a motor vehicle (or a simulator cab) while driving.

The most commonly used method to evaluate in-vehicle speech interfaces is the Wizard of Oz method ([18], [19], [24], [29], [61], [62], [63]), sometimes implemented using Suede [64]. In a Wizard of Oz experiment, subjects believe that they are interacting with a computer system, not a person simulating one. The “wizard” (experimenter), who is remote from the subject, observes the subject’s actions, and simulates the system’s responses in real-time. To simulate a speech-recognition application, the wizard would type what users say, or in a text-to-speech system, they read the text output, often in a machine-like voice. Usually, it is much easier to tell a person how to emulate a machine than to write the software to tell a computer to do it. The Wizard of Oz method allows for the rapid simulation of speech interfaces and the collection of data from users interacting with a speech interface, allowing for multiple iterations of an interface to be tested and redesigned.

2.5.2 Factors to Be Measured for User Performance

Dybkjær has written several papers on speech interface evaluation, the most thorough of which is Dybkjær, Bersen, and Minker [65]. That paper identified a number of variables that could be measured (Table 2-10), in part because there are many attributes to consider.

Table 2-10. Variables Used for Evaluating Entire Systems or System Modules

Module	*Variables*
Whole system	Task completion time, task completion rate, transaction success, number of interaction problems, query density, concept efficiency
Speech recognition	Word and sentence error rate, vocabulary coverage, perplexity
Speech synthesizer	User perception, speech intelligibility, pleasantness, naturalness
Language understanding	Lexical coverage, grammar coverage, real-time performance. concept accuracy, concept error rate

Walker, Kamm, and Litman [66] proposed a framework of usability evaluation of spoken dialogue systems, known as PARADISE (PARAdigm for DIalogue System Evaluation). (See [67] for criticisms.) Equations were developed to predict dialog efficiency (which depends on mean elapsed time and the mean number of user moves), dialog quality costs (which depends on the number of missing responses, the number of errors, and many other factors, and task success, measured by the Kappa coefficient and defined below).

$$\kappa = (P(A) - P(E)) / (1 - P(E)) \tag{2.1}$$

Where:

P(A) = proportion of times that the actual set of dialogues agree with scenario keys

P(E) = proportion of times that the dialogues and the keys are expected to agree by chance.

In terms of performance while driving, there is no standard or common method for evaluating speech interfaces, with evidence from bench-top, simulator, and on-road experiments being used. There are two important points to keep in mind when conducting such evaluations. First, in simulator and on-road experiments, the performance on the secondary speech interface task depends on the demand of the primary driving task. However, the demand or workload of that task is rarely quantified [68], [69]. Second, there is great inconsistency in how secondary-task performance measures are defined, if they are defined at all, making the comparison of evaluations quite difficult [70]. (See [71] for more information.) Using the definitions in SAE Recommended Practice J2944 [72] is recommended.

2.6 Summary

The issues discussed in this article are probably just a few of those that should be considered in a systematic approach to the design and development of speech interfaces.

As automotive speech interfaces move close to production, the safety and usability of those interfaces are usually assessed in a driving simulator, and sometimes on the road. The linguistics literature provides a long list of potential measures of the speech interface that could be used, with task time being the most important. Driving-performance measures, such as standard deviation of lane position, gap variability, and so forth, are measured as eyes-off-road time. These studies often have two key weaknesses: (1) The demand/workload of the primary task is not quantified, yet performance on the secondary speech task can depend on primary task demand, and (2) measures and statistics describing primary task performance are not defined. A solution to the first problem is to use equations being developed by the second author to quantify primary task workload. The solution to the second problem is to use the measures and statistics in SAE Recommended Practice J2944 [72] and refer to it.

Driver distraction is and continues to be a major concern. Some view speech interfaces as a distraction-reducing alternative to visual-manual interfaces. Unfortunately, at this point, actual use by drivers and data on that use is almost zero. There is some information on how to test speech interfaces, but technical standards cover only a limited number of aspects.

There is very little to support design other than guidelines. For most engineered systems, developers use equations and models to predict system and user performance, with testing serving as verification of the design. For speech interfaces, those models do not exist. This research will provide some guidelines needed to create those models.

CHAPTER 3

Survey of Navigation Systems and MP3 Players Use by Typical Drivers and Auto Experts

The major content of this chapter is reproduced from the previously published paper: "Where Do People Drive? Navigation System Use by Typical Drivers and Auto Experts" in *Journal of Navigation*, 2011 [25] .

3.1 Background

Infotainment devices are widely used by drivers and have been the subject of considerable recent research ([73], [74], [75], [76]). The visual-manual tasks associated with operating many of these devices require drivers to look away from the road for some time and crash risk is known to increase with eyes-off-road time ([5], [32], [12], [7], [6, 9]). One potential way to reduce crash risk is to use speech-controlled interfaces instead of visual-manual interfaces.

In the past decade, many projects focused on either the development of speech-controlled interfaces or the comparison of performance of speech-controlled interfaces with visual-manual interfaces ([10], [23], [18], [19], [24]). Common applications include navigation, music selection and cellular phone use. Although system developers claim these systems are “human-centered,” there is limited research, especially for speech-controlled interfaces, on what users want to do or how these systems are actually used.

A human-centered navigation system should support travel to the most common destinations. One important source of information about these destinations is the National Household Travel Survey (NHTS), a U.S. Department of Transportation (DOT) National Highway Traffic Safety Administration funded periodic study of personal travel in the United States [77]. Similar studies have been conducted in other countries (e.g., United Kingdom, Denmark) [78]. Common trip purposes in these surveys included “shopping/errands,” “go to school,” and “medical/dental services,” which could be used to define the point of interest (POI) categories for navigation systems. However, estimating the real use of navigation systems from these surveys may be speculative, as one would suspect navigation systems to be used to provide guidance to unfamiliar destinations, not all destinations, and these surveys do not address how destinations are entered into navigation systems. Furthermore, designers, in the absence of information, tend to think of users as being like themselves and, thus, in some instances, may design the systems for themselves. Therefore, data comparing the travel patterns of vehicle designers with the general public is needed.

This information is particularly important for assessing the compliance of navigation systems with existing design guidelines, in particular AAM guideline 2.1a [79] and SAE Recommended Practice J2364 [80]. These guidelines require that destination retrieval times not exceed specific maxima. However, designers often must compromise to meet these criteria: the goal is to facilitate the use of the most commonly used methods, for which at this point there are no published statistics. Furthermore, most navigation systems provide lists of recently visited destinations and saved favorites, from which users can search. Since there are no published data on the use or content of these lists, populating them requires guesswork, making safety and usability tests potentially unrealistic.

The use of radio, cassette, and CD ranked second among major causes of crashes involving distracted driving [3]. Nowadays, MP3 players is much more popular and widely use than the CDs and cassettes. However, less attention has been made on how the drivers use the MP3 players and how often various MP3 player features are used. Also, there are no specific guidelines concerning the design of in-vehicle music selection systems.

The purpose of this study was to provide the data on how and for which tasks drivers use navigation devices and MP3 players. The following hypotheses were examined.

1. *Drivers use their navigation system only for when they go to unfamiliar destinations.* This seems reasonable, as the purpose of these systems is to help drivers get to places where they do not know how to get there.
2. *There is no difference of the trip purpose based on the POI categories when drivers used their own navigation systems.*
3. *There are no differences in how frequently various entry methods used and destination types are reportedly selected and they are actually used (based upon transcribed data from in-vehicle devices).*
4. *There is no difference of the POI categories the records saved as drivers' favorites in their navigation devices.*

5. *Between auto experts and typical drivers, there is no difference on the numbers of songs, artists, and albums they have stored on their electronic music player.*
6. *There is no difference of the methods used to search a specific song by the drivers between the auto experts and typical drivers.*

Finally, these data will be used to identify test conditions for the next experiments, and as well as relevant variables and parameters for the simulation model to be built.

3.2 Methods and Materials

3.2.1 Subjects and Their Navigation Systems and MP3 Players

Thirty licensed drivers (16 F, 14 M; 28 ± 10 years) from southeast Michigan (*typical drivers*) who regularly use their navigation systems while driving were recruited via newspaper advertisements, web advertisements, and an email sent to students, faculty, and staff at the University of Michigan, friends of the authors, and members of the community. This sample was chosen largely for the convenience of accessing subjects and their vehicles. More than half (17/30) of the subjects in the group of typical drivers were students from a wide variety of academic disciplines. Only six of the thirty subjects were engineers (five engineering students and one mechanical engineer). Twenty-three of the thirty typical drivers were native English speakers. A second group of eleven licensed drivers (1F, 10 M; 39 ± 10 years) who regularly use their navigation systems while driving was recruited from the Nissan Technical Center in Farmington Hills, Michigan (*auto experts*). Most of the auto experts were engineers (mechanical, electrical, and project engineers). Ten of the eleven auto experts were native English speakers. All navigation systems set for English. It was expected that they would be different from typical drivers and they were selected to reinforce the principle that auto designers should not design for themselves, as they are different from typical drivers, the customers.

The typical drivers drove vehicles from a wide range of manufacturers, with Toyota (7/30) and Ford (6/30) being most common. Vehicles were typically of the 2003 and 2004 model years. All auto experts drove Nissan vehicles (most commonly of the 2007 and 2008 model years), reflecting an employee benefit. Data were collected in 2009.

Typical drivers predominately used portable aftermarket navigation systems made by Garmin (n=14), Tom Tom (n=5), and other manufacturers (n=7), but some used various types of built-in navigation systems (n=4) as well. For the auto experts, ten of the eleven subjects used Nissan built-in navigation systems, and one subject used a portable Garmin unit. The mean years of owning their current navigation devices were 3 ± 1.5 years for typical drivers and 2 ± 0.5 years for auto experts.

Omitting one Nissan employee who reported that she drove more than 128,000 miles/year, the typical drivers reported a mean of 10,900 miles/year, whereas the mean was 14,000 miles annually for the auto experts. To provide some context, the most recent population survey data indicated a mean annual mileage of about 14,000 miles in the United States [77], but only 3,500 miles in the United Kingdom [81], where public transportation is more readily available and gasoline is more expensive. Thus, the typical driver sample here drove somewhat fewer miles than in another larger survey, but the auto experts were quite close to the U.S. survey's mean.

3.2.2 Questionnaire

A seven-page questionnaire was designed to collect biographic information about the subjects as well as information on their navigation systems and MP3 players. This paper reports results for navigation systems. There were eight questions concerning biographical information, such as name, age, vehicle driven, miles driven per year, as well as information on the frequency of various types of trips. Using a modification of Hu and Reuscher's scheme [77], trips were categorized as business, vacation, religious, shopping, or school. There were seven questions concerning each subject's navigation system, such as the manufacturer, familiarity with each destination, percent and time using speech/manual input, the frequency of use of various destination entry methods, and error correction strategies for speech entries. Subject's familiarity with recently visited destinations was categorized as "Capable of getting to location without navigation directions (familiar)," "Capable of getting to location with some navigation directions (somewhat familiar)," and "Incapable of getting to location without navigation directions (unfamiliar)." Error correction strategies modified from Bourguet's scheme were categorized as "repeat exactly the same words," "rephrase or say it in different words," "spell the words out," and "correct by entering in the words manually." [56] There were eleven questions concerning each subject's MP3 players, such as the manufacturer, number of features in the MP3 players (such as songs, playlists, etc.), percent and time using speech/manual input, the frequency of use of MP3 players to various destinations, the frequency of use of various music selection methods, frequency of changing features while driving more than 30 minutes, and error correction strategies for speech entries.

3.2.3 Procedure

The typical drivers brought their vehicles and navigation systems to the University of Michigan Transportation Research Institute (UMTRI) in Ann Arbor, Michigan. They were given an overview of the study and signed the consent form that had been approved by the University of Michigan Health and Behavioral Sciences Institutional Review Board (IRB). Next, the previously described questionnaire was completed. While subjects answered the questions, the investigator transcribed and photographed the information stored in the subjects' navigation systems, including the content of their "Favorite" lists (destinations they entered and saved) and "History" lists (recent destinations selected). If there were more than nine records saved in their navigation system from either list, only the nine most recent trips and favorites from each list were transcribed for each subject due to time constraints. After the subjects finished the questionnaire, they were asked to confirm the trip purpose and entry method for the destinations stored on their "Favorite" and "History" lists. Finally, the typical drivers were paid $20 for their participation in this 40-minute survey.

At Nissan, a recruiting message, a consent form, and the questionnaire were distributed via email. A Nissan coordinator arranged for auto experts to participate, with interviews conducted at the Nissan Technical Center in Farmington Hills, Michigan. During the interviews, investigators checked to see that subjects answered all the previously disseminated questions, and then transcribed and photographed information identifying what was stored in each subject's navigation system. Other procedures were similar to the procedures for typical drivers. Since the auto experts were surveyed during business hours, they were not paid for their participation. The study required approximately 30 minutes per subject.

3.3 Results and Discussions

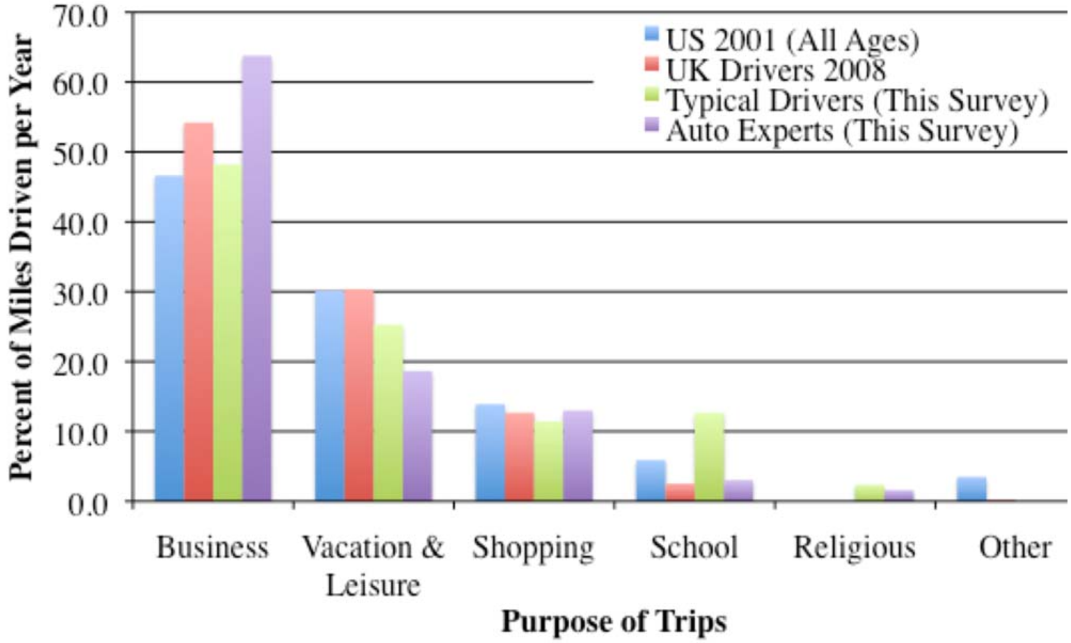
3.3.1 Number of Trips per Year

The total number of trips per year from the U.S. survey, 1388 trips/year/person [77], was three times more than reported here for both groups (369 trips/year/driver for typical drivers, and 409 trips/year/driver for auto experts), but both groups were almost the same as reported in the United Kingdom at 410 trips/year/driver [81]. It could be that the number of trips in the U.S. survey included all modes of transportation (private vehicle, public transit, walking, and others). Furthermore, the U.S. survey included 36 categories of trip purposes [77], but only five categories of trip purposes were used in the present survey. For example, the trip purpose "Medical/Dental Service" was not available to subjects in this study but was included in U.S. survey. Finally, the number of trips was estimated, and not obtained from a trip diary, which would have been more accurate.

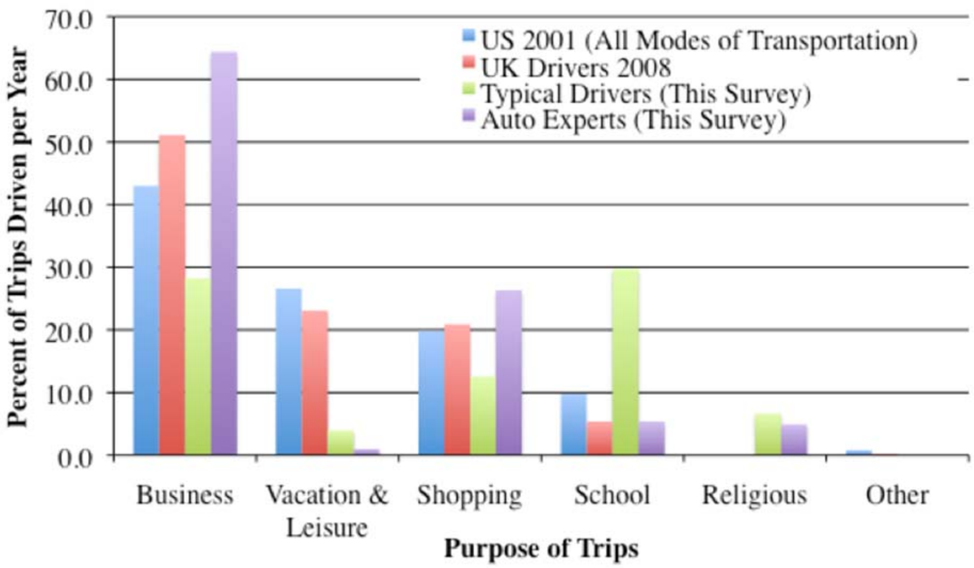
3.3.2 Purpose and Distances of the Trips

Each of the navigation systems and surveys used a different scheme for coding trip purpose and points of interest (POI). To ensure a sufficient number of responses for each POI/trip purpose category and to aid in consolidating the categories, POIs were classified using Garmin's scheme, which had the fewest number of categories. The investigators also included two more categories: friends' and relatives' houses, since they accounted for 15% and 4%, respectively, of the trips that subjects visited.

Figure 3-1(a) compares the distribution of miles driven by trip purposes from this survey with the U.K. [81] and U.S. [77] data. In the U.S. study, the results were the percentage of the annual miles travelled by people, not just by drivers. The distribution of miles driven by trip purpose for typical drivers is similar to the distribution from the U.K. study, except for school trips. This is not surprising, as 57% of the typical drivers were students. However, the auto experts in this study traveled relatively more miles for business trips and relatively fewer miles for vacation and leisure trips than those reported in the U.S. and U.K. studies, possibly because studies in the United States and United Kingdom included retired (age > 65) and younger (age < 16) subjects. Retired and young people are usually not engaged in business, and both groups have more time for vacation and leisure travel. Another possibility was that "vacation and leisure" are commonly thought of as having a longer duration, but subjects in this study did not count the one-day or two-day short trips to a recreation center, a friend's house, etc., as "vacation and leisure".



(a)



(b)

Figure 3-1. Comparison of (a) Percent of Miles Driven per Year and (b) Percent of Trips Driven Per Year for Both Groups and Other Studies.

Comparing the absolute number of trips, auto experts (409 trips/year/driver) drove more trips than typical drivers (369 trips/year/driver), which is consistent with having driven more miles. Further, their trip purpose distributions were quite different (Figure

3-1(b)). Auto experts drove more trips for business and shopping, whereas driving to school was common for typical drivers. Again, it is not surprising that there were fewer business trips and more school trips for typical drivers, as most of the study subjects were students. There is a large difference (28% vs. 64%) in the percentage of business trips between typical drivers and auto experts, but there is not much difference in the percentage of miles driven for business trips (48% vs. 64%). This might suggest that the average distance driven for one business trip is longer for typical drivers than the distance driven by auto experts.

The patterns of the trip purposes of each group were different from the studies in the United States and United Kingdom (Figure 3-1(b)). These differences are probably related to differences in the populations surveyed.

3.3.3 Necessity of Navigation Devices Use for Each Trip

When subjects used their navigation systems while driving, 61% (183/299) of the trips for typical drivers and 89% (212/239) of the trips for auto experts reported that they were familiar with the desired destinations; 21% (63/299) of the trips for typical drivers and 7% (16/239) of the trips for auto experts reported that they were somewhat familiar with the desired destinations; and only 18% (53/299) of the trips for typical drivers and 5% (11/239) of the trips for auto experts reported that they were unfamiliar with the desired destinations. Why drivers used their navigation systems for previously known destinations is unknown.

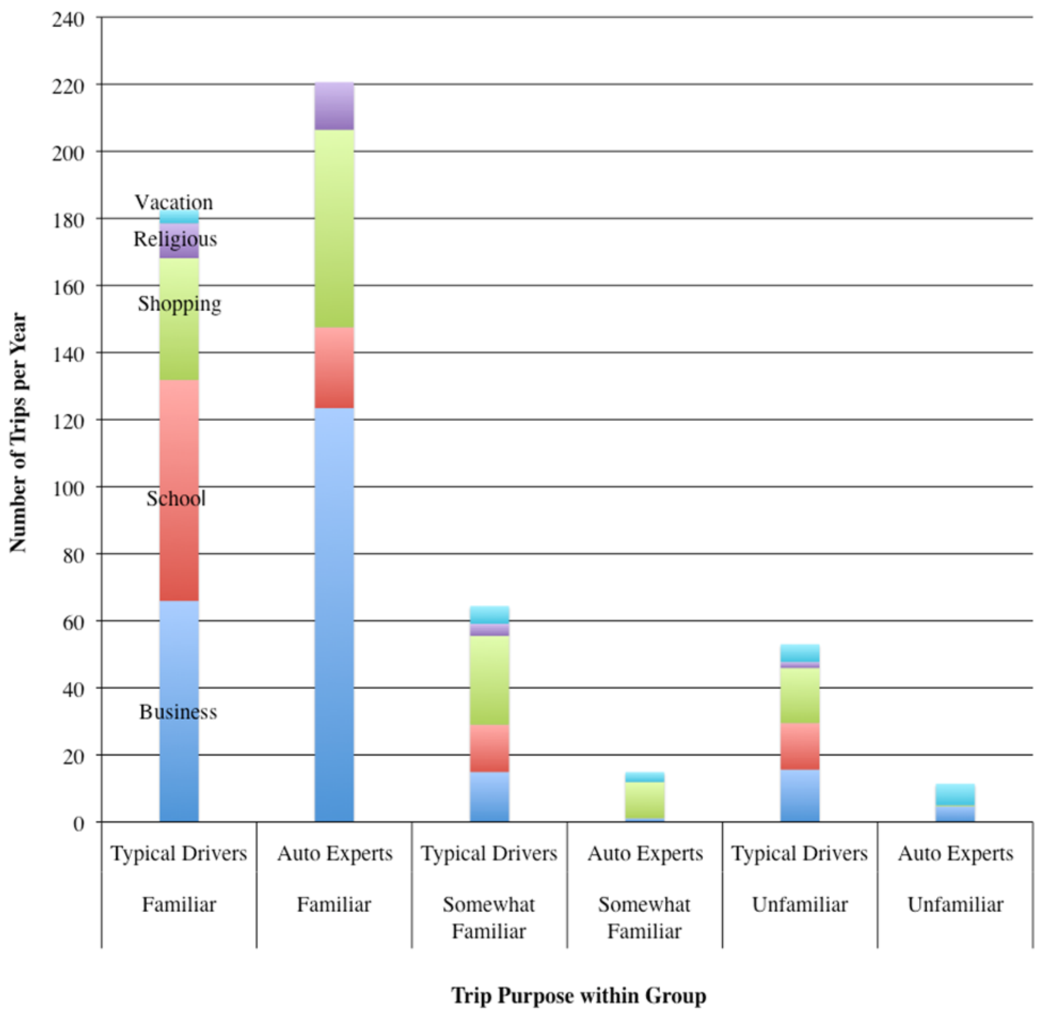


Figure 3-2. Reported Use of Navigation Assistance as a Function of Trip Purposes.

For the familiar destinations, the common purposes were "School" (36%) and "Business" (36%) for typical drivers, and "Business" (58%) for auto experts. For the category of "somewhat familiar with the destinations," the most common purpose was "Shopping" for both typical drivers (40%) and auto experts (69%). For unfamiliar destinations, there were no significant differences in the percentage for the purposes of "Business" (30%), School" (26%), and "Shopping" (30%) for typical drivers. For auto experts, "Vacation" (55%) and "Business" (45%) were the most common purposes. When auto experts used navigation systems, they did not report any trips for "School" and “Religious” purposes, for neither somewhat familiar nor unfamiliar destinations. In general, the results indicated that subjects from both groups still used navigation systems even when they were familiar with the destinations.

3.3.4 Reported Frequency and Task Completion Time of Manual and Speech Interfaces Use to Enter Destinations

Only two of the thirty typical drivers (8%), and two of the eleven auto experts (18%) reported that they had used speech interfaces to enter destinations. The estimated mean time to complete a destination entry task using the speech interface was 15 s for typical drivers and 158 s for auto experts (Table 3-1). Note that only two in each group reported using speech and so the data are limited.

Table 3-1. Reported Mean Time to Complete a Destination Entry Task Using Speech and Manual Inputs.

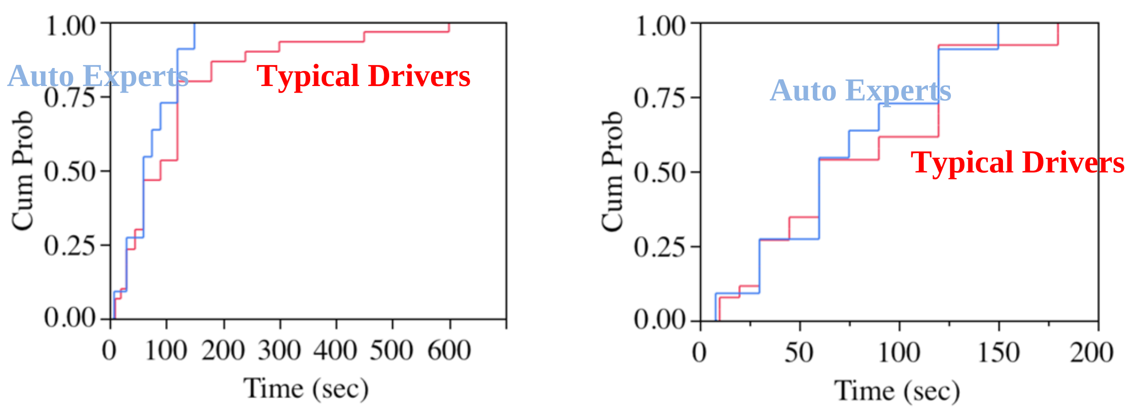
Interface	N	Typical Drivers Time (seconds) Mean ± S.D.	Typical Drivers Time (seconds) Range	N	Auto Experts Time (seconds) Mean ± S.D.	Auto Experts Time (seconds) Range
Speech	2	15 ± 0	15	2	158 ± 202	15 - 300
Manual	30	121 ± 130	10 - 600	11	73 ± 44	8 - 150
	26*	78 ± 49*	10 – 180*			

*: Without outlier

The estimated mean time to complete a destination entry task using the manual input was 121 ± 130 s for typical drivers and 73 ± 44 s for auto experts (Table 3-1). When the four outliers in the typical drivers were removed (an outlier is defined as any observation outside of the range: [Lower quartile – 1.5 (quartile difference), Upper quartile + 1.5 (quartile difference)], [82]), the estimated mean time to complete a destination entry task using the manual input was 78 ± 49 s. There were no outliers among the auto experts. There was no statistically significant difference between the two groups ($t_{(39)} = -1.20$, p=0.24, with outlier; $t_{(35)} = -0.34$, $p = 0.74$, without outlier) in the estimated mean time to complete a destination entry task using manual input.

The estimated destination entry task time distributions among typical drivers using manual input were quite different with outliers (lognormal distribution) than without outliers (exponential distribution). The estimated task times were normally distributed for the auto experts. When the outliers were removed from the data for typical drivers, the cumulative probability distributions were similar for both groups (Figure 3-3).

The estimated self-reported mean times to complete a destination entry task using the manual input for both groups were less than the measured times from other studies ([12], [31], [83]). When the outliers were removed from the data, the reported mean time to manually complete a destination entry task was a few seconds longer than the mean time from the Manstetten et al. study (73.2 s) [84]. The ranges of the time to manually complete a destination entry task were almost the same to those reported by Walls (24.7 s to 179 s) [83].



Cumulative Probability w/ Outlier

Cumulative Probability w/o Outlier

Figure 3-3. Cumulative Probability of Time Estimated to Complete a Destination Entry Task Using Manual Input from Both Groups.

3.3.5 Reported Methods Used to Enter Destinations by Trip Purpose

"History" (30%), "Street Address" (27%) and "POI" (26%) were the common methods reported by typical drivers to enter destinations for all purposes of trips. Auto experts reported that "Street Address" (48%) and "History" (24%) were common methods for destination entry for all purposes of trips. Figure 3-4 provides more detailed information on the method used by trip purpose. For business trips, the most popular method used by typical drivers was "History" and "Favorite," which accounted for a total of 60%. When typical drivers went shopping, the "POI" method was used almost half of the time, whereas auto experts preferred the "Street Address" method (62%). When going to school, typical drivers overwhelmingly searched the destination by "History" (69%). Auto experts did not report any school trips. Comparing the methods used for destination entry tasks, "Intersection" was the method both groups were least likely to use. As the reported numbers of trips for vacation and religious purposes were small, no conclusions could be drawn from the methods used.

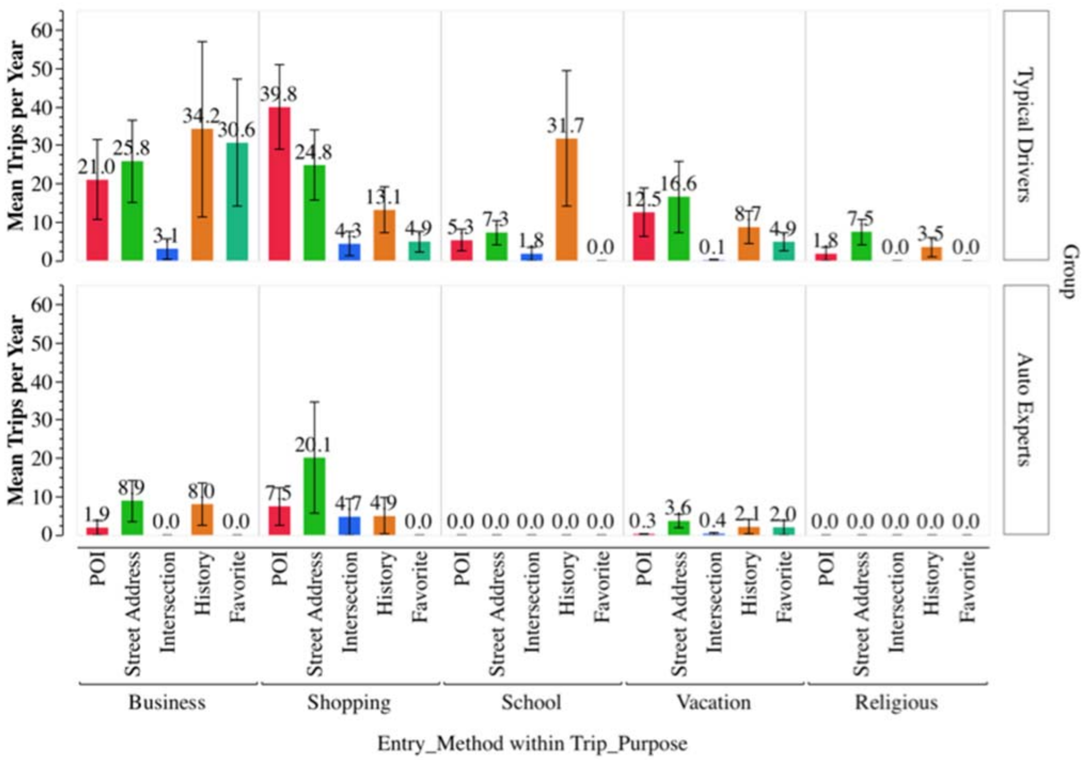


Figure 3-4. Reported Frequency of Destination Entry Method Use by Trip Purpose for Both Groups.

3.3.6 Reported Correction Strategies on Speech Entry Error Occurred

Fifty-seven percent of the time, typical drivers reported that they corrected a speech entry error by repeating exactly the same words, and thirty-seven percent of the time they used manual input to correct it (Table 3-2). On the other hand, sixty-three percent of the time auto experts corrected the error by entering the words manually. No subjects reported spelling the words out to correct an error. There are no empiric data to verify these claims, and there were only two subjects in each group that reported using this method.

Table 3-2. Mean Percentage (Ranges) of Error Correction Methods Using Speech.

Error Correction Method	Typical Drivers	Auto experts
Repeat exactly the same words	**57 ± 31 %** (30 - 90 %)	18 ± 28 % (0 - 50 %)
Manual input	37 ± 23 % (10 - 50%)	**63 ± 23 %** (50 - 90 %)
Rephrase or say it in different words	7 ± 12% (0 - 20 %)	18 ± 28 % (0 - 50 %)
Spell the words out	0 %	0 %

3.3.7 Frequency of Destinations That Drivers Actually Visited Recently Using Navigation Devices

There were 270 records transcribed from typical drivers and 91 records from auto experts. All subjects in the typical drivers group and nine of the eleven (82%) subjects in the auto experts group had more than nine records on the “History” lists in their navigation devices, so the data shown were from a self-selected data set, using only the first nine records in each device. As shown in Figure 3-5, “Shopping” and “Friends’ Houses” were the top two ranked destinations for typical drivers (17%). On the other hand, “Shopping,” “Community,” (which includes “School/University,” “Place of Worship,” Bank/ATM,” “Library”), “Food,” and “Others” were the top-ranked destinations for the auto experts. The high frequency for visiting “Friends’ Houses” and “Relatives’ Houses” is an important finding, because many destination entry studies use common entries, such as “POI” for restaurants or intersections for business offices to evaluate their navigation systems ([10], [18], [19], [24]). However, these data suggest the trials used for address entry task should include residential locations (friends’ houses and relatives’ houses), and furthermore, because these locations are for friends and family, they will be subject-specific.

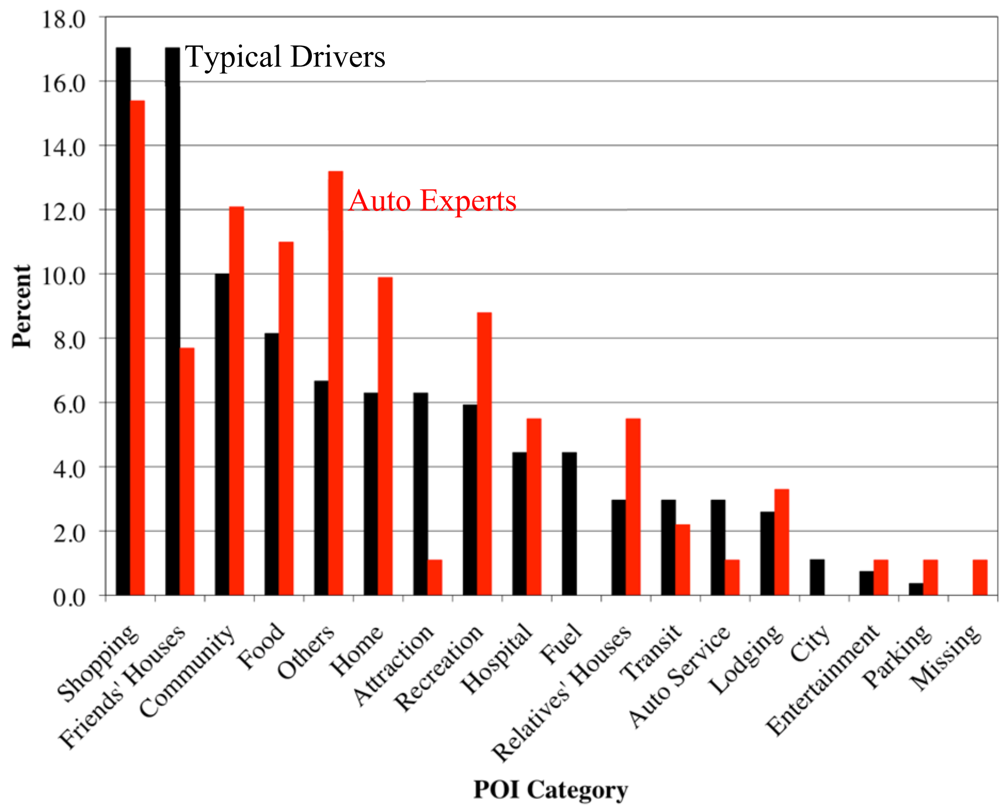


Figure 3-5. Frequency of POI Categories Visited for Both Groups.

3.3.8 Methods Drivers Actually Used to Enter Destinations

Table 3-3 compares reported methods with actual methods used by subjects to enter destinations. Keep in mind that destinations derived from the device history were only for trips in which destinations were entered, while the method estimated from the survey was for all trips. The most frequently used method to search for a destination was by "Street Address" for both groups, 43% for typical drivers and 55% for auto experts. The actual result for "Street Address" conflicted somewhat with the result from the survey for the typical drivers. "History" was only used once (0.3%) based on information retrieved from their devices. However, searching for a destination from "History" was the most frequently used method (30%) reported by typical drivers. Similarly, auto experts did not use the “History” method at all based on data retrieved from their devices, but it accounted for 24% of total use reported by auto experts. The results from actual use also contradict the results from the survey for both groups. "Favorite" was cited as the second most preferred method used by auto experts, but there was a substantial difference with the survey results, 20% vs. 3%, respectively. One possible explanation for the difference is that the method recorded in the device history is for part of the trips in which destinations are recently entered, but the method estimated from the survey is for all trips.

Table 3-3. Destination Entry Methods Recorded from Personal Devices Compared to Survey Results.

Entry Method	Typical Drivers (%)		Auto Experts (%)	
	Actual (when navi used)	Reported in Survey (all trips)	Actual (when navi used)	Reported in Survey (all trips)
Street Address	***43***	***27***	***55***	***48***
POI	35	26	18	16
Favorite	17	13	20	3
Intersection	3	3	2	8
City	1		0	
Default Emergency	0.3		0	
History	***0.3***	***30***	***0***	***24***
Map	0		2	
Near Different City	0		2	
Near Route	0		1	
Total	100% (270 trips)	100%	100% (91trips)	100%

Table 3-4 shows how destination types were distributed among the top three destination entry methods (“Street Address,” “Favorite,” and “POI”) from subjects’ devices. For typical drivers using the “Street Address” method, the most common destination was “Friends’ Houses” (31% of street address entries). There was no significant difference of destination on POI categories for auto experts using the “Street Address” method.

Table 3-4. Frequency of POI Categories by Top 3 Destination Entry Methods Used.

Entry Method	POI Category	Typical Drivers (%)	Auto Experts (%)
Street Address	***Friends' Houses***	***31***	14
	Community	12	18
	Other	12	12
	Recreation	8	10
	Home	6	0
	Hospital	6	8
	Shopping	6	14
	Relatives' Houses	4	8
	Attraction	4	0
	Food	3	10
	Auto Service	3	0
	Transit	2	0
	Lodging	1	4
	Entertainment	0	2
	Total Number	**100% (115)**	**100% (50)**
Favorite	***Home***	***19***	***50***
	Friends' Houses	***17***	0
	Shopping	***17***	0
	Community	6	0
	Food	6	0
	Recreation	6	0
	Transit	6	6
	Other	6	22
	Hospital	4	6
	Relatives' Houses	4	6
	Attraction	2	0
	Entertainment	2	0
	Fuel	2	0
	Parking	0	6
	Missing	0	6
	Total Number	**100% (47)**	**100% (18)**
POI	***Shopping***	***31***	25
	Food	***16***	***25***
	Fuel	11	0
	Attraction	11	6
	Community	8	13
	Lodging	6	0
	Recreation	4	***19***
	Auto Service	4	6
	Hospital	3	0
	Transit	3	6
	Entertainment	1	0

Entry Method	POI Category	Typical Drivers (%)	Auto Experts (%)
	Parking	1	0
	Total Number	**100% (95)**	**100% (16)**

"Shopping" and "Food" were the dominant destinations for both driver categories when using the "POI" entry method. Searching for "Recreation" (16%) was the third ranked most common destination when auto experts used the "POI" entry method. The most notable difference between driver categories was for destinations entered using the "Favorite" entry method: typical drivers used this method for "Home" (19%), "Friends' Houses" (17%), and "Shopping" (17%) with equal frequency, while auto experts used this method for "Home," but not at all for "Friends' Houses" and "Shopping." The relatively high frequency of visiting "Friends' Houses" or going "Home" as being destinations has not been identified previously in the literature. Since these residential locations are for social purposes, there may be an opportunity to aid drivers in selecting destinations by linking navigation systems to social networking sites, such as Facebook.

3.3.9 Frequency of Point of Interest (POI) Categories That Subjects Saved in "Favorite" Lists in Their Navigation Systems

Overall, there were 156 (5.2 records per subject) records for typical drivers and 39 (3.5 records per subject) records for auto experts on the "Favorite" lists saved in their navigation devices. Thirty-seven percent (11/30) of the typical drivers and eighteen percent (2/11) of the auto experts had more than nine records on their "Favorite" lists. Figure 3-6 shows the frequency of POI categories for these records. "Friends' Houses," "Home," "Shopping," and "Community" were the four top-ranked categories on the

"Favorite" lists of typical drivers. "Home" and "Other" were the most frequent categories on the "Favorite" lists of auto experts. These results confirmed that "Home" was the most common category for subjects when using their "Favorite" lists as the method to enter a destination.

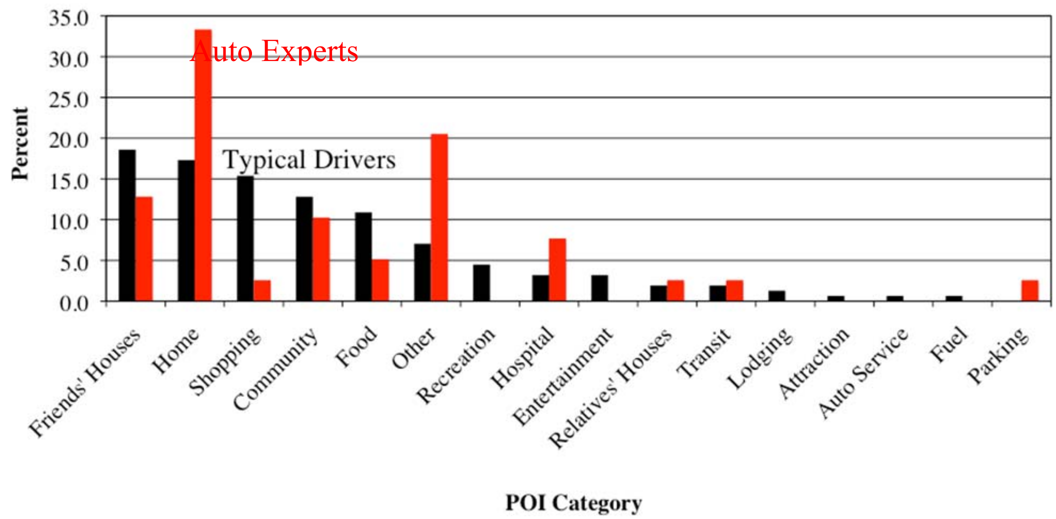


Figure 3-6. Frequency of POI Categories That Subjects Saved As Favorite in Their Navigation Devices.

3.3.10 Numbers of Each Feature in Subjects' MP3 Players

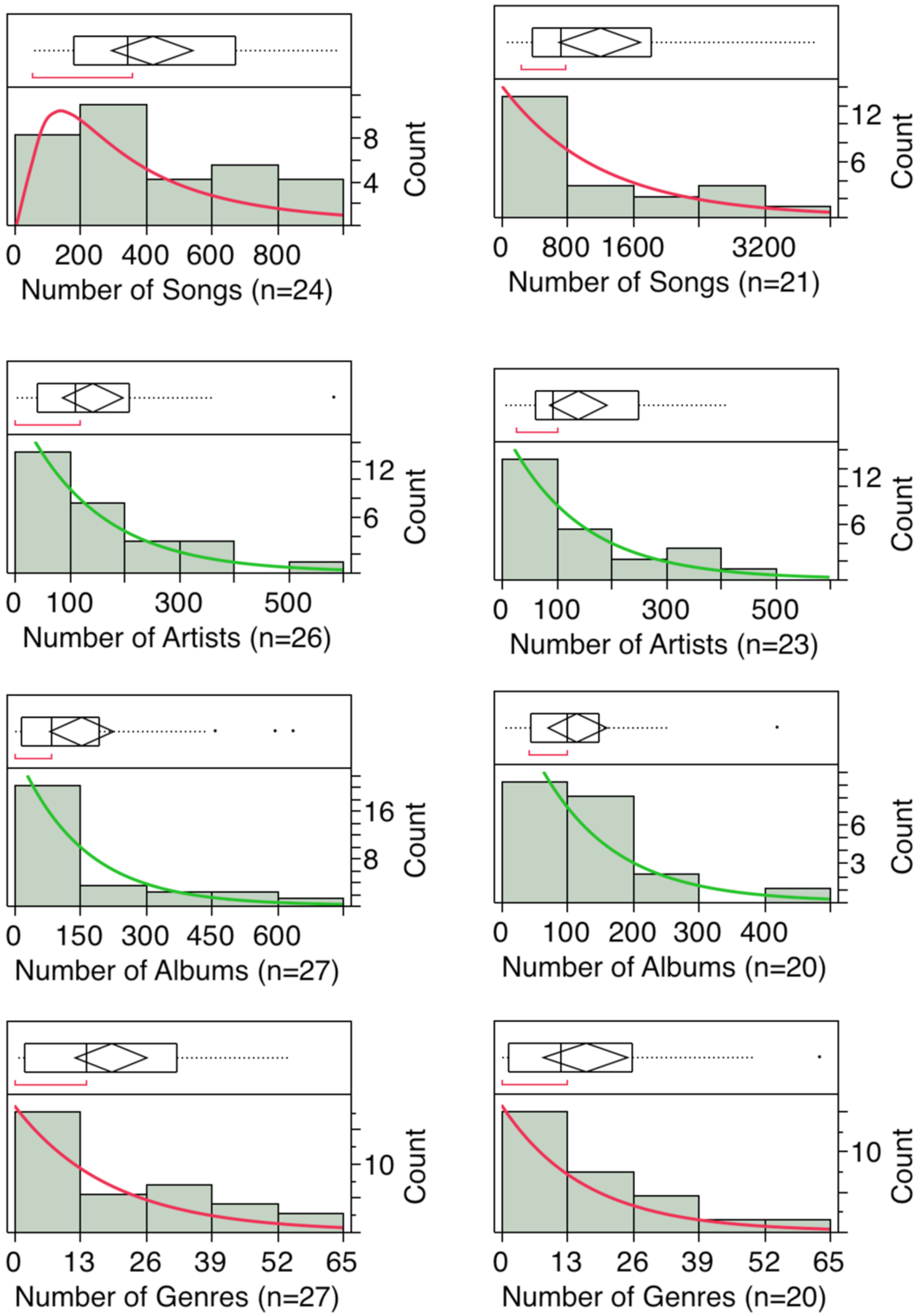
Table 3-5 shows the means, standard deviations, and ranges for the features, such as songs, artists, albums, playlists, and etc. on subjects' MP3 players. Distributions of songs, artists, and albums for both groups are shown in Figure 3-7. There were six outliers from typical drivers and two outliers from auto experts on the number of "Songs" which were not included in the calculation. The mean of number of songs were 420 for typical drivers and 1,200 for auto experts, statistically significantly different ($t_{(43)} = 3.37$, $p = 0.0016$). Features on videos and podcasts are relatively new, and the numbers for each category were small compared to music. Salvucci et al. conducted an experiment on iPod distraction [15]. In their study, the number of artists, podcasts, and videos were 99, 5, 5, respectively. Only the numbers of podcast in the MP3 player were similar to our results and the numbers of artists wert 40 less. The variation of numbers in each feature makes the comparison of results difficult.

Table 3-5. Mean Number of Features on Subject's MP 3 Players

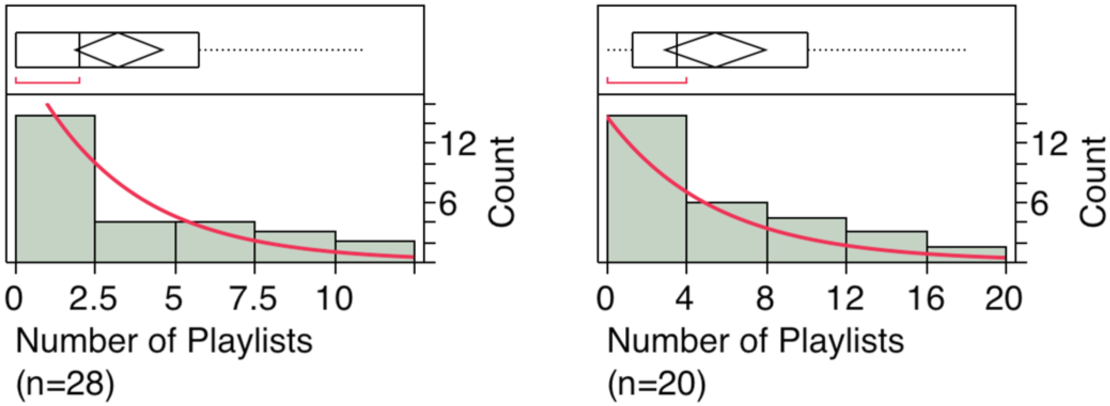
Category	Features	Typical Drivers		Auto experts	
		Mean (Range)	N	Mean (Range)	N
Music	Songs*	419 (52 - 982)	24	1193 (50 – 3800)	21
	Albums	151 (0 - 637)	27	114 (0 - 420)	20
	Artists	142 (0 - 584)	26	138 (0 - 413)	23
	Genres	19 (0 - 54)	27	17 (0 - 63)	20
	Composers	18 (0 - 138)	26	23 (0 - 136)	19
	Cover Flow	15 (0 - 200)	30	5 (0 - 100)	23
	Playlists	3 (0 - 11)	28	5 (0 - 18)	20
	Compilations*	1 (0 - 10)	25	4 (0 - 20)	21
	Audiobooks	0.6 (0 - 10)	30	1.4 (0 - 10)	23
Videos	Music Videos	0.6 (0 - 14)	30	1.3 (0 - 22)	23
	TV Shows	0.2 (0 - 4)	30	0.9 (0 - 10)	23
	Movies	0.1 (0 - 2)	30	0.7 (0 - 5)	23
	Video Playlists	0	30	0.1 (0 - 1)	23
	Rentals	0	30	0.04 (0 - 1)	23
Podcasts	Podcasts	4 (0 - 97)	30	9 (0 - 64)	23

*: Statistically significant difference, $p<0.05$

The distributions of songs that subjects owned for both groups followed lognormal distributions. The number of songs owned by the subjects for both groups in this survey is roughly 45 % fewer than the number from another study conducted by Lo. et al. [60]. However, the number of artists owned by typical drivers is two times more than the number from Lo, Walls, and Green's study [60]. There is no significant difference for the number of albums and genres owned by subject in both groups from this study and from the Lo, Walls, and Green's study [60].



0 200 400 600 800
Number of Songs (n=24)
4 8
Count
0 800 1600 3200
Number of Songs (n=21)
6 12
Count
0 100 300 500
Number of Artists (n=26)
6 12
Count
0 100 300 500
Number of Artists (n=23)
6 12
Count
0 150 300 450 600
Number of Albums (n=27)
8 16
Count
0 100 200 300 400
Number of Albums (n=20)
3 6
Count
0 13 26 39 52 65
Number of Genres (n=27)
10
Count
0 13 26 39 52 65
Number of Genres (n=20)
10
Count



(a) Typical Drivers (b) Auto Experts

Figure 3-7. Distribution of Number of Songs, Artists, and Albums that Subjects Own in Their MP3 Players from Both Groups (Without Outlier).

3.3.11 Reported Frequency of Using MP3 Players While Driving

Figure 3-8 and Table 3-6 show the average number of times per week for all features that subjects listened to while driving. On average, typical drivers listened to music about five times per week while driving and auto experts listened about seven times per week. Subjects seldom listened to audiobooks, podcasts, and watched videos while driving. The top two types of trips were “ Business” and “School” for typical drivers and “Business” and “Shopping” for auto experts.

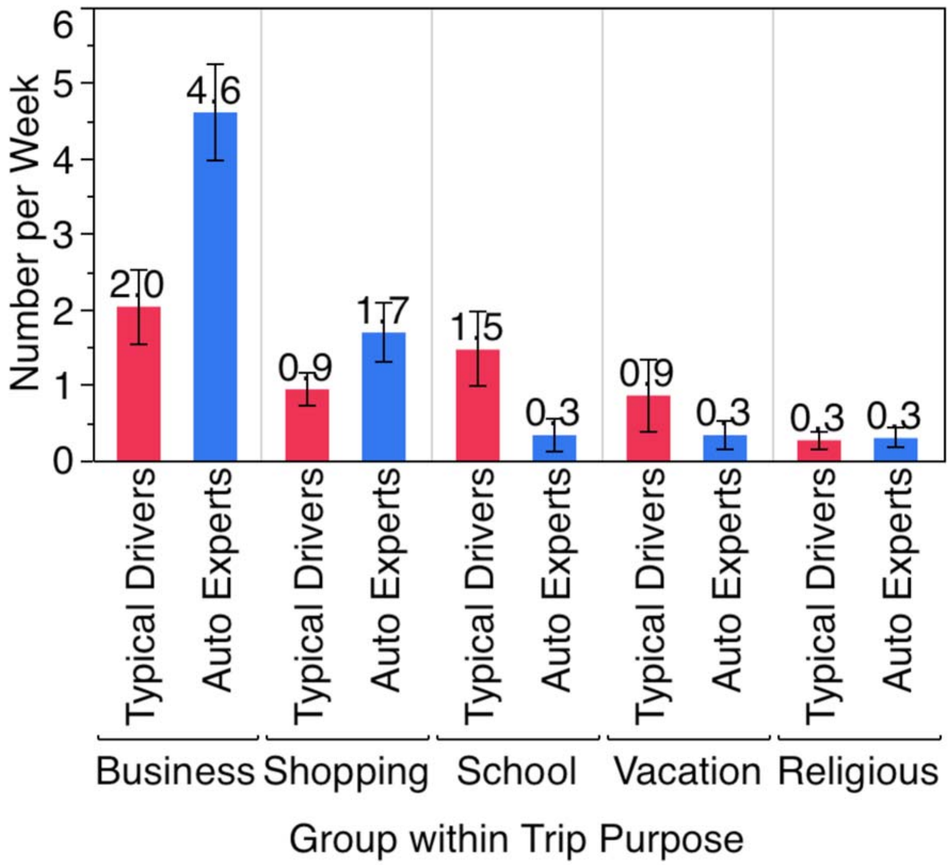


Figure 3-8. Number of Times per Week That Subjects Listened to Music While Driving

Table 3-6. Number per Week That Subjects Listened to Audiobooks, Videos, and Podcasts While Driving

Type of Trip	Audiobooks		Videos		Podcasts	
	Typical Drivers	Auto Experts	Typical Drivers	Auto Experts	Typical Drivers	Auto Experts
Vacation	0.5 ± 2.6	0.0 ± 0.2	0	0.0±0.01	0.1 ± 0.5	0.01 ± 0.04
Business	0	0.3 ± 0.9	0	0	0.1 ± 0.4	0.5 ± 1.1
School	0	0 ± 0.2	0	0.1 ± 0.6	0 ± 0.01	0.1 ± 0.6
Shopping	0	0	0	0	0.02 ± 0.09	0 ± 0.02
Religious	0	0	0	0	0	0
Total	0.5 ± 2.6	0.3 ± 0.9	0	0.1 ± 0.6	0.2 ± 0.6	0.5 ± 1.2

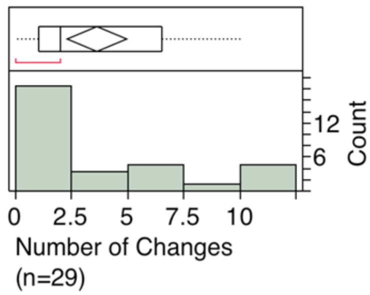
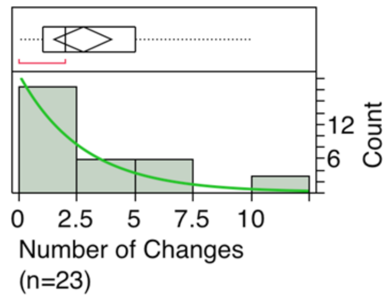
3.3.12 Interfaces Reported Used to Operate MP3 Players While Driving

Only one subject from the group of typical drivers operated his MP3 player using speech; others operated their MP3 players manually. However, it is surprising that most vehicles driven by auto experts did have a speech-controlled interface as standard equipment for music selection, but subjects did not use this function.

3.3.13 Reported Frequency of Changing the MP3 Player Features While Driving More Than 30 Minutes

When subjects listened to music while driving more than 30 minutes, data shows that subjects reportedly changed music 4 times per trip for typical drivers and 3.2 times for auto experts (Table 3-7). There is no statistically significant difference ($t_{(52)} = -0.80$, $p > 0.42$) between the two values. For other features (audiobook, video, and podcast), the means were small because these features are relatively new. Also, there was no significant difference between the number of changes from this survey and the number reported by Lo, Walls, and Green [60].

Table 3-7. Mean of Changes of Features on MP3 Players When Driving More Than 30 Minutes

Formats	Number of Changes: Mean ± S.D. (Range)		
	Typical Drivers	Auto Experts	Lo [60]
Music	3.6 ± 3.5 (0-10; 29) 	2.8 ± 2.9 (0-10) 	5
Audiobooks	0 (0)	0.1 ± 0.4 (0-2)	N/A
Videos	0 (0)	0.3 ± 1.2 (0-6)	N/A
Podcasts	0.1 ± 0.4 (0-2)	0.4 ± 1.1 (0-4)	N/A

3.3.14 Reported Frequency of Methods Used to Change the Music, Videos, or Podcasts

Figure 3-9 shows the results reported by the subjects on methods they used to search for other music. The three top-ranked methods were “Songs”, “Artists”, and “Playlists” for both typical drivers and auto experts. The only difference for these two groups was a reversal of the top two ranks. The results from this study were also the same as the survey and experimental results reported by Lo, et al. [60]. However, the results were not comparable with Salvucci et al. as they used “Artist” as the only method to search a specific song while assessing the iPod distraction [15].

When subjects want to watch movies, typical drivers searched using “Music Videos” and auto experts searched using “Music Videos” and “TV shows.” However, readers should keep in mind that watching movies is prohibited while driving, so the reported number was small.

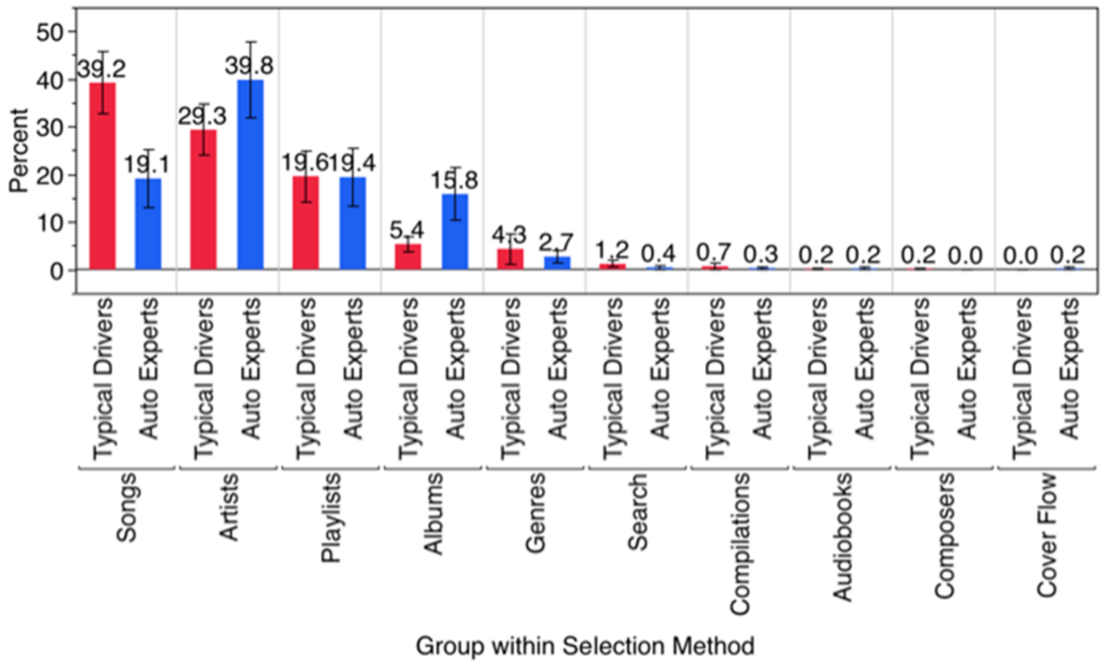


Figure 3-9. Mean Percent of Searching Methods for Music

Table 3-8. Mean Percent of Searching Methods for Videos

Method	Percent (Mean ± S.D.)	
	Typical Drivers	Auto Experts
Music Videos	6.7 ± 25.4	4.3 ± 20
Movies	1.7 ± 9.1	0
Rentals	1.7 ± 9.1	0
TV Shows	0	4.0 ± 19.4

3.3.15 Correction Strategies When Error Occurred Using Speech Interfaces

Only one typical driver reported that he used the speech interface to correct the errors while operating his MP3 player. Ninety-five percent of the time he repeated exactly the same words to correct the error and only five percent of the time he correct the error using manual selection. Other subjects from both groups did not use speech to control their MP3 players.

3.3.16 Frequency of Updating MP3 Players

Both typical drivers and auto experts usually connected their MP3 players once or twice per week to update music (Table 3-9).

Table 3-9. Number of Times Per Week to Update the Features on MP3 Players

Group	Number of Times Per Week to Update The Features (Mean ± S.D.)			
	Music	Audiobooks	Videos	Podcasts
Typical Drivers	0.9 ± 1.1	0.01 ± 0.05	0.00 ± 0.02	0.19 ± 0.57
Auto Experts	2.0 ± 4.5	0.04 ± 0.13	0.17 ± 0.50	0.62 ± 1.27

3.4 Conclusions

To design and evaluate in-vehicle navigation systems and travel information systems, one needs to know where real drivers typically go, a topic examined in this thesis. In fact, the authors do not know of any other current data on destination entry frequency or methods for contemporary navigation systems. The mean annual distance driven for subjects from both groups in this study were similar to the U.S. population, although students were overrepresented in the typical driver group. The comparison of the typical drivers to the general population likely means that there is a shift in the distribution of trip purposes from business to education, and given the comparatively lower income levels, vacation travel by car was more likely. Additional data would be needed to evaluate other major subsets of the U.S. population, such as retirees and others who are not auto experts.

Visiting “Friends’ Houses” (19%), “Home” (17%), “Shopping” (15%), and “Community” (13%) were the four top-ranked POI categories on the “Favorite” lists for typical drivers. “Home” (33%) was the most frequent POI category on the “Favorite” lists for auto experts. This also confirmed that “Home” was the most common POI category for subjects using the “Favorite” method for destination entry. The high frequency of visiting “Friends’ Houses” as a destination hints at the idea of linking social networking sites to navigation databases with the intent of reducing the effort to select a friend’s house as a destination, because the information would be more readily available.

The frequency data provide not only useful guidance for design, but data for assessment as well. For example, when checking compliance with AAM and SAE guidelines, the retrieval of destinations from guidance history and favorites lists should be assessed, and they need to be populated with subject specific data including their own home, local shopping, and addresses for their friends. In addition, these data emphasize the importance of POI lists, from which information is notoriously difficult to retrieve because of the uncertainty about which category contains the information desired by each driver.

The mean numbers of songs and albums on each subject’s MP3 player varied considerably. However, in studies of searching these devices, it is important to keep that number consistent across subjects, so the results from individual subjects can be compared. The most frequent methods used to search a specific song reported by subjects were by song title, artist name and playlists. Thus, it is important that designers attach priority to making these methods easy to use for music selection systems.

Good design of a user interface, be it for a navigation system in a vehicle, a travel information system on a desktop computer, or a system for any other purpose, requires data on who will use the system and the tasks to be accomplished by those users. In this case, the critical information is where people want to go and how they enter that information. The present investigation provides important data that addresses these questions.

CHAPTER 4

Destination Entry and Music Selection Using Speech in a Driving Simulator

4.1 Background

As was noted previously, time-consuming and error-prone in-vehicle tasks can distract drivers and degrade driving performance. This is particularly true for visual-manual interfaces, and is believed to be less of an issue for speech interfaces [5, 6, 8, 9, 11, 12]. For example, Tsimhoni et al. found that the task completion time decreased 82% for drivers entering an address using a word-based speech interface as opposed to using a touch-screen keyboard [12]. Also, Maciej and Vollrath found that speech interfaces on destination entry, phone dialing, and music selection tasks resulted in better driving performance, less gaze duration, and less distraction on subjective rating as opposed to manual input [8].

There are several studies that have specifically examined destination entry and music selection systems either on road or in simulator, the systems of interest here [5, 6, 8, 9, 11, 12, 15, 30]. In general, task performance, driving performance, driving behavior, and subjective rating on workload and preference are four major categories commonly used by researchers and summarized in [17, 28]. The results from these studies showed that driving performance was better and subject workload ratings were lower for destination tasks and music selection tasks when performed using speech interfaces than visual-manual interfaces [5, 6, 8, 9, 11, 12, 15, 30].

To assess the effects of operating navigation and music selection systems (secondary tasks) on driving performance, the ISO lane change test [44], vehicle following task [12], and peripheral detection tasks (PDT, [85]) are most commonly used. For example, Shutko et al. reported vehicle following tasks coupled with PDT were used to assess the driving performance, and the variables were lateral control and longitudinal control [11]. Also, Salvucci et al. used vehicle following tasks to assess iPod use with the variables of lateral deviation, and speed effect [15]. On the other hand, Maciej and Vollrath used lane change tasks to evaluate the use of speech interface vs. manual interface on driving performance [8]. To compare the results from these researchers is difficult due to lack of consistent standardized tasks and definitions of the variables used to assess the driving performance. There were some inconsistencies in the task completion times associated with the use of navigation and music selection tasks [17, 28].

To have a meaningful conversation, grounding, turn-taking [86], and *Conversational Maxims* [52] should be obeyed. Although using speech interfaces to find the address or play music is human-machine interaction, these human-human conversation principles can be applied to speech interactions. Utterances that violate the linguistics principles can result in errors. Errors can be made by either the machine or by subjects when subjects interact with the in-vehicle speech-controlled infotainment systems. Véronis [54] proposed a typology of errors for natural language dialogue systems, which also can be applied to partially explain the errors occurring in the interaction of human and command-based speech interfaces. Bourguet proposed a taxonomy of error-handling strategies for multi-modal human-computer interface [56]. From that taxonomy, lists of possible strategies that users will use are repeat, rephrase, spell out, etc.

Notice that missing from the literature are detailed analyses of driver use of speech interfaces, in particular, predictions of task times. Also missing from these studies is any examination of the types and frequencies of errors in performing in-vehicle tasks, something to be explored in depth here. That information is needed if speech interfaces are to be redesigned to reduce errors.

A final and major concern with the literature is the focus on evaluations of completed systems, rather than predictions of speech system performance before the system is fully developed. Those predictions could be made based on simulations. However, the construction of those simulations requires the understanding of interaction of drivers and a speech interface and detailed predictions of each step of driver use of an interface. The data necessary for that purpose was collected in this experiment.

The purpose of this experiment is to answer following the hypotheses:

1. *Hypothesis: There were no variables or distributions used to predict the drivers thinking and utterance time of any commands and phrases.* (How long do drivers need to think of and utter commands and phrases?)
2. *Hypothesis: There were no variables or distributions used to predict the speech interface's processing and prompt time of any commands and phrases.*

(How long does the speech interface need to process those commands or phrases by drivers and prompt the feedback?)

3. *Hypothesis: The error type of barge-in and time-out occurs most of the time.* (What are the types of errors that drivers make, and how often do they occur?)
4. *Hypothesis: Most of the time, subjects will re-try the same method.* (For each system response to an error, what is the driver's correction strategy?)
5. *Hypothesis: Driving workload will not affect the drivers' task performance* (How is performance of the speech task affected by the level of driving task workload?)

4.2 Method

4.2.1 Subjects

A total of 48 licensed drivers from southeast Michigan participated in this study. These subjects were equally distributed in each age (young (≤30): 23 ± 4; middle (40-50): 47 ± 3; and old (≥60): 68 ± 4) and gender (male and female) group. All subjects had normal or corrected-to-normal visual acuity and hearing.

4.2.2 Experiment Design

In this experiment, subjects performed destination entry tasks in three different conditions (while parked; while driving in low workload scenarios; and while driving in high workload scenarios). Each subject performed each destination entry task four times, and the order was counterbalanced across subjects. The two different driving workload scenarios were counterbalanced across subjects as well (Table 4-1).

Table 4-1. Counterbalanced Task Orders for Subjects in Each Age-Gender Group for Destination Entry Task

Subject 1		Subject 2		Subject 3		Subject 4	
Condition	Block - Trial	Condition	Block - Trial	Condition	Block - Trial	Condition	Block - Trial
Parked	A - 1	Parked	A - 2	Parked	A - 3	Parked	A - 4
Driving in low workload	B - 1	Driving in high workload	C - 2	Driving in low workload	B - 3	Driving in high workload	C - 4
Driving in high workload	C - 1	Driving in low workload	B - 2	Driving in high workload	C - 3	Driving in low workload	B - 4

4.2.3 Driving Simulator

The experiment was conducted in the third-generation University of Michigan Transportation Research Institute (UMTRI) driving simulator. This fixed-base simulator consists of a full-size cab, 11 computers, 6 video projectors (4 used in this study), 7 cameras (2 used in this study to record the driver's face and the central console), audio equipment, and other items. The simulator software (Vection and HyperDrive Authoring Suite, version 1.6.2) generated scene graphics, processed steering wheel, throttle, and brake inputs, provided steering wheel torque feedback, and saved the data. The raw data in the driving simulator was collected in 60 Hz, including subject's vehicle data (e.g. throttle and brake position, steering wheel angle, speed, longitudinal and lateral position) and traffic data (e.g. speed, position of other vehicles).

Figure 4-1 shows the partial simulator cab, forward scene, and front-right screen in this experiment. When using 6 projectors (5 forward channels and a rear channel), the simulator had a forward field of view of 200 degrees and a rear field of view of

40 degrees. In this study, with 4 projectors (3 forward channels and a rear channel), the forward field of view was 120 degrees. Each channel was 1024 x 768 and updated at 60 Hz. Depending on where the subject sat after adjusting the seat, the forward screen was 16 to 17 ft (4.9 to 5.2 m) from the driver's eyes.

Figure 4-1. Part of Simulator Cab, and Front Screen.

The simulator was controlled from an enclosure behind and to the left of the cab. The enclosure contained 4 quad-split video monitors that show the output of every camera and computer in the mockup. There was a display that shows the quad-split combination being recorded, as well as 3 sets of keyboards and LCD monitors for the driving simulator computers, and to control the instrument panel and warning and scenario control software (Figure 4-2). Also in the enclosure was a 19-inch rack containing audio and video equipment (audio mixers, video patch panel and switchers, distribution amplifiers, DVD recorder (used to record a quad split image of the driver), the quad splitter, etc.) and 2 separate racks for the instrument panel and touch-screen computers, the simulator host computers, and the 6 simulator image generators (4 used).

Figure 4-2. Simulator Operator's Workstation

The vehicle cab consisted of the A-to-B pillar section of a 1985 Chrysler Laser with a custom-made hood and back end mounted on casters for easy access. Mounted in the mockup were operating foot controls, a torque motor connected to the steering wheel (to provide steering force feedback), an LCD projector under the hood (to show the speedometer-tachometer cluster), a 10-speaker sound system (for auditory warnings), a haptic seat, a sub-bass sound system (to provide vertical vibration), and a 5-speaker surround system (to provide simulated background road noise). The 10-speaker sound system was from a 2002 Nissan Altima and was installed in the A-pillars, lower door panels, and behind each of the two front seats. The stock amplifier (from the 2002 Nissan Altima) drove the speakers. The speedometer-tachometer display was controlled by a

Macintosh computer running REALBasic and looked similar to those in an early 1990s Honda Accord.

The instrument panel and center console computers ran the Mac OS, the user interface to the simulator ran Windows, and the simulators ran Linux. Figure 4-3 shows a close-up of the cab interior. A unique feature of the simulator is the computer-generated, back-projected speedometer-tachometer cluster.

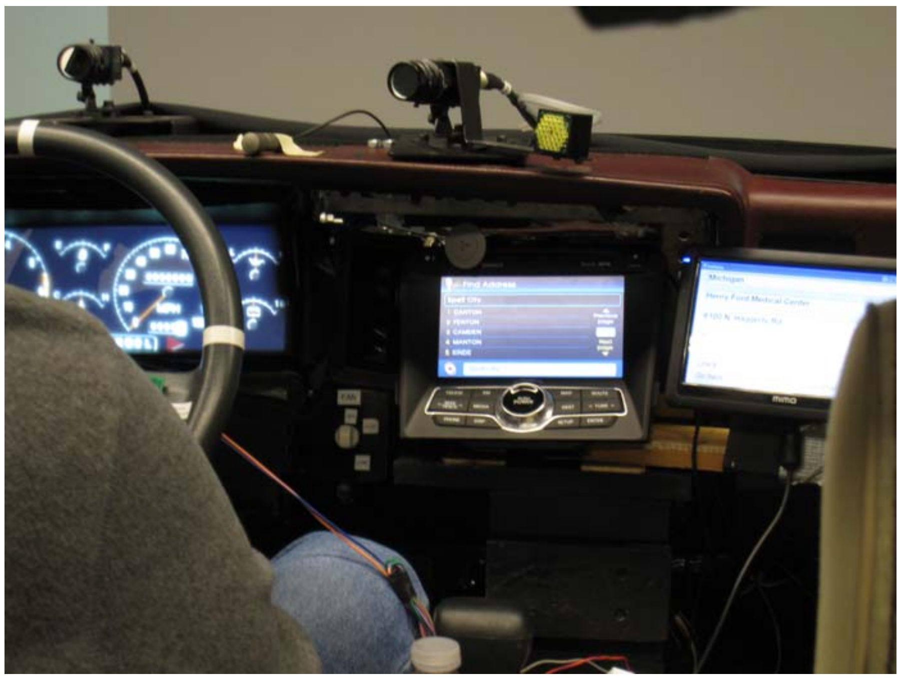


Figure 4-3. Navigation and Information Setting for the Experiment

4.2.4 Driving Scenarios

Subjects were asked to drive at 65 mph in the center lane. Passing the lead vehicle and changing lanes were not allowed. Otherwise, subjects were requested to drive normally. Two driving workload scenarios - low (workload = 2) and high (workload = 7.5) were determined using the equations from Schweitzer & Green (2007), described as follows. In those scenarios, the gap to the lead vehicle, the number of vehicles near subjects, and lead vehicle acceleration varied.

$$\text{Mean Workload Rating} = 8.86 - 3*(\text{LogMeanRange125}) + 0.47*(\text{MeanTrafficCount}) \quad (4.1)$$

Where

LogMeanRange125: Logarithm mean distance (m) to the same-lane lead vehicles over 30 s intervals. If no lead vehicle, mean distance = 125 m.

MeanTrafficCount: Mean number of vehicles detected (15 degree field of view) over 30 s intervals.

4.2.5 Navigation and MP3 Player

There were three systems that could potentially have been selected for this experiment. These systems are Ford SYNC, Nissan speech system, and Hyundai-Kia Genesis system, as implemented in mid-2012. All three systems for navigation support the commonly used methods to enter address, such as *Home*, *Street Address*, *Address Book*, *Previous Destination* and *Point of Interest (POI)*. However, the systems differ in many ways, and each system has its own unique structure, predetermined commands, and terms. For example, the number of pages that listed all commands used to enter the address was five for the Hyundai Genesis system, but only one for the Ford SYNC and Nissan system. The Nissan system uses *Places*, instead of *POI*, and the number of items listed on the screen is five per one page. The Hyundai Genesis lists all possible destinations from 1 to around 40 pages with 6 items per page. The Ford SYNC system normally requires the state and/or city information to be entered first, and then users choose the POI category. However, users can skip the step of city and state name and select the POI category first. The same procedures can be applied to street address entry for the Ford SYNC system. However, there is no such flexibility in the Nissan and Hyundai Genesis systems. The Nissan system even requires users to enter state name every time while using the *street address* method. For the history feature, the Nissan system again only provides five of the most recent records used by the user. The other two systems can provide as many records as had been entered until users delete them.

For music selection systems with speech interfaces, the Hyundai Genesis and Nissan system can only support simple control of a MP3 player, such as play all, play track (not a specific one), and play next/previous. On the other hand, the Ford SYNC systems can support more functions to control a driver's MP3 player. For example, users can select a specific song, artist, and album. After the author and his adviser spent three months attempting to borrow a portable in-vehicle speech interface without success, an alternative interface was selected for music selection tasks - an iPhone 4S without the Siri application. The system allows users to find a specific album and artist, but not a specific song. When users need to select a specific song, they can either select the *album* or *artist*. Users needed to say *next track* several times if there are more songs stored in the same album, or by the same artist until the specific song is played.

There were also difficulties in obtaining the most popular navigation systems with speech interfaces. Unfortunately, the manufacturers were not willing to provide them and in particular, in a portable form that could be installed in the UMTRI driving simulator. Fortunately, Mobis, the parts supplier for Hyundai, was willing to provide a system that could be temporarily installed in the simulator (Figure 4-3).

4.2.6 Destination Entry Tasks

As mentioned in the design session, subjects performed each destination entry task four times in each block (Table 4-2). Each trial included four pieces of information: destination name, street name and house number, city name, and state name (Figure 4-4). Among the four trials in each block, three trials were with correct information; one of them was related to subjects' personal information (residential address), and the others were selected by the experimenter based on the results from the survey study. In the incorrect information trial, an adjacent, incorrect city was given. First, subjects were requested to use the incorrect information to start the destination entry task even if he/she knew the city was incorrect. The goal of this particular trial was to collect the strategies that subjects used to correct errors. Also, there was one trial on each block with the information related to each subject (residential address). The purpose of this trial was to get the information that a subject would use to enter the destination, such as "Street Address," "Previous Destination," "Address Book," or "Destination Home." The order of the trials in each block was random.

Table 4-2. Destination Entry Task in Each Block.

	Block A	Block B	Block C
Destination address w/ correct information	Subject's Home	Subject's Friend's/Relative's Home	Subject's Friend's/Relative's Home
	Shopping Center/Mall	Shopping Center/Mall	Shopping Center/Mall
	Hospital	Church	Recreation
Destination address w/ incorrect information	Restaurant	University	Attraction

Figure 4-4. Information of Display for the Subject on Destination Entry Task

4.2.7 Music Selection Tasks

Similar to trials in the destination entry task, young and middle-aged subjects performed five trials in each condition - while the vehicle was parked, while driving in low-workload scenarios, and while driving in high-workload scenarios. Experimenters assumed that all subjects should be familiar with the songs in their own MP3 players, so information provided in these trials was always correct. Each block included three trials with complete information (song title, artist name, and album name), one trial for a specific album, and one trial for a specific artist (five trials total). The purpose of the three trials with complete information was to determine the method that each subject used to find a specific song. The five trials were randomized and then counterbalanced.

4.2.8 Performance Measures

To compare the driving performance of subjects during different workload scenarios and tasks conditions, eleven variables were examined. The brief definition of these variables is adapted from the SAE Recommended Practice J2944 [72] and shown in Table 4-3. Please see SAE Recommended Practice J2944 for the exact definition.

Table 4-3. The Brief Definition of Variables Used to Measure the Driving Performance (SAE Recommended Practice J2944 for the Exact Definition)

Performance measurement	**Brief Definition**
Task Completion Time (s)	The duration from when the experimenter finished saying "Next" and until the end of the prompt "Please proceed to the highlighted route and then the route guidance will start" prompted by machine.
Turns	The number of utterance sequences involving the machine, the user, or more commonly both, needed to exchange the information after the subject presses the voice-activation button. The "*total turns*" of a trial is the number of "*machine turns*" plus the number of "*user turns*."
Mean Speed	The mean speed averaged across the duration of a specific trial.
Speed Variation	The standard deviation of the speed from 65 mph across the duration of a specific trial.
Speed Difference	The maximum speed minus the minimum speed within the time period of a specific trial.
Maximum Speed	The largest speed within the time period of a specific trial.
Mean Time to Collision (TTC)	The mean time to collide with a lead vehicle in the travel path if the speed of vehicles were maintained with the gap at this moment.
Minimum Time to Collision	The minimum time to collide with a lead vehicle in the travel path if the speed of vehicles were maintained with the gap at this moment.
Mean Lateral Lane Position	The current vehicle position deviates from the center of lane.
Standard Deviation of Lateral Lane Position (SDLP)	The standard deviation of the current lane position deviates from the center of lane.
Mean Time-to-Lane-Crossing (TLC)	The time to reach the lane marking assuming the current heading angle and speed.
Minimum Time-to-Lane-Crossing (TLC)	The minimum time to reach the lane marking assuming the current heading angle and speed.

4.2.9 Procedure

The experiment was conducted in the UMTRI driving simulator. When recruited over the phone, subjects were told about the purpose of this experiment. Questions related to motion sickness were also asked to screen out susceptible subjects. Also obtained were three pieces of information concerning the subject's home address, and the home address of two friends or relatives. This information was entered into the navigation device in advance to save time. When subjects arrived at UMTRI, the experimenter greeted them and explained the purpose of this study again, as well as the procedure. Subjects then signed the consent form and completed a biographical form, which was used to collect some information about the subject, such as their age, years of driving, and driving habits. Vision and hearing tests were then administered to ensure that subjects could see and hear without any problems. If a subject had previous experience using a navigation device and/or MP3 player, questions related to each device were also asked.

The experimenter then gave a general overview of the driving simulator, pointing to the cameras that were used to record the performance during the experiment. The subject sat in the cab and adjusted the seat to the position that they felt comfortable, fastened the seatbelt, and also adjusted the mirrors. A practice driving session to check for motion discomfort and achieve stable driving performance lasted about five minutes. Next, the subject practiced using the navigation device with the speech interface. To begin, the experimenter described the structure of the navigation unit and the commands to be used later. Then the subject entered destinations using the following methods in this order – "Street Address," Point of Interest (POI)," Address Book" (created by the experimenter in advance), and "Previous Destinations." There were two practice trials for each method.

For the last two practice trials, the given state name was incorrect. This was to present subjects with a sense of what would happen when they entered incorrect information.

After 10 practice trials, when the subject was ready, the experiment started with destination entry tasks while the vehicle was parked. Subsequently, the subject performed the destination entry tasks while driving in low- and high-workload scenarios in a counterbalanced order. For the elderly subjects, a post-test questionnaire about their destination histories was completed. Because it took more than two hours for elderly subjects to complete the destination entry task, the music selection tasks were not included; it would have made the experiment excessively long for this group. Furthermore, most of the elderly subjects did not own and use MP3 players, so they would have had to learn to use one, an activity for which there was no time. For young and middle-age subjects, music selection tasks were included in the experiment, and the task orders were similar to the destination entry tasks – practice block, select music while vehicle parked, select music while driving in low-workload scenarios, and select music while driving in high-workload scenarios. See Table 4-4.

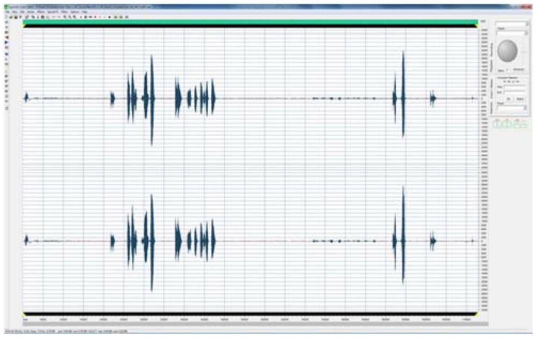
Table 4-4. Time Allocation of the Driving Simulator Experiment

#	Tasks	Comment	Duration (min)
1	Introduction	Greet subjects, sign consent, biographical forms, complete vision & hearing tests, answer questions about previous experience using navigations systems and MP3 players, walk to the simulator, and introduce the simulator	30
2	Practice Driving	Practice driving for five minutes to get used to the driving simulator and screen out those subjects who might be susceptible to motion sickness	5
3	Practice trials for navigation	10 tasks for subjects to get used to the interfaces and process	30-45
4	Block1 (while vehicle parked)	Test block. Four tasks for each block	10-15
5	Block2 (while driving in low workload scenarios)	Test block. The experimenter needed one minute to load the test scenario and activate the simulator. Wait one more minutes after subject reached the desired speed (65 mph), to begin collecting baseline-driving data. There were four tasks for each block.	10-15
6	Block3 (while driving in high workload scenarios)	Test block. The experimenter needed one minute to load and activate the simulator. They waited one more minute after subject reached the desired speed (65 mph), to collect baseline-driving data. There were four tasks per block.	10-15
7	Practice trials for music selection using iPhone 4	There were six tasks for subjects to get used to the interfaces and process	10
8	Block1 (while vehicle parked)	Test block. There were five tasks for each block	5
9	Block2 (while driving in low workload scenarios)	Test block. The experimenter needed one minute to load and activate the simulator. They waited one more minute after subject reached the desired speed (65 mph) to begin collecting baseline driving data. There were five tasks for each block.	5
10	Block3 (while driving in high workload scenarios)	Test block. The experimenter needed one minute to load and activate the simulator. They waited one more minute after subject reached the desired speed (65 mph), to begin collecting baseline driving data. There were	5

		five tasks for each block.	
11	Post test	The experimenter asked some questions concerning the subjects' history of visiting the destination just entered.	5
		Total	130-150

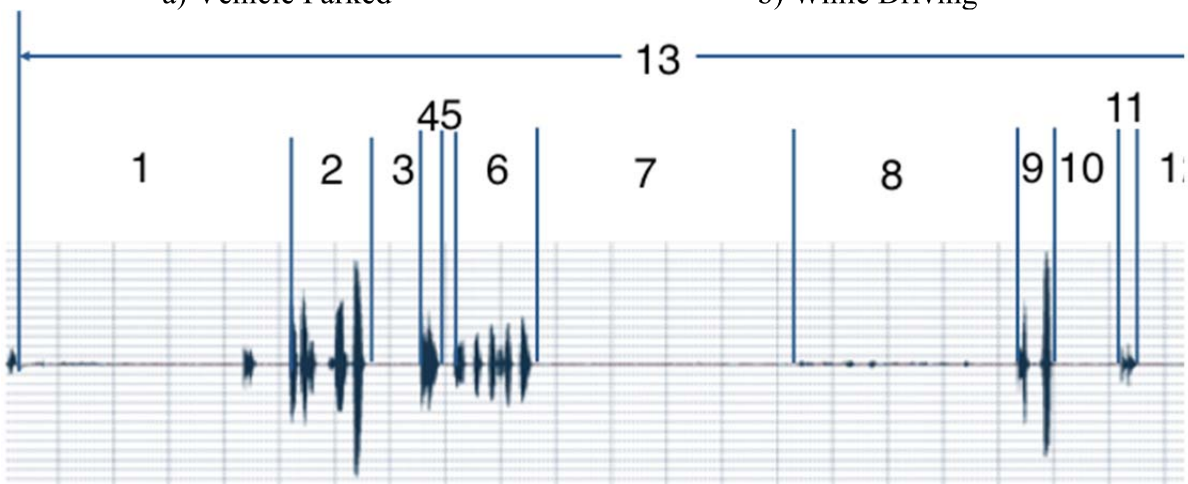
4.2.10 Data Reduction

All prompts for subjects and machine interaction for both the destination entry and music selection tasks were recorded and analyzed using the ExpStudio Audio Editor software. Examples of audio files and definitions are shown in Figure 4-5 and Table 4-5. All the prompts and time stamps were transcribed and corrected by a team of seven research assistants. This took many, many months, and all of their data was double checked, and in some cases, triple checked for accuracy.



a) Vehicle Parked

b) While Driving



c) Detailed Sound Wave

Figure 4-5. Audio Example of the Interaction Between Subject and the Speech Interface: Utterances for the Music Selection Task and an Explanation of the Segments Examined.

Table 4-5. Example of the Utterances for Music Selection Task and Associated Time Intervals

Area of Interest	Description
Example	Experimenter: *Next* Subject: ***Play Artist*** *+Sean Paul* iPhone: "Ding Ding" "***Playing Songs by*** + *Sean Paul*", music playing Subject: *Next Track* iPhone: "Ding Ding", music playing
1	Thinking and Response Time: The duration that subjects used to think about the query before saying it. It was the elapsed time between the end of the experiment saying "next" and the beginning of the utterance of the word.
2	Duration of the subject's utterance, including command words, pause, and the specific information
3	Duration of the pause after the subject's utterance and the beginning of chime from machine.

4	Duration of the sound from machine.
5	Duration of the pause after the chime from machine and the prompt of feedback from machine.
6	Duration of the prompt from machine's feedback.
7	Duration of the machine processing time to find and play the specific song.
8	Thinking and Response Time_2: the duration that the subject needed to determine the music played was correct or not.
9	Duration of the utterance for the subject to say the command words *Next Track*.
10	Duration of the pause after the subject's utterance and the beginning of the chime from machine.
11	Duration of the chime from the machine
12	Duration of the machine processing time to find and play next song from the same album or by the same artist.
13	Task Completion Time: Duration of the entire task from 1 to 12.

4.3 Results and Discussions

4.3.1 Destination Entry Task

A. Task Completion Rate, Task Completion Time, and Detailed Time Associated with Each Utterance

Task Completion Percentages

There were 576 trials performed by the 48 subjects in the destination entry task. Eight trials were incomplete (subjects balked), and the resulting task completion percentage was therefore 98.6% (Table 4-6). Among the 568 completed trials, there were 12 trials (2.1%; 12/568) that the subjects finished, but the location was incorrect, though they thought the final result was correct. Therefore, there were 556 trials that ended with the correct destination entered. Furthermore, 52% (300/576) of the trials were completed without any errors, neither from subjects nor from the speech interface. When the given information was correct, the task completion percentage (destination entry was correct at the end of the trial) was greater than the task completion rate when the information given was incorrect, 99 % (427/432) vs. 90% (130/144).

Table 4-6. Frequency of Task Completion for Destination Entry Task

Information	Complete / Give-up	Incorrect Final		Correct Final	
		Incorrect First	Correct First	Incorrect First	Correct First
Incorrect information	Complete	8	0	81	49
	Give-up	6	1	0	0
Correct information	Complete	4	0	176	251
	Give-up	1	0	0	0

However, there was still one trial, after four tries, in which the subject gave up, even though the information given was correct. She gave up because she said the prompt either too early (barge-in) or too late (time out) while driving. Therefore, the machine could not pick up the correct input and either provided incorrect feedback, or repeatedly asked for the desired information. For the trial in which the information given was incorrect, all six trials in which the subjects gave up occurred while driving, either in the low (4 trials) or high (2 trials) workload condition (Table 4-7). When the information given was correct, four trials were completed with an incorrect destination entered. For three of the four trials, subjects correctly entered the street name, but failed to enter the house number and start guidance. Two of the errors occurred because the subjects provided information in an invalid format. (Subjects said four thousand to enter the house number, but the system could only understand "four zero zero zero.")

Table 4-7. Frequency of Task Completion for Destination Entry Task by Scenarios

Driving Condition	Scenario	Task Completion	Incorrect Final		Correct Final	
			Incorrect First	Correct First	Incorrect First	Correct First
Parked	1* Recreation	Complete	4	0	23	20
		Give up	1	0	0	0
	2 Shopping	Complete	2	0	19	27
		Give up	0	0	0	0
	3 Home	Complete	0	0	24	24
		Give up	0	0	0	0
	4 Hospital	Complete	1	0	21	26
		Give up	0	0	0	0
Low Workload	5 Home	Complete	0	0	15	33
		Give up	0	0	0	0
	6* University	Complete	3	0	26	15
		Give up	3	1	0	0
	7 Shopping	Complete	0	0	22	25
		Give up	1	0	0	0
	8 Church	Complete	0	0	15	33
		Give up	0	0	0	0
High Workload	9 Recreation	Complete	0	0	16	32
		Give up	0	0	0	0
	10* Attraction	Complete	0	0	32	14
		Give up	2	0	0	0
	11 Home	Complete	0	0	20	28
		Give up	0	0	0	0
	12 Shopping	Complete	1	0	24	23
		Give up	0	0	0	0

*: Given information were incorrect

Task Completion Time

The mean task completion time of the trials that the subjects completed for the destination entry task was 123.47 ± 83.78 s. Of those trials with incorrect information in which the subjects gave up, the mean time was 322.64 ± 28.53 s (Table 4-8). For those trials that the subjects entered the final destination correctly, the mean task completion time when given incorrect information was almost double the task completion time when given correct information, for both correct and incorrect entries on the first attempt.

Furthermore, the mean task completion time for those trials without any errors was 76.65 ± 15.38 s, which was almost five times larger than the SAE 15-second rule, which is intended for visual-manual interfaces [80]. Although the interface tested was a speech interface, there were instances where subjects need to refer to a display screen to make a selection, so the interface test was not a purely speech interface. When errors occurred during the first attempt, the mean task completion time was almost twice as much (135 s vs. 76.7 s) compared with the time of trials without any errors. When the information given was correct and no errors occurred, age and driving conditions significantly affected the task completion time ($F_{(2, 232)} = 7.63$, $p = 0.001$ for age and $F_{(2, 232)} = 6.512$, $p=0.002$). The interaction of age and gender also affected the task completion time ($F_{(2, 232)} = 5.473$, $p = 0.005$). Post hoc test results revealed that the task completion time for young drivers was 7 s and 8 s less than the time for middle-aged and elderly subjects ($p<0.001$ and 0.001 with the Bonferroni adjustment). The task completion time while the vehicle was parked (81.5 s) was 6 s and 7 s longer than the task completion time while driving in low (75.5 s) and high (74.6 s) workload conditions ($p < 0.001$ and 0.001 with Bonferroni adjustment). There was no difference in the task completion time between the two driving workload conditions. Learning effects may explain the difference as subjects always performed the destination entry tasks while the vehicle was parked and the two driving conditions were counterbalanced.

Table 4-8. Mean and Standard Deviation of Task Completion Time for Destination Entry Task (in seconds)

Information	Complete / Give-up	Incorrect Final		Correct Final	
		Incorrect First	Correct First	Incorrect First	Correct First
Incorrect information	Complete	138.85 ± 70.07	0	233.27 ± 111.95	132.90 ± 24.88
	Give-up	322.64 ±	493.23		

		128.53			
Correct information	Complete	183.80 ± 131.53		135.00 ± 84.12	76.65 ± 15.38
	Give-up	195.52			

Table 4-9 shows the mean task completion time by scenarios for the various driving conditions.

Table 4-9. Mean Task Completion Time by Scenarios

Driving Condition	Scenario	Task Completion	Incorrect Final: Incorrect First	Incorrect Final: Correct First	Correct Final: Incorrect First	Correct Final: Correct First
Parked	1*	Complete	148.52		277.31	145.6
	Recreation	Give up				
	2	Complete	91.82		109.00	81.66
	Shopping	Give up				
	3	Complete			149.76	82.89
	Home	Give up				
	4	Complete	180.77		136.50	78.96
	Hospital	Give up				
Low Workload	5	Complete			114.29	68.51
	Home	Give up				
	6 *	Complete	122.74		225.73	132.90
	University	Give up	347.77	493.23		
	7	Complete			153.55	82.70
	Shopping	Give up	195.52			
	8	Complete			112.95	79.03
	Church	Give up				
High Workload	9	Complete			120.9	76.02
	Recreation	Give up				
	10*	Complete			207.73	114.75
	Attraction	Give up	284.93			
	11	Complete			138.58	65.32
	Home	Give up				
	12	Complete	370.79		155.69	78.06
	Shopping	Give up				

*: Given information were incorrect

Thinking and Response Time

The sum of these two times is the duration from the end of the "next" from the experiment's utterance to the beginning of machine's prompt. Thinking and response times were significantly affected by age and gender. The mean thinking and response time for elderly drivers was 1.9 s, which was significantly longer than the thinking time for young and middle-age drivers ($F_{(2, 522)} = 18.733$, $p < 0.001$). There also was a main effect of gender on the thinking and response time, $F_{(1, 445)} = 11.477$, $p < 0.001$. Men took 1 s longer to think of and respond to the information than women. No significant difference was found due to the main effect of driving conditions, correctness of given information, or information relevant to the subject.

To provide the data needed to build the simulation, half of the data was used to generate the distribution and its parameters. When subjects needed to enter a residential address, the thinking and response times were fitted by a lognormal distribution with the parameters of μ and σ equal to 1.48 s and 0.57 s ($n = 72$). On the other hand, the thinking and response times were fitted by a lognormal distribution with the parameters of μ and σ equal to 1.31 s and 0.61 s ($n = 214$), when the address that subjects entered was not a residential address.

Time to Utter the Commands by Subjects

Subjects needed to utter a variety of commands to advance to the next step. For example, to enter the street address, subjects need to say the command *Find Address*. To retrieve the saved information in the address book, subjects needed to say the command *Destination by Address Book*. When subjects thought they were ready to find the requested destination, they said *Start Guidance*. The time to utter these commands was significantly affected by the age ($F_{(2, 1061)} = 12.020$, $p < 0.001$). Post hoc tests revealed the mean time for elderly subjects to utter a command takes 0.12 s and 0.09 s greater than the time required by young and middle-age subjects ($p < 0.001$ and 0.008 with Bonferroni adjustment).

Using a stepwise regression, the time required to say a command can be predicted as follows ($R^2 = 0.374$, $F_{(3, 1075)} = 214.283$, $p < 0.001$).

$$\text{CMD Utterance Time} = 0.136 + 0.133 * \text{NSyllable} + 0.082 * \text{Age} + 0.094 * \text{NWord} \quad (4.2)$$

Where:

NSyllable: Numbers of syllables of the commands that subjects uttered.

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1. Elderly subjects = 2.

NWord: Number of words of the commands that subjects uttered.

Time to Think of and Utter the State Name by Subjects

The default setting for the state (e.g., Michigan) for the current destination is the state from the previous destination. If the state name is the same as previously used, drivers did not need to say the command – *Change State*.

On the other hand, subjects needed to think of and then provide state information if the information provided by the experimenter was different from the default state setting. The time to think of the state name was not affected by the variables of age, gender, driving conditions, and information relevant to the residential address, or their interactions. Using stepwise regression, the time to think of a state name can be predicted by age. However, the value of R-square was 0.078 ($F_{(1, 64)} = 5.356$, $p = 0.024$), which was too low to be meaningful. The time to think of the state name while driving was fit by a lognormal distribution with the parameters of μ and σ equal to -0.5 and 0.49 ($n = 35$, $p = 0.139$).

Age and gender significantly affected the time for subjects to utter the state name, $F_{(2, 62)} = 7.310$, $p = 0.002$ and $F_{(1, 62)} = 11.574$, $p = 0.001$, respectively . Using a stepwise regression, utterance times to enter the state name can be predicted using the equation shown below ($R^2 = 0.405$, $F_{(3, 58)} = 13.161$, $p < 0.001$).

$$\text{State Utterance Time} = -0.212 + 0.242 * \text{NSyllable} + 0.098 * \text{Age} - 0.104 * \text{Gender} \tag{4.3}$$

Where:

NSyllable: Numbers of syllables of the state name that subjects uttered.

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1. Elderly subjects = 2.

Gender: The gender group of subjects. Female = 0. Male = 1.

The mean time for middle-aged and elderly subjects was 0.1 s and 0.2 s greater to utter the state name than the time for the young drivers. Also the time for female drivers to utter the state name is 0.1 s longer than the time needed by male drivers.

Time to Think of and Utter the City Name by Subjects

Usually, the first information that subjects needed to provide was city name when subjects used the street address method, in which case the state was same as the default state. Subjects again needed to think of the city name before uttering it when the machine asked for the city information. Neither the age, gender, driving workload, information relevant to residential address nor their interactions, affected the time for subjects to think of the city name. The time for subjects to think of the city name was fit by a normal distribution, with a mean and standard deviation of 0.91 s and 0.78 s, respectively ($n = 351$).

The time to utter the city name was significantly affected by the age ($F_{(2, 302)} = 4.718$, $p = 0.01$). A post hoc test revealed that the time to utter the city name by elderly subjects was 0.17 s and 0.14 s longer than the time needed by the young and middle-age drivers, respectively.

Using a stepwise regression, utterance times to enter the city name can be predicted as follows ($R^2 = 0.432$, $F_{(3, 301)} = 75.544$, $p < 0.001$).

$$\text{City Utterance Time} = -0.103 + 0.172 * \text{NSyllable} + 0.089 * \text{Age} + 0.149 * \text{NWord} \tag{4.4}$$

Where:

NSyllable: Numbers of syllables of the city name that subjects uttered.

Age: The age group of subjects. Young subjects = 0. Middle-aged subjects = 1. Elderly subjects = 2.

NWord: Number of words of the city name that subjects uttered.

Time to Think of and Utter the Street Name by Subjects

Even though the experimenter emphasized during the practice and before the experimental trials that the subjects needed to utter all the information provided by the experimenter, 14.8% of the time (42/284) subjects only uttered the street name without saying the suffix (“road,” “street,” “avenue,” and etc.), which resulted in the incorrect feedback from the system or multiple feedback messages. Thus, subjects needed to perform an extra step to select the correct destination. Also, 12.9% of the time (21/163) subjects did not say the direction (east, west, south, and north), which also resulted in the incorrect feedback or multiple feedback messages that led subjects to perform an extra step to select the correct information. Both of these two cases can result in greater times to complete the task.

The time to think of the street name by the subjects was affected by age ($F_{(2, 291)} = 4.609$, $p = 0.011$) and information relevant to personal address ($F_{(1, 291)} = 3.788$, $p = 0.41$). Post Hoc tests revealed that the mean time to think of the street name for young subjects was 0.26 s and 0.29 s less than the time for middle-aged and elderly drivers. When subjects entered the street name of a friends’ or relative’s address (residential address), the time to think of the street name was 0.2 s less than the time when subjects tried to enter the non-residential address. Age can be used to predict the time to think of the street name using stepwise regression ($R^2 = 0.025$, $F_{(1, 329)} = 8.431$, $p = 0.004$). Again, the R-square was small. The time to think of the street name while driving was fit by a lognormal distribution with the parameters of μ and σ equal to 0.03 and 0.58 ($n = 260$, $p = 0.150$).

The time to utter the street name was significantly affected by the age and the residential address, $F_{(2, 284)} = 3.788$, $p = 0.024$ and $F_{(1, 284)} = 11.56$, $p = 0.001$, respectively. When subjects uttered the street name of a friend or relative, it took 0.3 s less than the time needed to utter the non-residential address. A post hoc test revealed that the time to utter the street name by elderly subjects was 0.27 s and 0.19 s greater than the time needed by the young and middle-age drivers, respectively.

Using stepwise regression, the utterance times for street names can be predicted using the following equation ($R^2 = 0.601$, $F_{(3, 283)} = 140.298$, $p < 0.001$). Again, the time to utter the street name increased as age increased and women took more time to say street names than men.

$$\text{Street Utterance Time} = 0.041 + 0.459 * \text{NWord} + 0.188 * \text{Age} - 0.103 * \text{Gender} \quad (4.5)$$

Where:

NWord: Number of words of the street name that subjects uttered.

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1. Elderly subjects = 2.

Gender: The gender group of subjects. Female = 0. Male = 1.

Time to Think of and Utter the House Number by Subjects

There were many ways in which subjects chose to utter house numbers. For example, subjects can say “four zero zero zero” or “four thousand” when the number is 4000. Subjects may say “seventeen seventeen,” instead of “one seven one seven” when the number is 1717. Also, subjects could say “seven thirty,” instead of “seven three zero” when the number was 730. However, the system could only recognize the utterance when subjects said the number one digit at a time. Thus, the system did not recognize “four thousand," "seven thirty," or "seventeen eighteen," which are commonly spoken combinations. This occurred in 8 % of trials during the experiment.

The time to think of a house number was affected by the main effect of age ($F_{(2, 227)}$ = 4.508, p = 0.012) and whether the information given was correct ($F_{(1, 227)}$ = 7.353, p = 0.007). The time needed by young subjects to think of the house number was 0.4 s less than the time needed by elderly subjects, based on a post hoc test with a Bonferroni adjustment. There were no statistically significant differences between the times for elderly and middle-age subjects or young and middle-age subjects. Although the house numbers given to subjects were always correct, thinking about whether the information given was correct affected the time to think of the house number. When given an incorrect city name, the time to think of the house number was 0.3 s longer than when the information given was correct. Using a stepwise regression, information correctness and age can be used to predict the time to think of the house number (R^2 = 0.087, $F_{(2, 270)}$ = 12.908, $p < 0.001$). Using the distribution fitting method, the inverse of thinking time for their house number (1/T) fit a Weibull distribution (n=271, p = 0.25) with parameters of α and β equal to 1.31 and 2.42.

The time to utter the house number was significantly affected by the age and information relevant to the private address, $F_{(2, 240)}$ = 8.095, $p < 0.000$ and $F_{(1, 284)}$ = 11.272, p = 0.001, respectively. When subjects uttered the residential (friends' or relative's) address, it took 0.36 s less than the time needed to utter the non-residential address. A post hoc test revealed that the time to utter the street name by elderly subjects was 0.46 s and 0.39 s longer than the time needed by the young and middle-age drivers, respectively.

Using a stepwise regression, utterance times to enter the house number can be predicted using the following equation ($R^2 = 0.527$, $F_{(4, 239)} = 65.486$, $p < 0.001$).

$$\text{House Number Utterance Time} = -0.805 + 0.681 * \text{NWord} + 0.201 * \text{Age} - 0.143 * \text{Workload} - 0.197 * \text{Gender} \quad (4.6)$$

Where:

NWord: Number of words of the house number that subjects uttered.

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1. Elderly subjects = 2.

Workload: Driving workload. Low workload = 0. High workload = 1.

Gender: The gender group of subjects. Female = 0. Male = 1.

Again, the time to utter the house number increased as a subject's age increased, and women took longer to say the street name than men. However, subjects spoke more quickly when driving than when not driving.

Time to Spell the State, City, and Street Name by Subjects

During the process of entering the destination, subjects might spell the state, city, or state name as an alternative method to providing the necessary information or to correct the information when the previously provided information was incorrect. When the subjects spelled the word(s), the mean number of characters was 7.9 ± 2.3 characters, corresponding to one (73.2%) or two (26.8%) words. Fitting the data discretely, the number of characters that subjects spelled followed Poisson distribution with parameter (λ) of 7.91 ($n = 56$, $p = 0.76$).

When subject spelled the state, city or street name, the utterance time was not significantly affected by age, gender, driving workload, or information relevant to the residential address. Results from stepwise regression revealed that the number of characters and the number of the words predict the utterance time ($R^2 = 0.704$, $F_{(2, 55)} = 62.973$, $p < 0.001$).

$$\text{Spelling Time} = -1.826 + 0.635 * \text{NCharacter} + 1.09 * \text{NWord} \quad (4.7)$$

Where:

NCharacter: Number of characters that subject spelled.

NWord: Number of words that subjects spelled.

Time for Barge-In by subjects

When subjects provided any information or uttered any commands before the speech interface signal (usually a beep), or without pressing the ASR button to interrupt the system, this is defined as User Barge-in. Most in-vehicle speech systems cannot recognize and process barge-in utterances, and therefore, present incorrect feedback as the system captures either none or some of the utterance. There were 53 occurrences of barge-in in this experiment, on average just over two per subject. The barge-in time was not significantly affected by age, gender, driving workload, and information relevant to personal related address. In fact, there were no variables that predicted the time of subject barge-ins. That barge-in time was normally distributed with the parameters of μ and σ equaling to 0.19 s and 0.11 s ($n = 53$, $p = 0.2291$).

Time for Time-Out by Subjects

When the system asks for information, subjects need to utter the information within a specific time, the time-out duration. If subjects did not say anything within this period, the system either repeats the previous prompt to ask for information, or deactivates the ASR function. Time-out also occurred when subjects start to provide the information within the normal time window, but continue beyond the time window, and the system cannot capture the entire user utterance. There were 76 time-outs in this experiment. The time-out duration was not affected by age, gender, driving workload, or information relevant to residential address. The time out duration was normally distributed with the parameters of μ and σ equal to 7.12 s and 1.8 s ($n = 78$, $p = 0.0829$).

Time for Various Prompts Uttered by the Machine

Unlike the speech system for music selection, the prompts by the speech interface and subjects are much more complex for the destination entry tasks. However, the time needed for the machine to say the same prompt should be always the same, but the duration varied with the content, depending upon the state, city, street name, and house number combinations. Most of the prompts that the speech interface uttered are shown in Table 4-10.

Table 4-10. The Prompts and Their Corresponding Duration by the Machine

Prompts	Time	Pause	Beep	Total Duration
Command Please	0.922	0.315	0.08	1.317
You can say, for example, destination help or say help at anytime	4.094	0.342	0.08	4.516
Destination help. Command Please	1.97	0.346	0.08	2.396
For example, say find nearest POI or say help at any time	3.867	0.351	0.08	4.298
Find Address. The city please	2.587	0.336	0.08	3.003
Please enter the state name	1.277	0.332	0.08	1.689
Sorry. Please enter the state name	2.442	0.327	0.08	2.849

Please select the respective line or start spelling	2.803	0.359	0.08	3.242
Please spell the name	1.05	0.323	0.08	1.453
Spell the name in blocks of letters or enter your destination again with change state or enter city	5.133	0.329	0.08	5.542
The city please	0.851	0.332	0.08	1.263
Sorry. The city please	1.996	0.344	0.08	2.420
What is the street	0.869	0.317	0.08	1.266
Sorry. What is the street	2.024	0.313	0.08	2.417
And house number or if you don't know that, please say show map or start guidance	4.798	0.304	0.08	5.182
Sorry. Your choice is not available at this point.	2.465	0.836	0.08	3.381
Show map or start guidance	1.492	0.338	0.08	1.910
Say show map, start guidance or say help at any time	3.298	0.333	0.08	3.711
Show Map	0.528			
Start Guidance	0.931	0.342	0.142	1.415
Click	0.142			
Please proceed to the highlighted route and then the route guidance will start	3.996			
Select previous destination. Line please.	2.598	0.322	0.08	3.000
For example say line two, next page, help, repeat, or back	4.152	0.328	0.08	4.560
Go Home	0.492	0.358	0.08	0.930
Please select a POI category	1.899	0.317	0.08	2.296
Please Select a POI sub-category.	2.192	0.302	0.08	2.574
Line Please	0.714	0.327	0.08	1.121
Please select user name	1.348	0.306	0.08	1.734
For example, say a user name like user one or next page, help or back	4.97	0.361	0.08	5.411
User one. Line please.	2.375	0.329	0.08	2.784
There is no database	1.301			
Please say Line and the line number. For example Line two	3.053	0.314	0.08	3.447
Sorry. I can't understand the command. Please say again or say help.	3.448	0.359	0.08	3.887

Note: The total duration = prompt time + pause time + beep time, as appropriate.

Using a stepwise regression, machine prompt times to say a command or provide information can be predicted using the following equation ($R^2 = 0.901$, $F_{(1, 32)} = 280.719$, $p < 0.001$).

$$\text{M-CMD Time} = 0.347 + 0.284 * \text{NWord} \tag{4.8}$$

Where:

NWord: Number of words that a speech system uttered.

After each speech interface prompt, there is a pause and then a beep to signal the subjects that the system is ready to capture speech. All the durations were assumed fixed during the data reduction so that detailed time stamps for other utterances and pauses in between could be determined.

Time to Process and Prompt the State Name by the Machine

After the subject's utterance, the ASR device processes the user's input, and then provides feedback to the subject. The machine processing time was not affected by subjects' age, gender, or driving workload. However, when the system correctly recognized the desired state name that the user had uttered (31% from the empirical data), the processing time of state name was 0.86 s less than when the system provided multiple choices for state name and required the subject to choose one of them ($F_{(1, 37)} = 220.151$, $p < 0.001$). Using a stepwise regression, the machine processing time of state name can be predicted using the following equation ($R^2 = 0.863$, $F_{(1, 56)} = 352.6$, $p < 0.001$).

$$\text{M-State Processing Time} = 2.185 + 0.89 * \text{MultipleChoice} \quad (4.9)$$

Where:

MultipleChoice: The speech system provides multiple choices of possible state. No = 0. Yes = 1.

Using a stepwise regression, prompt times to provide feedback for the state name to subjects can be predicted using the following equation ($R^2 = 0.713$, $F_{(2, 50)} = 59.672$, $p < 0.001$).

M-State Prompt Time = 0.117 + 0.163 * NSyllable + 0.117 * NWord (4.10)

Where:

NSyllable: Number of syllables of the state name that a speech system prompted.

NWord: Number of words of the state name that a speech system prompted.

Time to Process and Prompt the City Name by the Machine

Similar to the state name, the system needs to process the user utterances for the city name and provide feedback to the subject. There were no main effects of age, gender, and driving conditions on the machine processing time of city name. Again, there were two results after machine processing of the information. One result was that the machine correctly recognized the city name uttered by the subjects, in which the probability was 12.7 % and the mean processing time was 2.25 s. Another result was that the machine provided multiple choices after processing the information, and subjects needed to take one more step to choose the correct city. The processing time was 3.19 s. The processing time of city name for these two results were statistically significant difference ($F_{(1, 269)} = 543.623$, $p < 0.001$). Using a stepwise regression, the machine processing time of city name can be predicted using the following equation ($R^2 = 0.744$, $F_{(1, 290)} = 844.8$, $p < 0.001$).

M-City Processing Time = 2.241 + 0.95 * MultipleChoice (4.11)

Where:

MultipleChoice: The speech system provides multiple choices of possible city. No = 0. Yes = 1.

Using a stepwise regression, the machine utterance times for feedback for the city name can be predicted using the following equation ($R^2 = 0.342$, $F_{(1, 40)} = 20.264$, $p < 0.001$).

M-City Prompt Time = 0.211 + 0.212 * NSyllable (4.12)

Where:

NSyllable: Number of syllables of the city name that a speech system prompted.

Time to Process and Prompt the Street Name by the Machine

Again, the system needs to process the user utterances of street name and provide feedback to the subject. There were no main effects of age, gender, and driving conditions on the machine processing time of street name. The mean processing time for the street name was 0.76 s less when the street name was correctly recognized by the system than when the machine provided multiple choices of feedback and the subjects needed to choose the correct one ($F_{(1, 266)} = 669.68$, $p < 0.001$). Using a stepwise regression, the machine processing time of street name can be predicted using the following equation ($R^2 = 0.789$, $F_{(1, 288)} = 1078$, $p < 0.001$).

M-City Processing Time = 2.772 + 0.782 * MultipleChoice (4.13)

Where:

MultipleChoice: The speech system provides multiple choices of possible street. No = 0. Yes = 1.

Using a stepwise regression, machine utterance time to provide feedback for the street name to the subjects can be predict using the following equation (R^2 = 0.514, $F_{(1, 19)}$ = 19.067, $p < 0.001$).

M-Street Prompt Time = 0.448 + 0.194 * NSyllable (4.14)

Where:

NSyllable: Number of syllables of the street name that a speech system prompted.

<u>Time to Process and Prompt the House Number by the Machine</u>

Empirical data from the 24 subjects revealed that the machine processing time of house number uttered by the subjects was fitted by a Weibull distribution with the parameters of scale (α), shape (β), and threshold (θ) equal to 0.37, 2.44, and 2.71 (n = 251, p = 0.25), respectively.

Using a stepwise regression, prompt time to provide feedback for the house number to the subjects can be predicted using the following equation (R^2 = 0.686, $F_{(1, 19)}$ = 39.344, $p < 0.001$). From the equation, the result reveals that the time for machine to utter the house number was 0.55 s per word.

M-House Number Prompt Time = 0.169 + 0.553 * NWord (4.15)

Where:

NSyllable: Number of words of the house number that a speech system prompted.

<u>Time to Process the Route by the Machine</u>

The final step of the destination entry task was to wait for the navigation system to process the entered information and find the route to the requested destination. After the system found the route, the system prompts *Please proceed to the highlighted route and then the route guidance will start* to conclude the destination entry tasks. The route processing time was significantly affected by whether the state of the requested destination is Michigan ($F_{(1, 256)} = 277.204$, $p < 0.001$). The out-of-state processing time was 8 s greater than the processing time for an in-state destination. Using stepwise regression, the time for the system to process the route information can be predicted using the following equation ($R^2 = 0.520$, $F_{(1, 19)} = 39.344$, $p < 0.001$).

$$\text{M-Route Processing Time} = 3.182 + 7.936 * \text{Out\_State} \quad (4.16)$$

Where:

Out_State: The state entered by the subject was Michigan or not. Yes = 0. No = 1.

As a first cut, rough rule of thumb, the utterance time for subjects and the system can be estimated approximately 0.2 s per syllable and 0.5 s per word. According to the results from John's study [87], the estimated duration for a customer to say a syllable in an unpracticed sentence was 0.17 s, which is only 0.03 s different (30 ms) from the results here, a very small difference.

The subtasks for subjects to think of state, city, street, and house number can be divided into five elements: (1) silence signaling the end of the machine's turn, (2) moving the eyes to the information display screen, (3) complex visual signal (information display), (4) cognitive verification (determine the correct one), and (5) cognition initiation of response (utter the correct one). In the John's research, the corresponding durations were (1) 0.6 s, (2) 0.18 s, (3) 0.29 s, (4) 0.05 s, and (5) 0.05 s. Adding these durations together, the estimated time for subjects to complete the subtask tasks on state, city, street, and house number was 1.17 s. The observed mean times for thinking of state, city, street, and house number from this study were 0.89, 0.91, 1.19, and 1.11 s, respectively. The results from both studies are close, especially for thinking of the street and house number.

B. Number of Turns to Complete the Destination Entry Tasks.

Table 4-11 shows the mean number of turns to complete a destination entry task based on the correctness of the information given and whether the subjects completed the task, as well as the number of turns needed by the subjects and the navigation system. Overall, to complete a destination entry task correctly required 20 ± 10 turns when the information given was correct. On the other hand, the total turns were almost twice (37 ± 16) for those trials where incorrect information was given. For most trials, the machine required 2 more turns than the subjects needed to complete the destination task. This is not surprising given that the navigation system used in this experiment was system-initiated interface and that (1) the machine asks for the information first and (2) the navigation system also needs one more prompt to inform the subjects that the route guidance has started.

Table 4-11 Total Turns Needed to Complete Destination Entry Task on the Correctness of Information Given (Machine Turn; Subject Turn)

Information	Complete / Give-up	Incorrect Final		Correct Final	
		Incorrect First	Correct First	Incorrect First	Correct First
Incorrect information	Complete	23 ± 10 (12; 11)		37 ± 16 (19; 18)	27 ± 9 (15; 13)
	Give-up	57 ± 29 (27; 30)	102 (49; 53)		
Correct information	Complete	34 ± 23 (16; 19)		26 ± 13 (13; 12)	15 ± 2 (9; 7)
	Give-up	33 (18; 15)			

Table 4-12 shows the total turns needed by navigation system and subject to complete destination entry tasks partitioned by scenario and driving condition. There were main effects of total number of turns on the error occurrence when subjects entered residential address (home, friend's / relative's address), $F_{(1, 110)} = 52.157$, $p < 0.001$. When errors occurred, either due to subjects or the machine, the total number of turns needed to complete the destination task was 10 more turns than for those trials without errors. There was a marginal significant difference ($F_{(2, 110)} = 2.819$, $p = 0.064$) of driving conditions on the total number of turns. This occurred because the subjects' home address was not stored in the address book (Scenario 3). Instead, the navigation system had an alternative way to store the home address (go home). Surprisingly, no subjects used the method "go home" to perform the destination entry task while the vehicle was parked. When subjects tried to enter the home address using the "address book" method, there were no such records, and subjects changed the entry method to either "street address" or "previous destination." On the other hand, subjects (32%) could use the "address book" method to enter a friend's or relative's address while driving (Scenario 7 and 11) and this method required three fewer turns than when subjects used the "street address" method to enter a friend's or relative's house (when no errors occurred).

Table 4-12. Frequency of Total Task Turns for Destination Entry Task by Scenarios

Driving Condition	Scenario	Task Completion	Incorrect Final		Correct Final	
			Incorrect First	Correct First	Incorrect First	Correct First
Parked	1*	Complete	24		39	
	Recreation	Give up	49			
	2	Complete	18		21	16
	Shopping	Give up				
	3	Complete			26	17
	Home	Give up				
	4	Complete	35		26	17

	Hospital	Give up				
Low Workload	5	Complete			22	14
	Home	Give up				
	6 *	Complete	23		38	28
	University	Give up	56	102		
	7	Complete			29	15
	Shopping	Give up	33			
	8	Complete			22	16
	Church	Give up				
High Workload	9	Complete			23	15
	Recreation	Give up				
	10*	Complete			36	27
	Attraction	Give up	63			
	11	Complete			27	13
	Home	Give up				
	12	Complete	66		30	16
	Shopping	Give up				

Note: * - Trials with incorrect information

Subjects also performed the tasks of entering shopping address under several different driving conditions. When no errors occurred, the variables of gender, age, and driving conditions did not significantly affect the total turns required to complete this destination entry task. On the other hand, subjects and the navigation system required 10 more turns to complete the destination task for shopping centers when errors occurred or when subjects switched entry methods than the total turns required without any errors ($F_{(1, 104)} = 21.77$, $p < 0.001$).

C. Destination Entry Method Difference

On average, subjects needed to perform 1.7 ± 1.1 attempts to complete the destination entry tasks, with 1.3 ± 0.8 and 2.7 ± 1.4 attempts when given correct and incorrect information, respectively. Among those 576 trials, the most frequent method used by subjects was “street address” (507), followed by “address book” (39), “POI” (30), and “previous destinations” (10), respectively (Table 4-13). The method used for the first attempt could represent how subjects would find destinations when they used their own navigation system. When entering the personal address (Table 4-14), the percentage for elderly subjects using “address book” as the first attempted entry method was only 6% (3/48 trials). On the other hand, the probability of using “address book” to enter the personal address for young subjects was 54%, especially in both driving conditions. The reason for the difference in entry method for a residential address is that young subjects may use the navigation system more frequently. Another reason may be that it is difficult for elderly subjects to remember these methods, as some of the commands of methods are not listed on the first page. There were three pages of commands relating to all possible destination entry methods after subjects say “Destination Help.”

Table 4-13. Pooled-Frequency of Method Used by Subjects at the First Attempt for Destination Entry Task (*: w/ error information)

Scenario	Address			Address Book		POI				Previous Destination	
	Final Incorrect	Final Correct		Final Correct		Final Incorrect		Final Correct		Final Correct	
	First Incorrect	First Incorrect	First Correct	First Incorrect	First Correct	First Incorrect	First Correct	First Incorrect	First Correct	First Incorrect	First Correct
Home	0	39	58	13	26	0	0	1	0	6	1
Shopping	4	58	73	0	0	0	0	17	1	1	0
Recreation*	6	53	31	0	0	0	0	15	1	0	0
Hospital	0	17	26	0	0	1	0	22	0	2	0
University*	5	35	1	0	0	1	1	5	0	0	0
Church	0	11	33	0	0	0	0	4	0	0	0
Attraction*	2	42	3	0	0	0	0	1	0	0	0
Subtotal	17	265	225	13	26	2	1	25	2	9	1

Note: * - Trials with incorrect information

Table 4-14. Frequency of Method Used by Subjects at the First Attempt for Entering Personal Home Address

Age	Driving Conditions	Address	Address Book	POI	Previous Destination
Young	Parked	9	5	1	1
	Low Workload	5	11	0	0
	High Workload	5	10	0	1
Middle-Age	Parked	14	2	0	0
	Low Workload	13	3	0	0
	High Workload	9	5	0	2
Old	Parked	15	1	0	0
	Low Workload	13	2	0	1
	High Workload	14	0	0	2
	Subtotal	97	39	1	7

Although the “POI” entry method was ranked third in terms of what subjects used to enter a destination on the first attempt, this occurred on when a subject tried to find a destination for shopping, recreation, and hospital, whose probabilities (using “POI” entry method) were 13%, 16%, and 46%, respectively. However, the probability of using “POI” as an entry method for destinations such as universities, churches, and attractions was relative low and the reason is unknown.

After the first attempt, 40% of the trials ended with incorrect results. When errors occurred or when the subjects could not find the requested destination on the first attempt, subjects needed to correct the error and sometimes changed the entry method. Table 4-15 shows the frequency of methods that subject used for the second attempt to enter destinations. Clearly, “street address” was the preferred method when errors occurred (88%).

Table 4-15. Frequency of Method Used by Subjects on the Second Attempt for Destination Entry Task While the Result from the First Attempt Was Incorrect

First Entry Method	Second Entry Method					
	Address	Address Book	POI	Previous Destination	Show Map	Previous Start Point
Address	174	3	8	1	2	0
Address Book	9	2	0	0	0	0
POI	18	1	8	1	0	0
Previous Destination	2	2	0	0	0	1
Total	203	8	16	2	2	1

After the second attempt, for 96% of those 232 trials, subjects still could not correctly find the requested destinations (Table 4-16). Again, “street address” was still the most frequently method used to re-attempt to enter a destination.

Table 4-16. Frequency of Method Used by Subjects on the Third Attempt for Destination Entry Task While the Result from the Second Attempt Was Incorrect

Second Entry Method	Third Entry Method				
	Address	Address Book	POI	Previous Destination	Show Map
Address	46	2	3	5	2
Address Book	4	1	0	0	0
POI	1	0	0	0	0
Previous Destination	1	0	0	0	0
Previous Start Point	1	0	0	0	0
Show Map	2	0	0	0	0
Total	55	3	3	5	2

The maximum number of destination entry attempts was 10. The frequency of each method used for each attempt is an important aspect of simulation model to be developed.

D. Errors and User Correction Strategies

There are two sources of errors when subjects interact with a speech-controlled navigation system to enter the destination – machine and human. The mechanism or reasons behind why the machine cannot recognize a user's prompt is beyond the scope of this research. However, the frequency of different types of machine errors and how subjects corrected these errors are provided.

There are three categories of human errors that occurred when subjects entered destinations: (1) information relevance, (2) system commands and entry method relevance, and (3) subject's knowledge of the related navigation. Véronis' proposed typology of error and Grice's conversation maxims may be applied to partially explain the listed types of errors that occurred in this study (Table 4-17). This list is not mutually exclusive.

Table 4-17 Categories of Human Errors and Examples Based on the Command–Based In-Vehicle Navigation System.

Category	Error Type	Example	Linguistic Principle
Information Relevant	Time out		Maxim of quantity
	Barge-in		Turn-Taking
	Stammer the prompt / command	Dixboro North Dixboro Road	Word insertion
	Provide incorrect information	City was Ann Arbor, not Ypsilanti	Maxim of quality
	Provide incomplete information	South Beyer Road, not Beyer	Word deletion
	Provide invalid information	No POI sub-category named "College"	Maxim of quality
	Provide invalid format of information	four zero zero zero, not four thousand	Word substitution
	Forgot to change the state name	State was Ohio from previous trial, but the state was Michigan for the current trial	Maxim of quality
	Pick the wrong choice	Street name was S. Beyer Road, not N. Beyer road	Maxim of quality
	Miss the correct information	Correct choice was shown on the list, but did not select	
	Ask experimenter for confirmation	" Is it in Ann Arbor?"	Maxim of Relevance
	Say unnecessary words	Umm North Dixboro Road	Word insertion
	Murmur	"Don't know what to do"	Maxim of Relevance
	Fail to find the correct information	Did not enter the house number and start route guidance	
Command and Entry Method Relevant	Cannot find the information by specific method	Use "Previous Destination" as entry method. After several steps and cannot find the information, the user change the method using "Street Address"	
	Change Entry method	Select "Find nearest POI" as the entry method, then say "Cancel"	
	Cannot determine the entry method	After going through the lists for all method, the user failed to pick the one and restart the trial.	Maxim of Manner
	Forget to say command word	LINE three, not three	Word deletion
	Say incorrect command	Find Nearest POI, not Nearest POI	Word deletion
	Say invalid command	Find Address, instead of Find Destination	Word substitution
Knowledge	Did not know the system reach the first step of the entry method	"Please enter the state name" was the first prompt when users entered the destination using "Street Address". The user tried to say "Go Back" to change the information	
Subject's Relevant	Did not know the ASR function has been deactivated	After saying "Show Map", the ASR function automatically deactivated. The user tried to say "Go Back" to change the information	
	Give-up		

For human errors relevant to the information provided by subjects, most of the error types violated the Grice's conversational maxims, especially for the "Maxim of Quality." When subjects forgot to change the state name, the information for city name is incorrect. An alternative way to explain this is that the system and subjects were not on a common ground. When subjects provide only incomplete information, such as "Beyer" to "South Beyer Road," one can argue that the error was due to "word deletion." Another way to explain such an error is that the type of error violates the Maxim of Quantity.

Human error types relevant to system command words can be easily explained as word deletion or substitution.

<u>Frequency of errors</u>

There were 1,088 errors distributed among 323 trials (56% of total 576 trials), averaging 3.4 errors/trial. Excluding those trials in which incorrect information was intentionally given, there were 487 errors in 183 trials (42% of the 432 trials with correct information). The mean number of errors dropped to 2.7 errors/trial. Table 4-18 presents descriptive statistics of the errors for each scenario. The number of errors was significantly affected by the age, driving conditions, and the correctness of information given (Table 4-19). The number of errors for elderly subjects was 1.3 times greater than the young and middle-age subjects when entering the destination, with means of 3.5, 2.2, and 2.2 times per trial, respectively ($F_{(2, 540)} = 11.832$, $p < 0.001$). When subjects performed a destination entry task while driving in high- and low-workload conditions, the mean of total errors was 1.0 and 0.8 times greater than the number of total errors that occurred while the vehicle was parked, with values of 3.0, 2.8, 2.0 times per trial, respectively ($F_{(2, 540)}$ = 6.007, p = 0.003). Thus, overall, driving did not lead to more errors with the speech interface, and if anything, the number of errors was the same or fewer.

Table 4-18. Descriptive Statistics of Errors by Scenarios

Scenario	Number of Trials	Sum	Mean ± Std. Dev.	Range (Min. – Max.)
1. Recreation*	48	151	3.1 ± 2.5	1 - 14
2. Shopping	20	28	1.4 ± 0.6	1 - 3
3. Home	24	46	1.9 ± 1.2	1 – 5
4. Hospital	22	60	2.7 ± 2.4	1 – 11
5. Home	15	35	2.3 ± 1.7	1 – 7
6. University*	48	227	4.7 ± 4.9	1 – 23
7. Shopping	24	88	3.7 ± 4.6	1 – 21
8. Church	15	28	1.9 ± 1.0	1 – 4
9. Recreation	16	36	2.2 ± 2.5	1 – 9
10. Attraction*	46	223	4.8 ± 3.6	1 – 14
11. Home	21	71	3.4 ± 2.9	1 – 10
12. Shopping	26	95	3.7 ± 3.0	1 – 12
Total	325	1088	3.4 ± 3.3	1 – 23

Note: * - Trials with incorrect information

Table 4-19. ANOVA of Total Number of Errors Occurred by Age, Gender, Driving Conditions, and Information Correctness for Destination Entry Task

Source	df	F	p-value
Age (A)	2	12.136	**< 0.001***
Gender (G)	1	0.545	0.461
Driving (D)	2	6.054	**0.003***
Information Correctness (I)	1	147.464	**< 0.001***
A * G	2	3.843	**0.022***
A * D	4	1.171	0.323
A * I	2	2.959	0.053
G * D	2	0.741	0.477
G * I	1	0.000	0.985
D * I	2	2.997	0.051
A * G * D	4	0.889	0.470
A * G * I	2	3.012	**0.050***
A * D * I	4	3.791	**0.005***
G * D * I	2	2.269	0.104
A * G * D * I	4	0.884	0.473
Error	540		

Note: * - Statistically significant, p < 0.05

When given an incorrect city name, the mean number of errors that occurred was 3 times greater than the number of errors that occurred when subjects performed the destination entry tasks with correct information, with values of 4.1 and 1.1 errors per trial, respectively ($F_{(1, 540)} = 148.12$, $p < 0.001$).

Table 4-20 shows that a total of 89 machine errors occurred among the 63 trials from 29 subjects with a mean of 1.4 times per trial and range of 1 to 7. The ASR error rate was 11% (58/576). Age, driving conditions and information correctness affected the mean number of machine-caused errors. The mean number of machine errors that occurred when elderly subjects entered the destination task was 1.1 times greater than the young and middle-age subjects ($F_{(2, 540)} = 4.746$, $p = 0.009$). The mean number of machine errors that occurred when subjects entered the destination task with incorrect information was 1.1 times greater than the mean of errors with correct information ($F_{(1, 540)} = 15.506$, $p<0.001$). When driving in high-workload conditions, machine errors increased 0.17 times than no driving ($p = 0.007$ with Bonferroni adjustment). There was no difference in mean number of machine errors between the high- and low-workload conditions.

Table 4-20. Descriptive Statistics of Machine Errors by Driving Conditions and Scenarios

Driving	Scenarios	Number of Trials	Sum of Errors	Mean ± Std. Dev.	Min. – Max.
Parked	Recreation*	7	10	1.4 ± 0.8	1 – 3
	Shopping	1	1	1	1
	Home	1	1	1	1
	Hospital	3	3	1	1
Low Workload	Home	3	4	1.3 ± 0.6	1 – 2
	University*	12	16	1.3 ± 0.5	1 – 2
	Shopping	5	5	1	1
	Church	2	2	1	1
High Workload	Recreation	2	2	1	1
	Attraction*	13	18	1.4 ± 0.8	1 – 3
	Home	6	19	3.2 ± 2.4	1 – 7
	Shopping	8	8	1	1

Note: * - Trials with incorrect information

Most errors that occurred on destination entry tasks were human errors. There were 999 human errors occurred among 317 trials), averaging 3.2 errors/trial. Table 4-21 shows the frequency of errors that occurred at different scenarios with driving conditions. Apparently, there was not one scenario that ended without any errors occurring. Obviously, most of the human errors occurred among those trials with incorrect information given, accounting for 56% of the total human errors.

Table 4-21. Descriptive Statistics of Human Errors by Driving Conditions and Scenarios

Driving	Scenarios	Number of Trials	Sum of Errors	Mean ± Std. Dev.	Min. – Max.
Parked	Recreation *	48	141	2.9 ± 2.2	1 – 11
	Shopping	20	27	1.4 ± 0.6	1 – 3
	Home	24	45	1.9 ± 1.2	1 – 5
	Hospital	21	57	2.7 ± 2.2	1 – 10
Low Workload	Home	14	31	2.2 ± 1.7	1 – 7
	University*	48	211	4.4 ± 4.6	1 – 23
	Shopping	23	83	3.6 ± 4.5	1 – 20
	Church	14	26	1.9 ± 0.9	1 - 4
High Workload	Recreation	14	34	2.4 ± 2.7	1 – 9
	Attraction*	44	205	4.5 ± 3.4	1 – 14
	Home	21	52	2.5 ± 1.6	1 – 7
	Shopping	24	87	3.6 ± 2.8	1 – 11

Note: * - Trials with incorrect information

The number of errors that occurred was significantly affected by age, driving conditions, and information correctness (Table 4-22). The mean of human errors that occurred among elderly subjects was 1.2 times greater than that of the young and middle-age subjects when entering the destinations, with trial means of 3.2, 2.1, and 2.1 times, respectively ($F_{(2, 540)} = 10.707$, $p < 0.001$). When subjects entered the destinations while the vehicle was parked, the mean number of human errors was 0.8 times that of the mean number of human errors that occurred while driving in either low- or high-workload conditions ($F_{(2, 540)} = 5.166$, $p = 0.006$). Again, the mean number of human errors was 2.8 times greater when subjects entered the destination given incorrect information than when given correct information ($F_{(1, 540)} = 148.957$, $p < 0.001$).

Table 4-22. ANOVA of Total Number of Errors Occurred by Age, Gender, Driving Conditions, and Information Correctness for Destination Entry Task

Source	df	F	p-value
Age (A)	2	10.707	**< 0.001***
Gender (G)	1	0.32	0.572
Driving (D)	2	5.166	**0.006***
Information Correctness (I)	1	148.957	**< 0.001***
A * G	2	2.655	0.071
A * D	4	0.893	0.468
A * I	2	2.779	0.063
G * D	2	1.468	0.231
G * I	1	0.001	0.976
D * I	2	3.161	**0.043***
A * G * D	4	0.741	0.565
A * G * I	2	2.213	0.110
A * D * I	4	2.239	**0.012***
G * D * I	2	3.009	**0.050***
A * G * D * I	4	0.646	0.630
Error	540		

Note: * - Statistically significant, p < 0.05

Types of Errors and Type of Correction Strategies

Table 4-23 shows the frequency of six types of machine errors and user's correction strategies. "Machine cannot recognize the user's prompt" and "Machine misrecognized the users prompt" were the two major categories and accounted for 66% of the machine errors. Eighty-three percent of the machine errors occurred when the subject entered the destination while driving. When the error "Machine cannot recognize the user's prompt" error occurred, subjects *repeated the prompt* or *repeated the prompt slower* to correct the errors in greater than 50% of the instances. The error of "machine misrecognized the user's prompt" is defined as: the speech interface provides incorrect feedback when the user's input is correct.

Table 4-23 Frequency of Machine Error by Type and User Correction Strategies for Destination Entry Task

Error Type	Example	User Correction Strategy	Frequency
Cannot recognize the user's prompt (36)	S: Six one zero zero. M: Sorry and house number or if you don't know that, please say show map or start guidance."	Repeat the prompt	14
		Repeat the prompt slower	5
		Spell the word	2
		Spell the word slower	1
		Rephrase the prompt	2
		Say "Next Page"	3
		Say "Enter City"	1
		Say "Go Back/Back"	2
		Provide requested information	4
		Deactivate the ASR and restart the trail	1
		Ask the experimenter for information	1
Misrecognize the user's prompt (23)	S: Two zero two eight. M: Two zero two THREE	Repeat the prompt	1
		Spell the word	5
		Spell the word slower	1
		Say "Next Page"	5
		Say "Go Back/Back"	5
		Provide requested information	2
		Deactivate the ASR and restart the trail	2
		Change the entry method	1
Cannot find the matching information from the database (10)	M: There is no database	Change the entry method	2
		Restart the trail	2
		Spell the word	2
		Repeat the prompt	1
		Say "Go Back/Back"	1
		Provide incorrect information	1
Deactivate the ASR function (15)		Reactivate the ASR	15
Machine failure (4)		Reactivate the ASR	3
		Use manual input	1
No response (1)		Repeat the prompt	1

Table 4-24 to Table 4-26 shows the frequency of the human errors that occurred relevant to three categories, as well as the frequency of error correction strategies used by the subjects. As the navigation system used in this experiment was a command-based interface, this restricted the strategies subjects could use to correct the errors. The information subjects can provide should exactly follow the requests or guidance provided by the system. The two types of errors, time-out and barge-in, accounted for one-third of human errors, and the correction strategy *repeat the prompt / repeat the prompt slower* was the most common method used to correct an error (45%).

Another major type of human error relevant to the information provided occurred when subjects provided incorrect information (26%). In this experiment, three of the trials involved providing incorrect information, (which some subjects did not realize was incorrect), that the experimenter requested to use for the first attempt. These trials were included to determine how subjects found the destination when some of the information was incorrect. The most frequently used strategy by subjects to correct this type of error was to say the command *go back / back*.

Table 4-24. Types and Frequency of Human Errors Relevant to Information Provided and User Correction Strategies for Destination Entry Task

Error Type	User Correction Strategy	Frequency
Time out (218)	Repeat the prompt /slower	68
	Provide requested information	59
	Say “Go back /Back”	21
	Say “Enter City”	7
	Say “Change State/City”	5
	Say “Next Page”	6
	Say “Show Map”	1
	Say “Help”	10
	Say “Start Guidance”	1
	Spell the word	5
	Rephrase the prompt	1
	Reactivate the ASR	13
	Cancel and restart the trial	1
	Ask the experiment for information	1
	Cancel and give up	4
	Time out again	15
Barge-in (121)	Repeat the prompt /slower	86
	Provide the requested information	10
	Spell the word	6
	Say “Go back /Back”	4

	Say “Enter City”	2
	Say “Next Page”	4
	Say “Help”	2
	Say “Start Guidance”	1
	Rephrase the prompt	1
	Reactivate the ASR	2
	Cancel and restart the trial	1
	Time out	2
Stammer the prompt or command (38)	Provide requested information	10
	Repeat the prompt /slower	9
	Say “Go Back”	6
	Say “Enter City”	1
	Say “Next Page”	1
	Say “Start Guidance”	1
	Spell the word	1
	Reactivate the ASR	5
	Time out	2
	(Machine corrected errors)	2
Provide incorrect information (264)	Say “Go Back/Back/Correct”	109
	Say “Change State/City/Street”	23
	Say “Next Page”	16
	Say “Start Guidance”	14
	Say “Enter City”	7
	Say “Help”	1
	Spell the word	23
	Provide requested information	17
	Repeat the prompt	13
	Rephrase the prompt	1
	Say “Show Map” to restart	7
	Cancel and restart the trial	7
	Say “Destination Help” to restart	2
	Reactivate the ASR	4
	Murmur	3
	Time out	16
	(Machine corrected errors)	1
Provide incomplete information (71)	Spell the word	22
	Provide requested information	21
	Say “Go Back”	11
	Say “Next Page”	9
	Say “Enter City”	1
	Repeat the prompt	5
	Rephrase	1
	Cancel and restart the trial	1
Provide invalid information (28)	Say “Go Back/Back”	10
	Say “Enter City”	2
	Say “Help”	2
	Say “Next Page”	1
	Provide requested information	5
	Repeat the prompt	2
	Spell the word	1

	Say “Show Map” to restart	1
	Cancel and restart the trial	2
	Reactivate the ASR	1
	Give up	1
Provide invalid format of information (75)	Rephrase	27
	Repeat the prompt	13
	Spell the word/slower	8
	Provide requested information	3
	Say “Go Back”	11
	Say “Enter City”	3
	Say “Next Page”	3
	Say “Change State”	1
	Say “Help”	1
	Say “Start Guidance”	1
	Reactivate the ASR	3
	Time out	1
Forgot to change the state name (10)	Say “Go Back”	5
	Say “ Change State”	3
	Repeat the prompt	1
	Spell the word	1
Pick the wrong choice (1)	Say “Go Back”	1
Miss the correct information (1)	Spell the word	1
Ask experimenter for confirmation (1)	Repeat the prompt	1
Say unnecessary words (6)	Repeat the prompt	3
	Spell the word	2
	Reactivate the ASR	1
Murmur (4)	Say “Go Back”	2
	Say “Show Map” and restart	1
	Repeat the prompt	1
Fail to find the correct information (10)	Restart the trial	2

The error type of providing incomplete information commonly occurred when the system asked for a street name. As described earlier, 14.8 % of the time subjects would not say the suffix, and 12.9 % of the time subjects did not say the direction. As it is common for people not to provide the suffix and direction, the interface designers should keep this in mind when designing speech interfaces for destination entry.

Table 4-25 shows the human errors related to the entry methods used to enter the destinations and commands accepted by the system. Some 42% of the time the error “cannot find the information by specific method” occurred when subjects entered the destination using "Find nearest POI" as the entry method. The system listed possible results based on the distance from the default location, which was Farmington Hills, Michigan. If the requested destination was not close to the default location, subjects needed to repeat the command "next page" several times to find the destination. Also, how well subjects knew the POI categories and subcategories affected their acceptance to use this method. For example, the subcategory of "higher education" was listed under the category of "community." When subjects tried to find the destination of "Washtenaw Community College," errors occurred when subject uttered "college" as the subcategory of POI. All of the errors resulted in subjects changing the entry method.

Table 4-25. Types and Frequency of Human Errors Relevant to Entry Method and System Commands and User Correction Strategies for Destination Entry Task

Error Type	User Correction Strategy	Frequency
Cannot find the information by specific method (33)	Change entry method	32
	Give up	1
Change Entry method (3)		3
Cannot determine the entry method (3)	Say "Go Back"	2
	Cancel and restart the trial	1
Forget to say command word (8)	Say "Go Back"	3
	Say "Enter City"	2
	Spell the word	1
	Provide requested information	1
	Time out	1
Say incorrect command (55)	Provide requested information	10
	Say "Go Back"	3
	Say "Next Page"	1
	Reactivate ASR	1
	Time out	1
	(Machine corrected errors)	39
Say invalid command (25)	Say "Go Back/Back"	9
	Say correct command	2
	Say "Help"	2
	Provide requested information	2
	Repeat the prompt	1
	Spell the word	1
	Cancel and restart the trial	3
	Reactivate ASR	1
	(Machine corrected errors)	4

When subjects were not familiar with the structures and predefined command of the navigation system, errors occurred frequently (Table 4-25 and 26). When the list or lists of machine feedback were shown on the screen and the subject was requested to choose the correct one, subjects might say "one," instead of "LINE one" which caused the machine to present another list or lists. Also the command word *Go Back* is invalid when the system asks for state name while using *find address* as the entry method. Subjects may not understand and repeat the command again.

Table 4-26. Frequency of Other Types of Human Errors and User Correction Strategies for Destination Entry Task

Error Type	User Correction Strategy	Frequency
Did not know the system reach the first step of the entry method (13)	Provide requested information	9
	Repeat the prompt	1
	Restart the trial	1
	Say invalid command	1
	Time out	1
Did not know the ASR function has been deactivated (3)	Restart the trial	3
Give-up (8)		

E. Driving Performance

Ten common measures of driving performance were examined -- mean speed, speed difference (maximum speed – minimum speed), maximum speed, variation of speed, mean TTC, minimum TTC, mean lane position, standard deviation of lateral position (SDLP), mean and minimum time to lane crossing (TLC). As was described earlier, the 48 subjects were divided into two groups. The data from one group was used to build the simulation model to predict the task performance and the data from another group was used to validate the model. To verify that the two groups are similar, their driving performance was compared.

Difference of Driving Performance Between the Model-Build and Validation Groups

A nonparametric two-independent test (Wilcoxon Rank Sum Test) was used to compare 10 driving performance variables with the 10 tasks for the two groups: model-build and model-validation groups. There were no statistically significant differences between these two groups ($p > 0.05$). Thus, one can assume that the subjects of these two groups were from the same population.

Driving Performance - Mean Speed

To identify the factors that significantly affect the driving performance of mean speed, a repeated-measure ANOVA was run with the independent variable of age (three age groups) and gender (male and female) as between-subject factors, and workload (low and high), and tasks (no secondary + four destination entry tasks) as within-subject factors. Data from one older subject was dropped because he stopped the vehicle during the experiment, as he could not maintain a safe gap during the high workload condition. Table 4-27 shows the repeated measures ANOVA, a 3 (age) X 2 (gender) X 2 (workload) X 5 (tasks) mixed design. The effects of age, gender and their interaction were not significant. However, the effect of mean speed on workload and tasks were significant, as well as the interaction of workload and tasks.

Table 4-27. ANOVA of Mean Speed for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	1.756	2	0.786
	Gender (G)	1.154	1	0.692
	A * G	2.584	2	0.702
	Error		41	
Within-Subjects	Workload (W)	49.089	1	**<0.001***
	W * G	0.008	1	0.929
	W * A	2.883	2	0.067
	W * G * A	1.206	2	0.310

Error (W)		41	
Tasks (T)	7.512	4	**<0.001***
T * G	0.587	4	0.672
T * A	1.568	8	0.138
T * G * A	0.392	8	0.924
Error (T)		164	
W * T	6.665	4	**<0.001***
W * T * G	0.465	4	0.761
W * T * A	0.572	8	0.799
W * T * G * A	0.225	8	0.986
Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

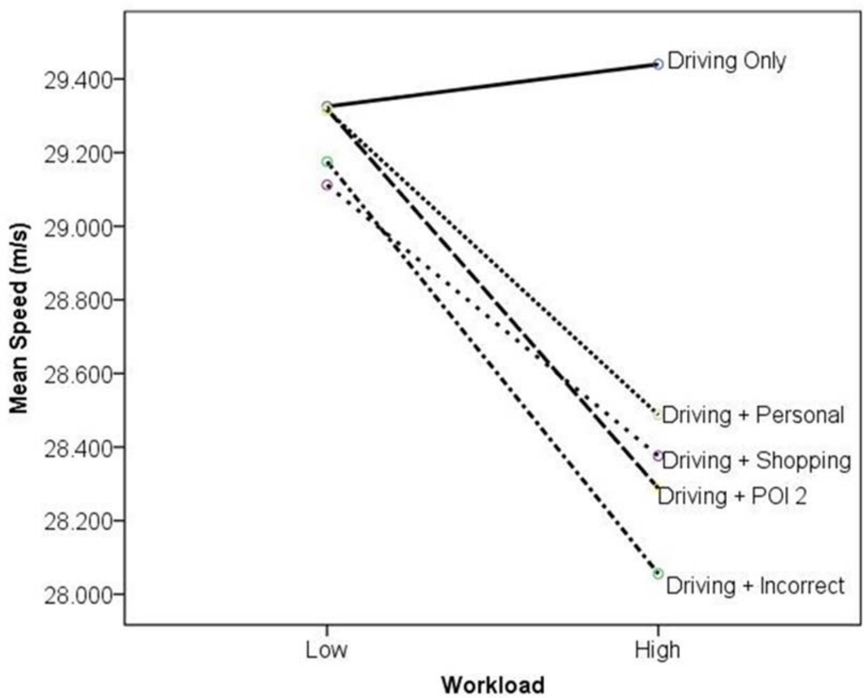


Figure 4-6. Mean Speed on Different Workload and Secondary Task Combinations

There was no statistically significant difference between the two workload conditions in which subjects drove (paired $t_{(-.811, 46)} = 0.421$) and did not perform a destination entry task. However, the mean speed while driving low-workload scenarios was statistically significantly higher than the mean speed while driving high workload scenarios when performing the destination entry tasks, including giving both correct and incorrect information (Figure 4-6). It also can be seen that there were no statistical differences in the mean speed among the five different task conditions while driving in low workload tasks. However, the mean speed was significantly less when subjects performed any destination tasks than versus no task in the high-workload scenarios. This confirms that argument in Oslund et al. (2005) that visual distraction leads to decreased travel speed [88]. Control theory would suggest that the driver reduces speed in order to gain time to complete the tracking control loop.

Driving Performance - Speed Difference

Table 4-28 shows a repeated measure ANOVA for speed difference (speed drop), a 3 (age) X 2 (gender) X 2 (workload) X 5 (tasks) mixed design. The effects of age, gender and their interaction were not statistically significant. However, the effect of speed difference on workload and tasks were significant difference, as well as the interaction of workload and tasks.

Table 4-28. ANOVA of Speed Difference for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	1.435	2	0.250
	Gender (G)	0.295	1	0.590
	A * G	2.239	2	0.119
	Error		41	
Within-Subjects	Workload (W)	93.283	1	**<0.001***
	W * G	0.903	1	0.347
	W * A	0.917	2	0.408
	W * G * A	0.759	2	0.475
	Error (W)		41	
	Tasks (T)	30.562	4	**<0.001***
	T * G	0.735	4	0.569
	T * A	1.100	8	0.366
	T * G * A	1.028	8	0.417
	Error (T)		164	

W * T	3.127	4	**<0.001***
W * T * G	1.487	4	0.208
W * T * A	0.274	8	0.974
W * T * G * A	0..508	8	0.849
Error (W * T)		164	

Note: * - Statistically significant, $p < 0.05$

There were no statistically significant differences in speed difference when subjects drove in the two different workload conditions (paired $t_{(-.976,\ 46)} = 0.334$) with no destination entry task. However, the speed differences while driving low-workload scenarios were statistically significantly less than the speed difference while driving high workload scenarios when performing the destination entry tasks, including for both the correct and incorrect information conditions.

For low-workload scenarios (Figure 4-7), the speed difference was two times greater when subjects performed destination entry task with incorrect information than the value for no secondary task (0.93 m/s vs. 1.96 m/s, $p < 0.05$). There were no significant differences when the information was correct while driving in the low-workload scenarios.

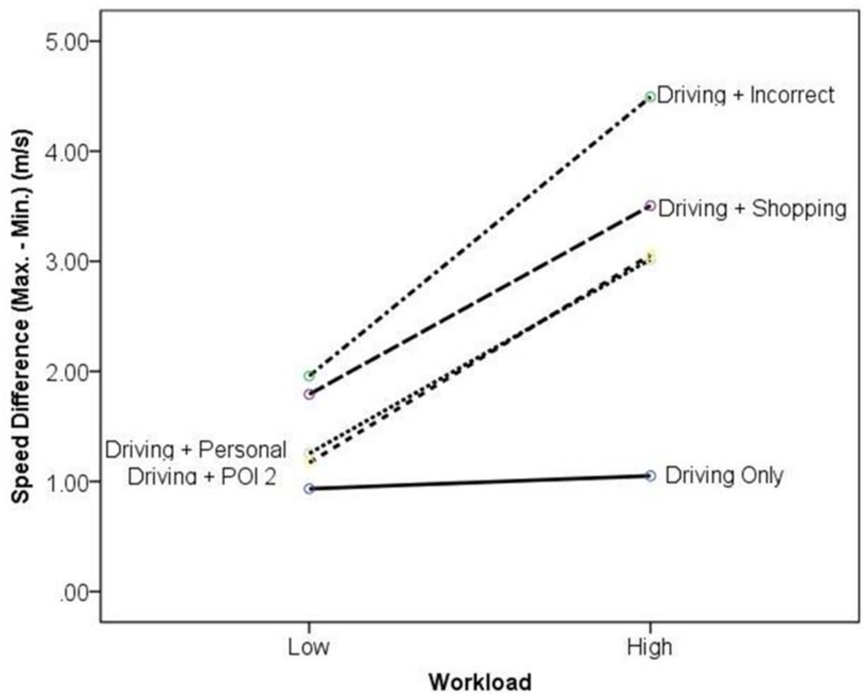


Figure 4-7. Speed Difference on Different Workload and Secondary Task Combinations

While driving in the high-workload scenarios, the speed difference when performing destination entry tasks was at least three times greater than the speed difference for driving only ($p < 0.05$).

While driving in both the low- and high-workload scenarios, the speed difference was statistically significantly greater for incorrect information than the value of correct information for the personal trial and POI 2 trial ($p < 0.05$). There was no difference when comparing the speed difference for correct information of the shopping task with incorrect information.

Driving Performance - Maximum Speed

A repeated-measure ANOVA of the maximum speed -- a 3 (Age) X 2 (Gender) X 2 (Workload) X 5 (Tasks) design -- is shown in Table 4-29. There were no statistically significant main effects of the age, gender, workload, or tasks. There were also no statistically significant differences due to the interaction of age with gender, workload or tasks. However, there was a statistically significant interaction of tasks and age on maximum speed ($F_{(8, 164)} = 2.366$, $p = 0.02$).

Table 4-29. ANOVA of Maximum Speed for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	0.170	2	0.845
	Gender (G)	0.011	1	0.918
	A * G	1.570	2	0.220
	Error		41	
Within-Subjects	Workload (W)	0.000	1	0.996
	W * G	0.168	1	0.684
	W * A	1.362	2	0.267
	W * G * A	0.282	2	0.756
	Error (W)		41	
	Tasks (T)	2.085	4	0.085
	T * G	0.543	4	0.704
	T * A	2.366	8	**0.020***
	T * G * A	1.439	8	0.184
	Error (T)		164	
	W * T	0.472	4	0.756
	W * T * G	0.237	4	0.917
	W * T * A	0.633	8	0.749
	W * T * G * A	0.286	8	0.970
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

Figure 4-8 shows the interaction between age and tasks on maximum speed. For elderly subjects, the value of maximum speed was greater when only driving than when they performed destination tasks with driving. Apparently for young subjects, the maximum speed when performing a destination entry task with incorrect information while driving was significantly greater than the maximum speed when performing destination entry task with correct information while driving or driving only.

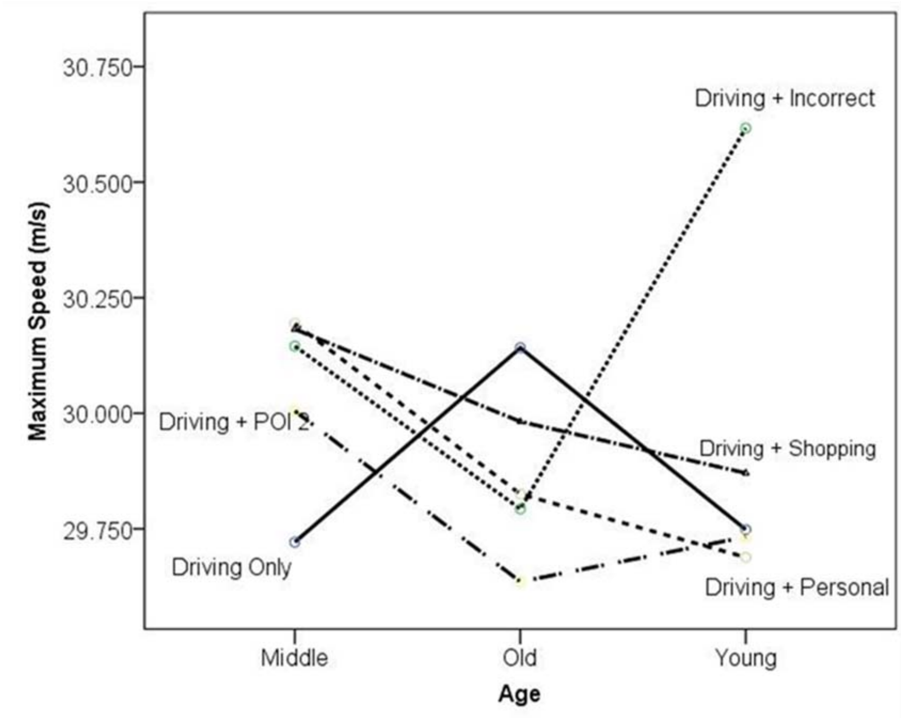


Figure 4-8. Maximum Speed on Different Age and Task Combinations

There were no statistically significant differences in maximum speed when subjects drove in the two different workload conditions with no destination entry task. Also, the maximum speed while driving low-workload scenarios was not statistically significant from the high-workload scenarios while performing the destination entry tasks.

Driving Performance - Speed Variation

A repeated measure ANOVA of speed variation -- a 3 (Age) X 2 (Gender) X 2 (Workload) X 5 (Tasks) design -- is shown in Table 4-30. There were no main effects of the gender or age, or the age and gender interaction. The main effects of workload ($F_{(1, 41)} = 77.999$, $p < 0.001$) and tasks ($F_{(4, 164)} = 24.859$, $p < 0.001$) on speed variation were statistically significant. The interaction of workload and tasks on the speed variation was also statistically significant ($F_{(4, 164)} = 10.665$, $p < 0.001$).

Table 4-30. ANOVA of Speed variation for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	2.349	2	0.108
	Gender (G)	0.206	1	0.652
	A * G	2.388	2	0.104
	Error		41	
Within-Subjects	Workload (W)	77.999	1	**<0.001***
	W * G	0.791	1	0.379
	W * A	0.786	2	0.463
	W * G * A	0.775	2	0.467
	Error (W)		41	
	Tasks (T)	24.859	4	**<0.001***
	T * G	0.739	4	0.567
	T * A	0.723	8	0.671
	T * G * A	0.974	8	0.458
	Error (T)		164	
	W * T	10.665	4	**<0.001***
	W * T * G	1.693	4	0.154
	W * T * A	0.067	8	1.000
	W * T * G * A	0.672	8	0.715
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

There were no statistically significant differences of speed variation when subjects drove in the two different workload conditions (paired $t_{(-.682,\ 46)} = 0.499$) with no destination entry task. However, the speed variation while driving the low-workload scenarios were statistically significantly less than the speed variation while driving the high-workload scenarios while entering destinations, for both correct or incorrect information. While subjects drove in the low-workload scenarios, the speed variation when performing destination entry task with incorrect information was statistically greater than the speed variation of driving only (0.28 vs. 0.544, p < 0.05). There was no difference in speed variation between the three destination entry tasks with correct information and driving only condition. On the other hand, the speed variation was at least 2.5 times greater when subjects performed any destination entry tasks while driving in the high-workload scenarios than speed variation of driving only without performing destination entry tasks (Figure 4-9).

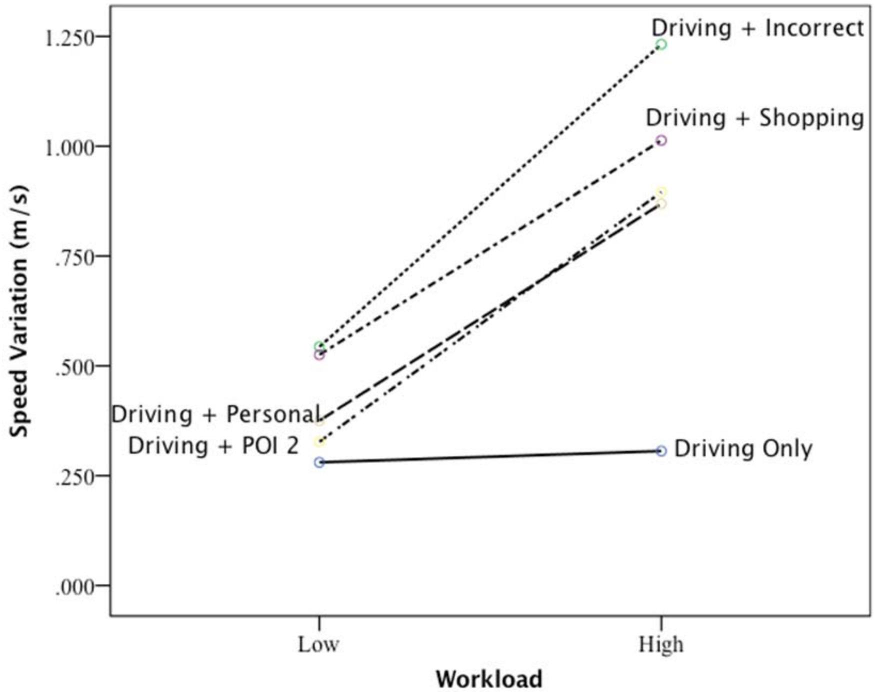


Figure 4-9. Speed Variation on Different Workload and Task Combinations

Driving Performance - Mean of Time-to-Collision (TTC)

A repeated measure ANOVA of the mean time-to-collision (TTC) with 3 (age) X 2 (gender) X 2 (workload) X 5 (tasks) is shown in Table 4-31. There were no gender and age effects on the mean TTC and also no effect of the gender and age interaction. The mean TTC while driving in the low workload scenarios was greater than the mean TTC while driving in high workload ($F_{(1, 41)} = 10800.567$, $p < 0.001$). This is reasonable as the lead vehicle was designed to be closer to the subject vehicle in high workload scenarios. Tasks also affected the mean TTC ($F_{(4, 164)} = 3.605$, $p = 0.008$). However, there was an interaction between the workload and tasks ($F_{(4, 164)} = 73.305$, $p < 0.001$), Figure 4-10. The mean TTC for driving only was significantly less than the mean TTC when subjects performed destination entry tasks while driving in low workload scenarios (0.4 s less, $p < 0.05$). On the other hand, the mean TTC for driving only was significantly higher than the mean TTC when subjects performed destination entry tasks while driving in high workload scenarios (0.3 s more, $p < 0.05$). Without performing destination entry task, the mean TTC when subjects drove in the low workload scenarios was statistically significantly greater than the mean time of time to collision when subjects drove in the high workload scenarios (paired-$t_{(46)} = 44.000$, $p < 0.001$).

Also, the mean TTC while driving low workload scenarios was statistically significantly greater than the TTC while driving the high workload scenarios when performing the destination entry tasks, for both the correct and incorrect information conditions.

Table 4-31. ANOVA of Mean Time-to-Collision (TTC) for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	2.646	2	0.083
	Gender (G)	0.048	1	0.828
	A * G	0.021	2	0.979
	Error		41	
Within-Subjects	Workload (W)	10800.567	1	**<0.001***
	W * G	0.003	1	0.958
	W * A	1.474	2	0.241
	W * G * A	1.875	2	0.166
	Error (W)		41	
	Tasks (T)	3.605	4	**0.008***
	T * G	0.339	4	0.852
	T * A	1.102	8	0.364
	T * G * A	0.555	8	0.814
	Error (T)		164	
	W * T	73.305	4	**<0.001***
	W * T * G	0.970	4	0.425
	W * T * A	0.663	8	0.723
	W * T * G * A	0.591	8	0.785
	Error (W * T)		164	

Note: * - Statistically significant, $p < 0.05$

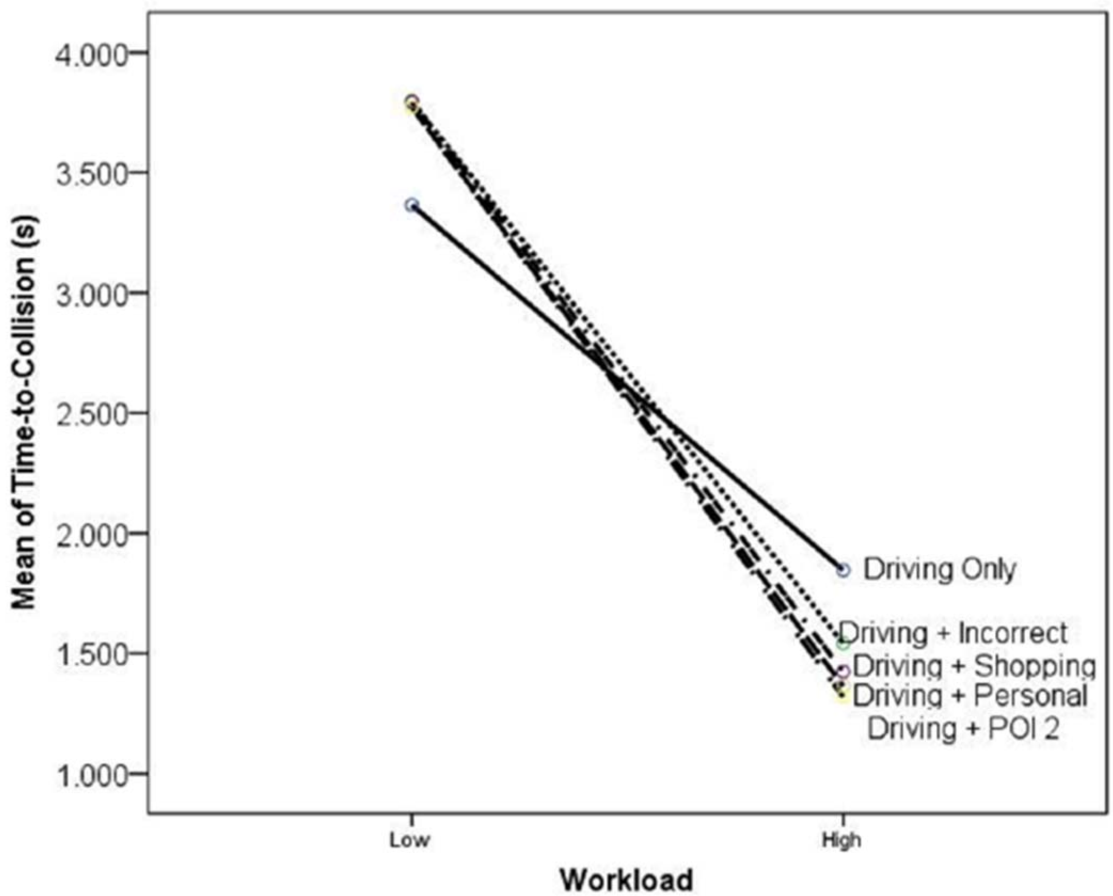


Figure 4-10. Mean of Time-to-Collision on Different Workload and Task Combinations

Driving Performance - Minimum of Time-to-Collision (TTC)

A repeated measure ANOVA of the minimum time-to-collision (TTC) with 3 (age) X 2 (gender) X 2 (workload) X 5 (tasks) is shown in Table 4-32. There were no gender and age effects on the minimum TTC and also no effect of the gender and age interaction. The minimum TTC while driving in low workload scenarios was greater than the mean TTC while driving in high workload ($F_{(1, 41)}$ = 20168.916, $p < 0.001$). Again, this is because the lead vehicle was designed to be closer to the subject vehicle in high workload scenarios. Tasks also affected the mean TTC ($F_{(4, 164)}$ =12.613, $p < 0.001$). However, there was interaction between the workload and tasks ($F_{(4, 164)}$ = 317.251, $p < 0.001$), Figure 4-11.

Table 4-32. ANOVA of Minimum Time-to-Collision (TTC) for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	0.355	2	0.703
	Gender (G)	0.006	1	0.939
	A * G	0.421	2	0.660
	Error		41	
Within-Subjects	Workload (W)	20168.916	1	**<0.001***
	W * G	0.220	1	0.642
	W * A	7.382	2	**0.002***
	W * G * A	2.818	2	0.071
	Error (W)		41	
	Tasks (T)	12.613	4	**<0.001***
	T * G	0.755	4	0.556
	T * A	1.261	8	0.267
	T * G * A	2.210	8	**0.029***
	Error (T)		164	
	W * T	317.251	4	**<0.001***
	W * T * G	1.361	4	0.250
	W * T * A	1.093	8	0.370
	W * T * G * A	1.183	8	0.312
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

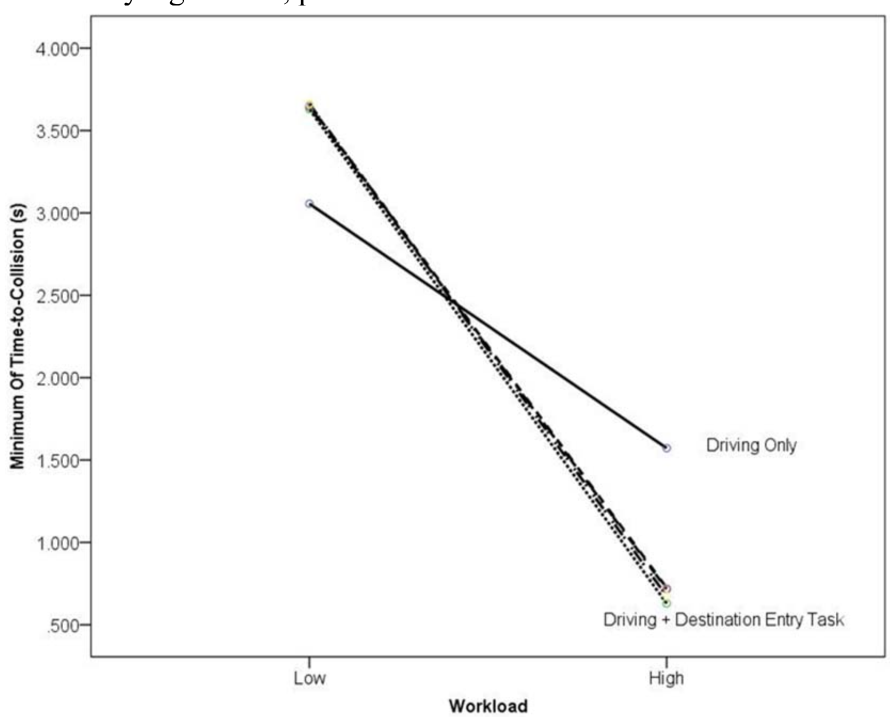


Figure 4-11. Minimum of Time-to-Collision on Different Workload and Task Combinations

Without performing destination entry tasks, the minimum time of time to collision when subjects drove in low workload scenarios was significantly longer than the minimum time of time to collision when subjects drove in high workload scenarios (paired-$t_{(46)}$ = 30.809, $p < 0.001$). Also, the minimum time of time to collision while driving low workload scenarios were significantly longer than the minimum time of time to collision while driving high workload scenarios when performing the destination entry tasks, including giving both correct and incorrect information.

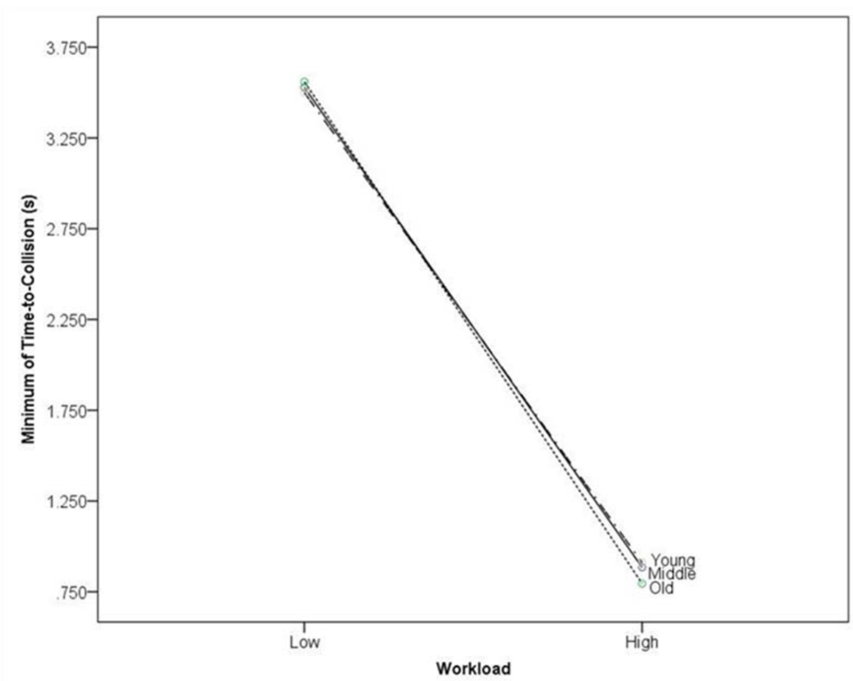


Figure 4-12. Minimum of Time-to-Collision on Different Workload and Age Combinations

There was also an interaction of workload and age on minimum TTC ($F_{(2, 41)}$ = 7.382, $p = 0.002$). Elderly subjects maintained longer TTCs while driving in the low workload scenarios, but the minimum TTC was shortest while driving in the high workload scenarios when compared with middle-aged and young subjects (Figure 4-12).

Driving Performance - Mean Lateral Lane Position

The positive value of lateral lane position means that subjects drove the vehicle towards the right hand side of the lane. A repeated-measure ANOVA of the mean of the lateral lane position was computed (Table 4-33). There were no age or gender effects on the mean of lateral lane position or an interaction of age and gender. Workload had a major effect on the mean of lateral lane position ($F_{(1, 41)}$ =8.428, p = 0.006).

Table 4-33. ANOVA of Mean of Lateral Lane Position for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-vlaue
Between-Subject	Age (A)	2.195	2	0.124
	Gender (G)	0.104	1	0.749
	A * G	0.579	2	0.565
	Error		41	
Within-Subjects	Workload (W)	8.428	1	**0.006***
	W * G	0.004	1	0.952
	W * A	0.038	2	0.963
	W * G * A	2.242	2	0.119
	Error (W)		41	
	Tasks (T)	0.218	4	0.928
	T * G	1.067	4	0.374
	T * A	0.693	8	0.697
	T * G * A	0.323	8	0.957
	Error (T)		164	
	W * T	2.948	4	**0.022***
	W * T * G	0.271	4	0.896
	W * T * A	1.585	8	0.133
	W * T * G * A	1.063	8	0.392
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

While driving in the low-workload scenarios, subjects drove toward the right compared with driving in the high-workload scenarios which was toward the left (paired-$t_{(46)}$ = 2.748, p = 0.009), except the condition when entering personal destination (Figure 4-13). Also, there was a marginal difference of the mean lane deviation to the right when subjects drove in the low workload scenarios than the mean lane deviation when subject drove in the high workload scenario when performing the destination entry task with incorrect information (paired-$t_{(46)}$ = 1.845, p = 0.071). There was a significance difference of the mean deviation to the right while subjects drove in the low-workload scenarios than the mean lane deviation while subjects drove in the high-workload scenario when performing the destination entry task using the POI method for shopping (paired-$t_{(46)}$ = 3.837, $p < 0.001$). There was no difference while driving in the low and high workload scenarios when entering a destination with their relatives'/friends' address or a community POI.

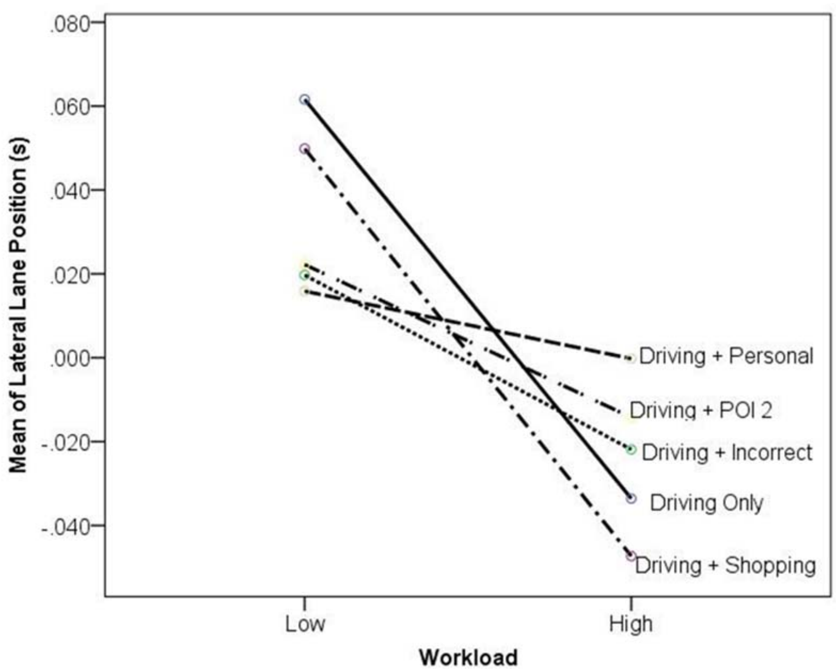


Figure 4-13. Mean of Lateral Lane Position on Different Workload and Task Combinations

Driving Performance - Standard Deviation of Lateral Lane Position (SDLP)

Some researchers use the standard deviation of lane position (SDLP) to represent the variation of the lateral position deviated from the center of lane, an indicator of how well the driver is controlling the vehicle [11].

A repeated-measure ANOVA of the standard deviation of lateral lane position was computed (Table 4-34). There were no statistically significant effects of age or gender effects nor an interaction of age and gender. The effects of Workload ($F_{(1, 41)}$ =27.605, p<0.001) and Tasks ($F_{(4, 164)}$ =22.932, p < 0.001) were highly significant (Figure 4-14). Also significant was the interaction of workload and tasks.

Table 4-34. ANOVA of Standard Deviation of Lane Position for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	2.807	2	0.072
	Gender (G)	0.206	1	0.652
	A * G	0.591	2	0.559
	Error		41	
Within-Subjects	Workload (W)	27.605	1	**<0.001***
	W * G	1.551	1	0.220
	W * A	0.220	2	0.803
	W * G * A	0.189	2	0.829
	Error (W)		41	
	Tasks (T)	22.932	4	**<0.001***
	T * G	1.586	4	0.180
	T * A	2.562	8	**0.012***
	T * G * A	1.313	8	0.241
	Error (T)		164	
	W * T	0.621	4	0.648
	W * T * G	0.685	4	0.603
	W * T * A	1.470	8	0.172
	W * T * G * A	0.828	8	0.579
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

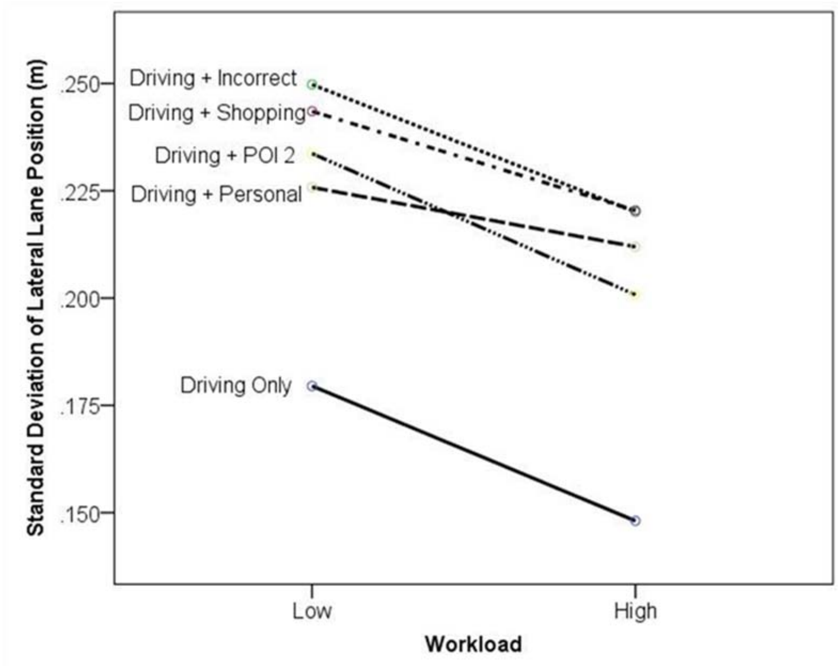


Figure 4-14. Standard Deviation of Lane Position on Different Workload and Task Combinations

Without performing destination entry task, the value of SDLP when subjects drove in the low workload scenarios was statistically significantly greater than when subjects drove in the high workload scenarios (paired-$t_{(46)}$ = 2.514, p = 0.014). The SDLP was statistically significantly less for subjects only driving when compared with the SDLP for conditions in which subjects performed destination entry tasks in both the low and high workload scenarios. Also, there was statistically significantly larger SDLP when subjects drove in the low workload scenarios than the standard deviation of lane position from the center of lane while subject drove in the high workload scenario when performing the destination entry task with incorrect information (paired-$t_{(46)}$ = 3.417, p = 0.001) . There was significantly larger SDLP when subjects drove in the low workload scenarios than the SDLP while subjects drove in the high workload scenario when performing the destination entry task with POI of shopping (paired-$t_{(46)}$ = 2.384, p = 0.021) and POI 2 (paired $t_{(46)} = 3.226$, $p = 0.002$). There was no difference while driving in low and high workload scenarios when entering destination with their relatives'/friends' address.

The interaction of age and tasks was significant ($F_{(8, 164)} = 2.562$, $p = 0.012$, Figure 4-15. When driving only, the middle-age subjects had better vehicle control (smaller SDLP) when compared with the young subjects. However, when performing destination entry tasks while driving, the value of SDLP was greater for middle-aged subjects than for young subjects. All of these effects were not statistically significant. The value of SDLP was significant greater for elderly subjects entering a shopping address when compared with the value of SDLP for young subjects entering a shopping address while driving (0.284 m vs. 0.190 m, $p < 0.05$).

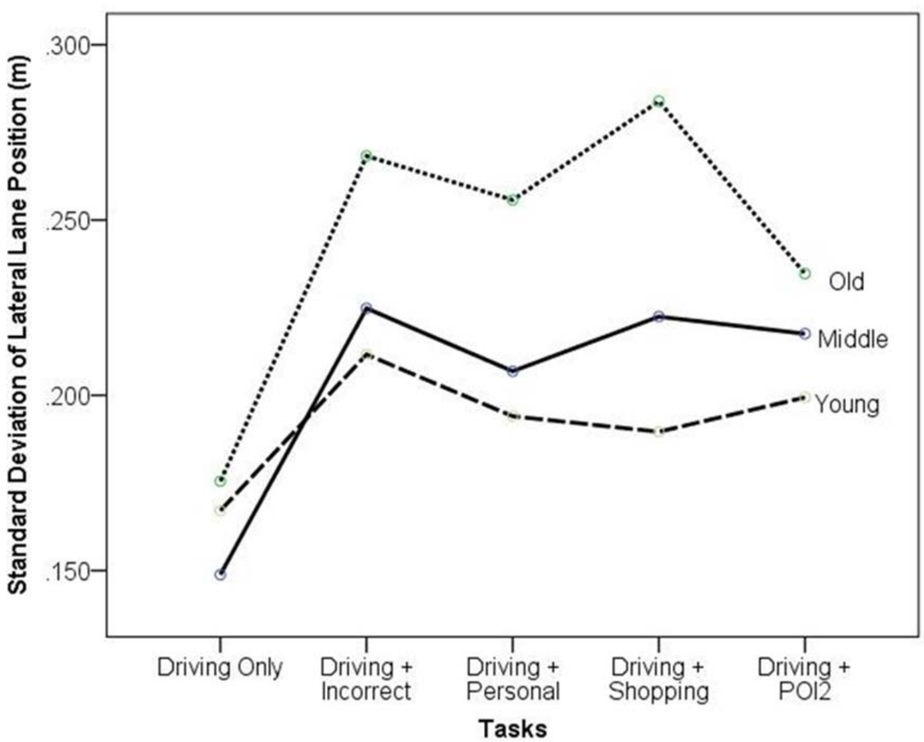


Figure 4-15. Stand Deviation of Lane Position on Different Workload and Age Combinations

Driving Performance - Mean of Time-to-Lane-Crossing (TLC)

A repeated measure ANOVA of mean time-to-lane-crossing (TLC) with 3 (age) X 2 (gender) X 2 (workload) X 5 (tasks) is shown in Table 4-35. There were no significant of gender and age effects or their interaction. There was marginal effect of the mean TLC on workload ($F_{(1, 41)} = 3.268$, $p = 0.077$). Tasks ($F_{(4, 164)} = 29.267$, $p < 0.001$) and the interaction of workload and tasks ($F_{(4, 164)} = 3.186$, $p = 0.015$) were statistically significant. The mean TLC while driving only was statistically significantly greater than the mean TLC when subjects performed destination entry tasks while driving in the high workload condition.

Table 4-35. ANOVA of Mean of Time-to-Lane-Crossing (TLC) for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p-value
Between-Subject	Age (A)	2.429	2	0.101
	Gender (G)	0.283	1	0.598
	A * G	0.631	2	0.537
	Error		41	
Within-Subjects	Workload (W)	3.286	1	0.077
	W * G	0.267	1	0.608
	W * A	0.559	2	0.576
	W * G * A	0.175	2	0.840
	Error (W)		41	
	Tasks (T)	29.267	4	**<0.001***
	T * G	0.821	4	0.514
	T * A	0.504	8	0.852
	T * G * A	0.623	8	0.757
	Error (T)		164	
	W * T	3.186	4	**0.015***
	W * T * G	1.099	4	0.359
	W * T * A	0.821	8	0.585
	W * T * G * A	1.364	8	0.216
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

The mean TLC when subjects drove in the low workload scenarios was statistically significantly less than for the high workload scenarios (paired-$t_{(46)} = -2.969$, $p = 0.005$)

without performing destination entry task (Figure4-16). However, there was no statistically significant difference of the mean TLC when subjects drove in the low workload scenarios than the mean TLC when subjects drove in the high workload scenarios while performing any destination entry tasks.

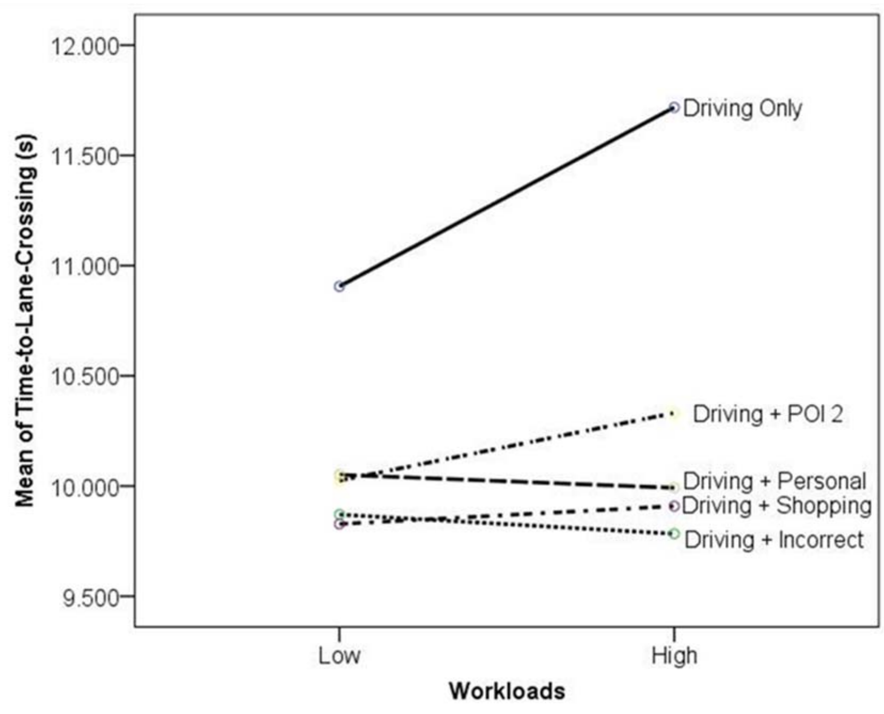


Figure 4-16. Mean TLC on Different Workload and Task Combinations

Driving Performance - Minimum Time-to-Lane-Crossing (TLC)

TLC was positive when approaching the right edge of the lane and negative when approaching the left edge of the lane. A repeated measure ANOVA of minimum TLC was computed (Table 4-36). There were no significant effects of age and gender effects on the mean of minimum TLC nor an interaction of age and gender. Tasks had a significant effect on the minimum TLC ($F_{(1, 41)} = 26.296$, $p < 0.001$, Figure 4-17).

Table 4-36. ANOVA of Minimum of Time-to-Lane-Crossing (TLC) for the Effect of Age, Gender, Workload, and Tasks

Effect	Source	F	d.f.	p
Between-Subject	Age (A)	2.815	2	0.072
	Gender (G)	0.1.039	1	0.314
	A * G	0.945	2	0.397
	Error		41	
Within-Subjects	Workload (W)	3.771	1	0.059
	W * G	0.381	1	0.540
	W * A	0.767	2	0.471
	W * G * A	0.041	2	0.960
	Error (W)		41	
	Tasks (T)	26.296	4	**<0.001***
	T * G	2.021	4	0.094
	T * A	1.418	8	0.192
	T * G * A	1.971	8	0.053
	Error (T)		164	
	W * T	0.523	4	0.719
	W * T * G	2.626	4	**0.037***
	W * T * A	2.428	8	**0.017***
	W * T * G * A	1.661	8	0.112
	Error (W * T)		164	

Note: * - Statistically significant, p < 0.05

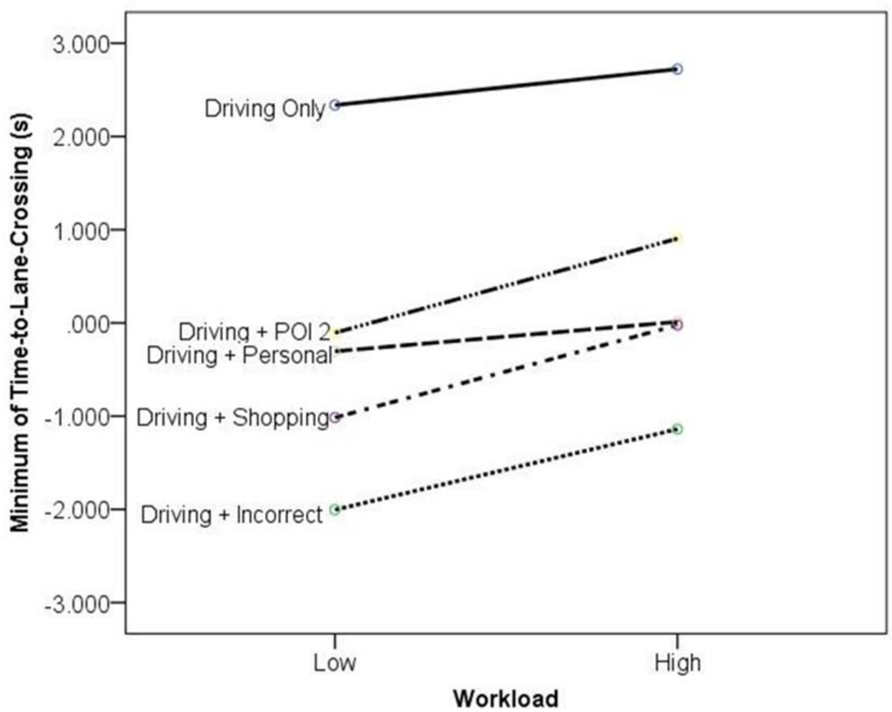


Figure 4-17. Minimum TTC on Different Workload and Task Combinations

The minimum of time-to-lane crossing (TLC) when subjects drove in the low-workload scenarios was not statistically significantly different from the minimum TLC when subjects drove in the high-workload scenarios (paired $t_{(-1.018,\ 46)} = 0.314$) without performing a destination entry task. Also, there was no statistically significant difference of minimum TLC while subjects drove in the low workload scenarios than the minimum TLC while subjects drove in the high workload scenarios when performing any destination entry tasks.

Table 4-37 summarizes the results on the driving performance when subjects performed the destination entry task. Surprisingly, age and gender did not have any significant effects on driving performance for all measures. However, different driving workload did affect the driving performance when entering an address.

Table 4-37. Effect of Statistical Significant Difference ($p < 0.05$) on the Driving Performance Measurement

	Age (A)	**Gender (G)**	**Workload (W)**	**Task (T)**	**W * T**	**W * A**	**T * A**
Mean Speed			X	X	X		
Speed Difference			X	X	X		
Max. Speed							X
Speed Variation			X	X	X		
Mean TTC			X	X	X	X	
Min TTC			X	X	X	X	
Mean Lateral Lane Position			X		X		
SDLP			X	X			X
Mean TLC				X	X		
Min TLC							

4.3.2 Music Selection

A. Task Completion Rate, Task Completion Time, and Detailed Time Associated with Each Utterance

Task Completion frequencies

There were a total of 480 trials performed by the 32 subjects in the music selection task with an equal number in each age (young and middle-age) and gender group. Subjects completed 99.0% of the trials (Table 4-38), giving up (balking) on just five trials. Among the 475 completed trials, there were 6 trials (1.3%; 6/475) that subjects thought they had finished correctly, but in fact, were incorrect. Therefore, there were 469 trials that ended with the correct music being selected. Furthermore, there were 84% (392/469) that the subjects completed the music selection tasks without any errors, neither from subjects nor from the MP3 player.

Table 4-38. Frequency of Task Completion for Music Selection Task at Different Driving Conditions

Driving	Information Given	Complete / Give-up	Incorrect Final	Correct Final	
			Incorrect First	Incorrect First	Correct First
Parked	Artist	Complete	0	6	26
		Give-up	0	0	0
	Album	Complete	0	1	31
		Give-up	0	0	0
	Song/Artist/Album	Complete	0	19	75
		Give-up	2	0	0
Low Workload	Artist	Complete	0	2	30
		Give-up	0	0	0
	Album	Complete	0	2	30
		Give-up	0	0	0
	Song/Artist/Album	Complete	2	28	65
		Give-up	1	0	0
High Workload	Artist	Complete	0	1	31
		Give-up	0	0	0
	Album	Complete	1	1	30
		Give-up	0	0	0
	Song/Artist/Album	Complete	3	17	74
		Give-up	2	0	0

Sub-Total	11 (2.3%)	77 (16.0%)	392 (81.7%)

Task Completion Time

The mean task completion time for all trials for the music selection task was 24.71 ± 18.52 s (n = 475). For trials that ended with the correct music selection, and there was at least one error in the process: the mean task completion time was 51.11 ± 24.97 s, which was 2.6 times longer than the mean task completion time for those trials without any errors (19.44 ± 11.08 s). The task completion time for error-free trials was longer than the 15-second rule (SAE J2364, 2004) allows [80].

As shown in Table 4-39, there were significant differences due to age, driving conditions, and information given, but not gender. The task completion time to select music for young drivers was 7.2 s, which was significantly shorter than the time for middle-age drivers (19.5 s vs. 26.7 s; $F_{(1, 445)} = 22.45$, $p < 0.001$). There was also a main effect of driving condition on task completion time, $F_{(2, 445)} = 4.262$, $p=0.015$ (Table 4-40). Post hoc tests showed that the task completion time while driving in the low workload condition (26.2 s) was significantly greater than the time while the vehicle was parked (22.1 s, $p = 0.035$) or while driving in the high workload condition (21.0 s, $p=0.005$). Task completion times for when subjects were given a specific artist or album (one piece of information) to select a specific song were statistically significantly less than the time when given 3 pieces of information (song title, artist name, and album name) and subjects selected a specific song (14.8 s vs. 31.4 s $F_{(1, 445)} = 118.82$, $p < 0.001$). Table 4-40 also shows that the task completion time when selecting a specific song was almost twice as long as the task completion time when selecting an artist or album during any driving conditions when no errors occurred.

Table 4-39. ANOVA of Task Completion Time on the Effect of Age, Gender, Driving Conditions, and Number of Information Given

Source	df	F	p-value
Age (A)	1	22.452	**<0.001***
Gender (G)	1	0.091	0.763
Driving (D)	2	4.262	**0.015***
Numb of Information Given (I)	1	118.823	**<0.001***
A * G	1	0.434	0.510
A * D	2	0.827	0.438
A * I	1	2.493	0.115
G * D	2	0.162	0.850
G * I	1	1.043	0.308
D * I	2	1.845	0.159
A * G * D	2	0.414	0.661
A * G * I	1	0.248	0.619
A * D * I	2	0.139	0.870
G * D * I	2	0.029	0.972
A * G * D * I	2	0.491	0.612
Error	445		

Note: * - Statistically significant, $p < 0.05$

Table 4-40. Task Completion Time for Music Selection Task at Different Driving Conditions (in seconds)

Driving	Information Given	Complete / Give-up	Incorrect Final	Correct Final	
			Incorrect First	Incorrect First	Correct First
Parked	Artist	Complete		30.25±18.61	12.70±2.33
		Give-up			
	Album	Complete		60.61	12.8±3.26
		Give-up			
	Song/Artist/Album	Complete		50.82±27.72	23.71±10.47
		Give-up	36.62±33.43		
Low Workload	Artist	Complete		50.19±19.04	13.22±1.78
		Give-up			
	Album	Complete		65.16±19.16	12.92±3.07
		Give-up			
	Song/Artist/Album	Complete	29.84±23.72	59.01±24.80	26.84±13.93
		Give-up	89.73		
High Workload	Artist	Complete		29.18	13.20±3.18
		Give-up			
	Album	Complete	16.22	26.46	12.86±3.16
		Give-up			
	Song/Artist/Album	Complete	35.63±28.43	46.43±22.62	24.20±12.79
		Give-up	18.78±7.56		

Thinking and Response Time

Thinking and Response Times were significantly affected by age and gender. Among the 469 trials completed for selecting the correct music, the mean thinking time for middle-age drivers was 1.8 s significantly longer than the thinking time for young drivers (8.8 s vs. 7.0 s; $F_{(1\ 445)} = 43.986$, $p < 0.001$). There also was a main effect of gender on the thinking time, $F_{(1,\ 445)} = 6.409$, $p = 0.012$ (Table 4-41). Male drivers needed 0.67 s more time to perform the music selection task than female drivers. Driving workload conditions played a marginal effect on the thinking time ($F_{(2,\ 445)} = 2.682$, $p = 0.070$). Post hoc tests showed that the thinking time while driving in the high-workload condition (8.22 s) was significantly greater than the time while the vehicle was parked (7.50 s, p=0.040 with Bonferroni adjustment). There was no statistically significant difference in thinking time while driving in the low-workload condition compared with while the vehicle was parked.

Table 4-41. ANOVA of Thinking and Response Time for the Effect of Age, Gender, Driving, and Number of Information Items Given

Source	df	F	p-value
Age (A)	1	43.986	**<0.001***
Gender (G)	1	6.409	**0.012***
Driving (D)	2	2.862	0.07
Numb of Information Given (I)	1	1.395	0.238
A * G	1	1.034	0.310
A * D	2	0.643	0.526
A * I	1	0.121	00.728
G * D	2	0.185	0.831
G * I	1	0.276	0.600
D * I	2	1.394	0.249
A * G * D	2	0.458	0.633
A * G * I	1	0.833	0.326
A * D * I	2	1.063	0.346
G * D * I	2	0.944	0.390
A * G * D * I	2	0.154	0.858
Error	445		

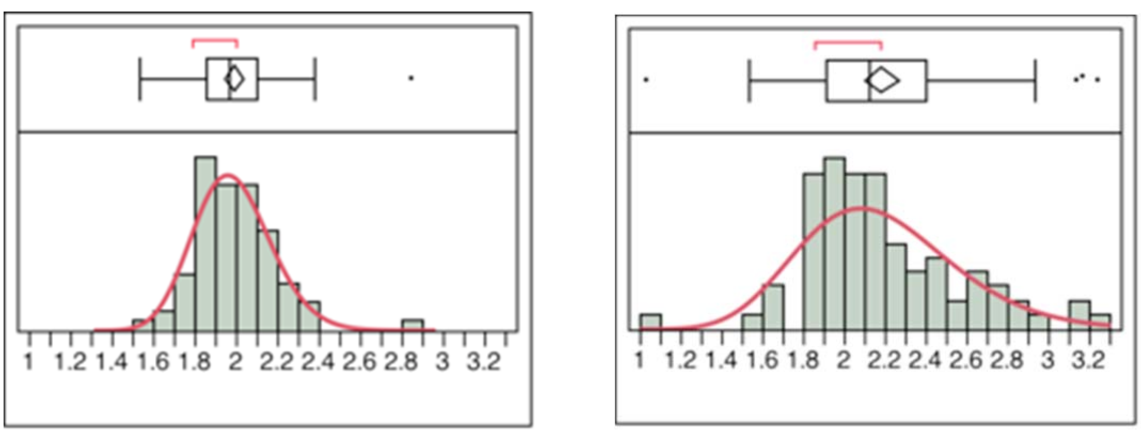
Note: * - Statistically significant, p < 0.05

To generate useful data for next step, building the simulation model for music selection, distributions for thinking time were determined. As the system is used while driving, the data collected in this experiment while the vehicle was parked were excluded. Further, the data were split in half for the subjects in each age and gender group, with half being used to create the model and the other half of the subjects to validate the model (as described in the following chapter). After taking natural logarithm transformation, the thinking times for both young and middle-age groups both appeared normally distributed, but having different means and standard deviations (Figure 4-18). Stepwise regression reveals that age is a significant variable to predict the thinking and response time ($R^2 = 0.1$, $F_{(1, 159)} = 17.468$, $p < 0.001$).

$$\text{Thinking and Response Time} = 5.323 + 2.131 * \text{Age} \qquad (4.17)$$

Where:

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1.



LogNormal Distribution (n=80)
$(\mu,\sigma) = (0.68, 0.09)$
Goodness-of-fit test: $p > 0.08$

a) Young-age Driver

LogNormal Distribution (n=80)
$(\mu,\sigma) = (0.77, 0.18)$
Goodness-of-fit test: $p > 0.05$

b) Middle-age Driver

Figure 4-18 Distribution Fitting of Thinking Time (lnT) and its Parameters. a) Young-age Drivers b) Middle-age Drivers

Time to Utter Command Words and the Desired Music Name

There are three valid commands for subjects to utter to perform music selection tasks using iPhone 4S -- "*Play Album* + album name," "*Play Artist* + artist name," and "*Play songs by* + artist name." The commands are still valid when the subjects do not say "Play" or say "Find/Finding" instead of "Play." For example, subjects can say either "Artist Michael Jackson" or "Find(ing) artist Michael Jackson" and the system can recognize and play the correct music. However, the speech interface would not accept "*Play song* + song title" which will be confusing to people used to selecting music by song title.

Using a stepwise regression, utterance times to select the music while driving were predicted using the following statistically significant equation ($R^2 = 0.4$, $F_{(4, 181)} = 30.223$, $p < 0.001$).

$$\text{Utterance Time (s)} = 0.342 + 0.137 * \text{Syllable} + 0.165 * \text{Word} - 0.221 * \text{Workload} + 0.146 * \text{Age} \tag{4.18}$$

Where:

NSyllable: Number of syllables of the artist name or album name that subjects uttered.

NWord: Number of words of the artist name or album name that subjects uttered.

Workload: Driving workload. Low workload = 0. High workload = 1.

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1.

The observed time to say a syllable was about 0.03 s greater than the value reported by John [87].

ASR Processing Time

After the subjects' utterance, the ASR system needs time to recognize the users input and perform the search. Presumably, the gender, age, whether the information provided is correct or not should affect the ASR processing time. A stepwise regression identifies significant effects of the information provide correctly and age ($R^2 = 0.124$, $F_{(2, 268)} = 18.991$, $p < 0.001$).

$$\text{ASR Processing Time (s)} = 1.711 - 0.49 * \text{InfoCorrect} - 0.079 * \text{Age} \tag{4.19}$$

Where:

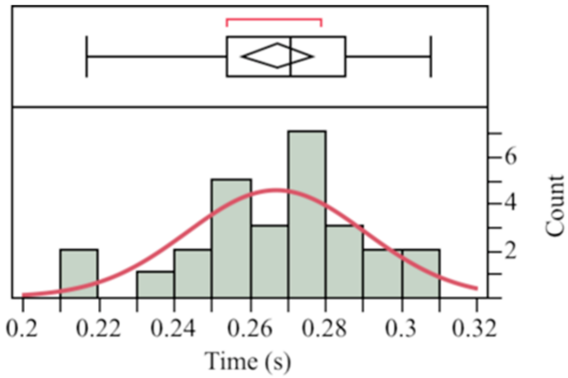
InfoCorrect: The information provided by the subjects. Correct = 1. Incorrect = 0.

When the information provided is correct, the recognition time is 0.49 s less than the time when the information is incorrect.

Machine Feedback and Music Playing

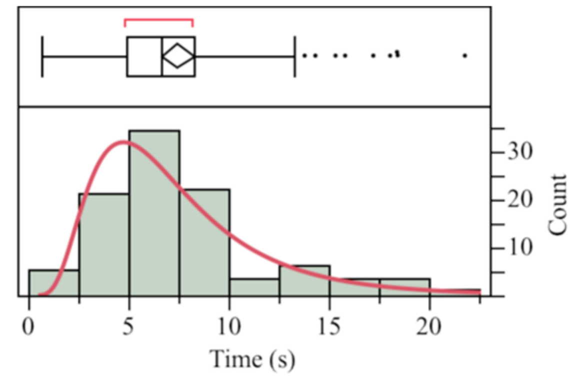
Various machine function times—such as chimes, machine feedback utterance time, the time between the chime and machine feedback prompt time, and music play time were determined by analyzing the audio file recorded when the author using the interface in a quiet environment with limited background noise. The machine feedback starts with a chime, whose duration is fixed (0.4 s).

The duration between the end of chime and the beginning of machine prompt from the empirical data (n=27) follows normal distribution with parameters of $(\mu,\sigma) = (0.267, 0.024)$, $p > 0.686$ (Figure 4-19 a)).



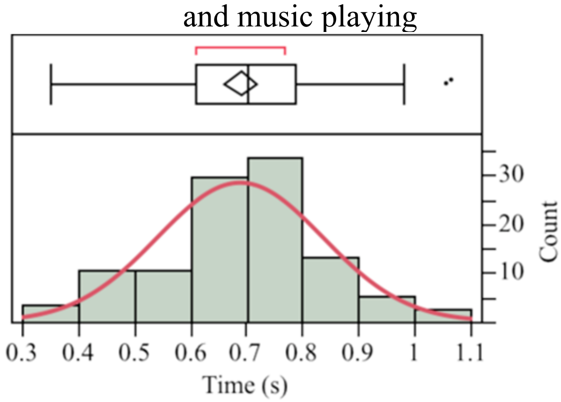
Normal Distribution (n=27)
(μ,σ) = (0.267, 0.024)
Goodness-of-fit test: p > 0.686

a) Time between machine feedback and music playing



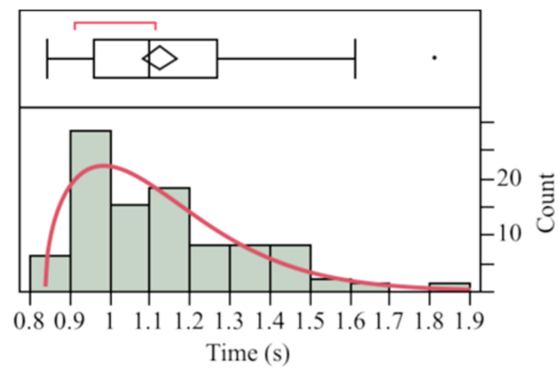
LogNormal Distribution (n=98)
(μ,σ) = (1.863, 0.551)
Goodness-of-fit test: p > 0.07

b) Thinking time 2



Normal Distribution (n=105)
(μ,σ) = (0.690, 0.148)
Goodness-of-fit test: p > 0.262

c) Time to say next track



Weibull Distribution (n=95)
(μ,σ) = (0.690, 0.148)
Goodness-of-fit test: p > 0. 262

d) Machine processing time 2

Figure 4-19. Distributions Fitting from the Data for Half the Subject

The machine feedback prompt time is the time starting with the beginning of machine prompt, and ending with last word of the prompt. For example, machine will provide the feedback by prompting "*Playing Songs by* Michael Jackson" or "*Playing album* History Continue" when the subjects say "*Play Artist* (or songs by) *Michael Jackson*" or "*Playing album* History Continue." Again, individual, driving, or task differences should not affect this machine feedback. The results from stepwise regression show that machine utterance time was predicted using the number of syllables of the prompt and the number of pauses ($R^2 = 0.678$, $F_{(2, 25)} = 26.324$, $p < 0.001$), approximately one-tenth of a second per syllable and per word.

$$\text{Machine Feedback Prompt Time (s)} = 0.559 + 0.091 * \text{NSyllable} + 0.084 * \text{NWord} \quad (4.20)$$

Where:

NSyllable: Number of syllables of the artist name or album name that a system prompted.

NWord: Number of words of the artist name or album name that a system prompted.

When the speech interface provides feedback, subjects immediately recognize whether the speech interface will play the correct selection. Therefore, there is no need to wait for the music to play to confirm their selection and the task completion time will not include the duration between the end of machine's feedback and when the music begins to play. On the other hand, when the subjects try to select a specific song, it is not always the case that the music playing is the music they selected after the first attempt. They need to wait for the music to play to confirm whether it is the music requested by the experimenter. Therefore, the task completion time should include the duration between the end of speech interface's feedback and the beginning of music playing. Also, the duration is song-specific, depending on the MP3 file saved in the database. For example, it always takes 7.6 s to play the first song when searching the songs performing by Sugar Ray or 2.6 s to play the first song when searching the songs performing by Sean Paul.

As described in the previous paragraph, the first song played by the MP3 player is sometimes not the target song. Sometimes, subjects need to listen to a song to determine if it is the target song and press the ASR button to activate the speech system if it is not the target song, which is defined as *Thinking Time 2*. Using the data collected while driving, no variables, such as age, gender, or information given, were found to significantly predict the thinking time 2. However, those times were fit by a LogNormal distribution with parameters of $(\mu,\sigma) = (1.863, 0.551)$, $p > 0.07$ (Figure 4-20 b)).

Subjects had the option of saying the command “Next Track” when the playing song was not the target song and may say it several times to find the correct song. They also may change the selection method. Using the data while driving, the time to say *next track* was predicted using age (young vs. middle-age) and gender ($R^2 = 0.227$, $F_{(2, 69)} = 26.324$, $p < 0.001$). The stepwise regression is

$$\text{NextTrack Time (s)} = 0.429 + 0.131 * \text{Age} + 0.092 * \text{Gender} \tag{4.21}$$

Where:

Age: The age group of subjects. Young subjects = 0. Middle-age subjects = 1.

Gender: The gender group of subjects. Female = 0. Female = 1.

This time was normally distributed with parameters of $(\mu,\sigma) = (0.690, 0.148)$, $p > 0.262$ (Figure 4-19 c)).

After the subjects said the command “next track,” the speech interface ASR needs time to process the prompt and provide feedback (chime) to the subjects. The time from the empirical data (n=95) followed a Weibull distribution with a threshold. The distribution parameters are $(\mu,\sigma) = (0.690, 0.148)$, $p > 0.262$ (Figure 4-19 d)).

Again, the time for the machine to play the chime is fixed and the value is 0.4 s. The time for the MP3 player to play the music is song specific.

The number of songs in each album, or by the same artist, depends on the number of songs stored in a personal MP3 player. This also affects the number of times to say the command "*Next Track*" to find the correct song. Regression analysis did not identify any significant predictors for this time. The mean to say the command "*Next Track*" were 1.55 ± 0.83 and 1.56 ± 0.84 for selecting a specific song using "*Play Album*" (n = 29) and "*Play Artist*" (n = 32), respectively.

B. Turns to Complete the Music Selection Tasks

Table 4-42 shows the total number of turns to complete a music selection task based on the information given and different driving conditions, as well as the number of turns needed by subjects and the MP3 player. Overall, to complete a music selection task required 3.3 ± 2.3 turns, which subjects needed to say the prompt 1.7± 2.3 times and the MP3 players needed to provide feedback 1.6 ± 1.1 times. Apparently, the total number of turns needed to select a specific song (given three pieces of information) was greater than the total number of turns needed to select a specific artist or album on all driving conditions.

Table 4-42. Total Turns Needed to Complete Music Selection Task at Different Driving Conditions (Machine Turn; Subject Turn)

Driving	Information Given	Complete / Give-up	Incorrect Final	Correct Final	
			Incorrect First	Incorrect First	Correct First
Parked	Artist	Complete		3±1(2, 2)	2 (1, 1)
		Give-up			
	Album	Complete		8 (4, 4)	2 (1, 1)
		Give-up			
	Song/Artist/Album	Complete		7±4 (4, 3)	3±2 (2, 2)
		Give-up	6±5 (2, 4)		

Low Workload	Artist	Complete		6±3 (3, 3)	2 (1, 1)
		Give-up			
	Album	Complete		5±1 (3, 3)	2 (1, 1)
		Give-up			
	Song/Artist/Album	Complete	3±1 (2, 2)	7±3 (3, 4)	3±2 (2, 2)
		Give-up	8 (3, 5)		
High Workload	Artist	Complete		4 (2, 2)	2 (1, 1)
		Give-up			
	Album	Complete	2 (1, 1)	4 (2, 2)	2 (1, 1)
		Give-up			
	Song/Artist/Album	Complete	4±3 (2, 2)	6±2 (3, 3)	3±1 (2, 2)
		Give-up	2 (1, 1)		

For those 469 trials correctly selecting the music, there were no main effects of age, gender, driving conditions, or the interactions on the total turns to complete the task. Selecting specific music requires two more turns to complete the task than selecting for a specific artist or album ($F_{(2, 412)} = 11.667$, $p < 0.001$). There was no difference in the total turns required to select a specific artist versus the total turns to select a specific album. When errors occurred during the first attempt of a music selection task, 3.3 more turns were required to complete the task ($F_{(1, 412)} = 42.73$, $p < 0.001$).

C. Music Selection Method

The method used to select music by the subjects is shown in Table 4-43. When given the artist name only and requesting subjects to find the specific artist, subjects used the command "*Play Artist*" 99% of the time to select the specific artist. However, the command "*Play Songs by*" is a valid command that only occurred once when given the artist name. On the other hand, subjects started with the command "*Play Artist*" twice when the given the album name and the machine feedback was incorrect.

Table 4-43. Frequency of Method Used by Subjects on the First Attempt by Information Given and Driving Conditions

Information Given	Driving	First Selection Method			
		Play album	Play artist	Play songs by artist name	Song
Album	Parked	31	1		
	Low Workload	31	1		
	High Workload	32			
Artist	Parked		31	1	
	Low Workload		32		
	High Workload		32		
Song/Album /Artist	Parked	42	45	6	3
	Low Workload	38	47	4	7
	High Workload	50	40	4	2

Of the most interest were the 288 trials with all information given. Although the experimenter already informed the subjects that "*Play Song* + song title" was not a valid method to select a specific song, this method still occurred in 12 trials (4%) in all driving conditions. The probability using the commands "*Play Album*" and using "*Play Artist*" were similar, 45.1% and 45.8%, respectively. The percentages of the methods will be used to build simulation model to predict the drivers' performance on task completion time in the next chapter.

When the subject failed to select the correct music, 37% of the time subjects used the same method to perform the task on the next attempt (Table 4-44). For those trials which "*Play Album*" was the first attempt, 42% of the time subjects switched to the "*Play Artist*" method. For those trials that the method of "*Play Artist*" was the initial attempt, 59% of the time subjects switched to the "*Play Album*" method.

Table 4-44. Frequency of Method Used by Subjects on the Second Attempt for the Music Selection Task While the Result from the First Attempt Was Incorrect

First Selection Method	Second Selection Method		
	Play album	Play artist	Play song
Play album	7	5	0

Play artist	20	14	0
Play songs by artist name	4	0	0
Play song	3	7	2
Total	34	26	2

Subjects completed 83.9% (52/62) of the music selection tasks after the second attempt. There were still 10 trials that ended without selecting the correct music (Table 4-45). When the subject failed to select the correct music, 30% of the time subjects would use the same method to perform the task again, and all trials occurred when the subjects used the command "*play artist*" for the previous attempt.

Table 4-45. Frequency of Method Used by Subjects on the Third Attempt for the Music Selection Task While the Result from the Second Attempt Was Incorrect

Second Selection Method	Third Selection Method			
	Play album	Play artist	Play songs by artist name	Play song
Play album	0	1	1	0
Play artist	2	3	0	1
Play song	1	1	0	0
Total	3	5	1	1

Subjects completed 80% (8/10) of the music selection tasks after the third attempt. All trials ended with selection of the correct music (Table 4-46). When the subject failed to select the correct music, there was a 50% chance that the subjects would use the same method to perform the task on the next attempt. After four attempts to select the music, all subjects completed the music selection tasks.

Table 4-46. Frequency of Method Used by Subjects on the Fourth Attempt for the Music Selection Task While the Result from the Third Attempt Was Incorrect

Third Selection Method	Fourth Selection Method	
	Play album	Play artist
Play album	1	0
Play song	0	1
Total	1	1

D. Errors and Correction Strategies

Frequency of Error

Although the information given was always correct, and the specific relationships between the information and the artist name, album name, or song title were presented on the screen, there were 97 trials (20%) with at least one error occurring among the total of 480 trials. For example, the presented information on the screen was "Album: Life for Rent," and subjects immediately recognized that they were requested to select the album *Life for Rent* and play the music. Overall, the total numbers of errors that occurred either from subjects or from the MP3 player was 154, with mean and standard deviation of 1.6 and 0.9, respectively (Table 4-47). The total number of errors that occurred was significantly affected by the age group and the number of information items given (Table 4-48). The mean of total errors that occurred among middle-age subjects was 0.3 times greater than the young subjects when selecting a music, 0.4 and 0.1 times, respectively ($F_{(1, 456)} = 21.016$, $p < 0.001$). When the subjects were requested to select a specific song (given three pieces of information), the mean of total errors was 0.3 times greater than the mean of total error when the subjects were requested to select a specific album or artist, 0.4 and 0.1 times, respectively ($F_{(1, 456)} = 21.966$, $p < 0.001$).

Table 4-47. Descriptive Statistics of Total Errors by Driving Conditions and Information Given

Driving	Information Given	Number of Subjects	Sum of Errors	Mean ± Std. Dev.	Min. – Max.
Parked	Album	1	3	3	3
	Artist	6	7	1.2 ± 0.4	1 – 2
	Song/Album/Artist	14	31	2.2 ± 1.4	1 – 6
	Song/Album/Artist	4	6	1.5 ± 0.6	1 – 2
	Song/Album/Artist	6	6	1.0	1
Low Workload	Album	3	4	1.3 ± 0.6	1 – 2
	Artist	2	5	2.5 ± 0.7	2 – 3
	Song/Album/Artist	12	20	1.7 ± 1.1	1 – 4

	Song/Album/Artist	7	10	1.4 ± 0.8	1 – 3
	Song/Album/Artist	14	26	1.9 ± 0.8	1 – 3
High Workload	Album	3	4	1.3 ± 0.6	1 – 2
	Artist	2	2	1.0	1
	Song/Album/Artist	8	9	1.1 ± 0.4	1 – 2
	Song/Album/Artist	8	9	1.1 ± 0.4	1 – 2
	Song/Album/Artist	7	12	1.7 ± 1.0	1 – 3
Total		97	154	1.6 ± 0.9	1 – 6

Table 4-48. ANOVA of Total Number of Errors Occurred by Age, Gender, Driving Conditions, and Number of Information Given

Source	df	F	p-value
Age (A)	1	21.016	**<0.001***
Gender (G)	1	0.032	0.858
Driving (D)	2	1.867	0.156
Numb of Information Given (I)	1	21.966	**<0.001***
A * G	1	0.998	0.318
A * D	2	0.611	0.543
A * I	1	2.771	0.097
G * D	2	0.520	0.595
G * I	1	0.111	0.739
D * I	2	0.946	0.389
A * G * D	2	0.336	0.714
A * G * I	1	0.289	0.591
A * D * I	2	0.052	0.950
G * D * I	2	0.103	0.902
A * G * D * I	2	0.124	0.883
Error	456		

Note: * - Statistically significant, $p < 0.05$

Table 4-49 shows that a total of 9 machine errors occurred among the 8 trials with a mean of 1.1 times. The ASR error rate was 1.7% (8/480), which was low. The error occurred only once when a subject attempted to select a specific album while the vehicle was parked. Although the data shows that errors occurred when subjects selected music while driving, there was no statistically significant difference.

Table 4-49. Descriptive Statistics of Machine Errors by Driving Conditions and Information Given

Driving	Information Given	Number of Subjects	Sum of Errors	Mean ± Std. Dev.	Min. – Max.
Parked	Album	1	1	1	1
	Artist				
	Song/Album/Artist				
	Song/Album/Artist				
	Song/Album/Artist				
Low Workload	Album				
	Artist				
	Song/Album/Artist				
	Song/Album/Artist	1	2	2	2
	Song/Album/Artist	2	2	1	1
High Workload	Album				
	Artist				
	Song/Album/Artist				
	Song/Album/Artist	3	3	1	1
	Song/Album/Artist	1	1	1	1
Total		8	9	1.1 ± 0.4	1 – 2

Apparently, 94% (145/154) of the errors occurred from 27 subjects. The number of errors that occurred was significantly affected by the age group and the number of information items given (Table 4-50 and Table 4-51). The mean of total errors that occurred among middle-age subjects was 0.3 times greater than that of the young subjects when selecting a music, 0.1 and 0.4 times, respectively ($F_{(1, 456)} = 21.468$, $p < 0.001$). When the subjects were requested to select a specific song (given three pieces of information), the mean of total errors was 0.3 times greater than the mean of total error when the subjects were requested to select a specific album or artist, 0.4 and 0.1 times, respectively ($F_{(1, 456)} = 20.029$, $p < 0.001$).

Table 4-50. Descriptive Statistics of Human Errors by Driving Conditions and Information Given

Driving	Information Given	Number of Trials	Sum of Errors	Mean ± Std. Dev.	Min. – Max.
Parked	Album	1	2	2	2
	Artist	6	7	1.2 ± 0.4	1 – 2
	Song/Album/Artist	14	31	2.2 ± 1.4	1 – 6
	Song/Album/Artist	4	6	1.5 ± 0.6	1 – 2
	Song/Album/Artist	6	6	1.0	1
Low Workload	Album	3	4	1.3 ± 0.6	1 – 2
	Artist	2	5	2.5 ± 0.7	2 – 3
	Song/Album/Artist	12	20	1.7 ± 1.1	1 – 4
	Song/Album/Artist	7	8	1.1 ± 0.4	1 – 2
	Song/Album/Artist	12	24	1.9 ± 1.0	1 – 3
High Workload	Album	3	4	1.3 ± 0.6	1 – 2
	Artist	2	2	1.0	1
	Song/Album/Artist	8	9	1.1 ± 0.4	1 – 2
	Song/Album/Artist	5	6	1.1 ± 0.4	1 – 2
	Song/Album/Artist	7	11	1.6 ± 1.0	1 – 3
Total		92	145	1.6 ± 0.9	1 – 6

Table 4-51. ANOVA of Number of Human Errors Occurred by Age, Gender, Driving Conditions, and Number of Information Given

Source	df	F	p-value
Age (A)	1	10.756	**<0.001***
Gender (G)	1	0.001	0.958
Driving (D)	2	1.023	0.131
Numb of Information Given (I)	1	10.035	**<0.001***
A * G	1	0.272	0.461
A * D	2	0.282	0.570
A * I	1	1.606	0.074
G * D	2	0.205	0.665
G * I	1	0.001	0.958
D * I	2	0.486	0.380
A * G * D	2	0.192	0.682
A * G * I	1	0.089	0.674
A * D * I	2	0.044	0.916
G * D * I	2	0.067	0.875
A * G * D * I	2	0.051	0.959
Error	456		

Note: * - Statistically significant, $p < 0.05$

Error Type and User Correction Strategy

Table 4-52 shows the frequency of two types of machine errors– "cannot recognize the user's prompt" or "misrecognized the user's prompt." The error of "speech interface misrecognized the user's prompt" is defined as the speech interface provides incorrect feedback when the user's input is correct. All 7 errors occurred when the subjects performed the music selection tasks while driving. Subjects repeated the utterance or repeated the utterance slower to correct the errors were more than 50% of the time. The error "cannot recognize the user's utterance," occurred one time when the subject said the command "next track" and another time was when the subject said "*play album Lost and Found.*" Both of the cases resulted in the incorrect feedback from the machine and subjects repeated the utterance to correct the errors.

Table 4-52. Frequency of Machine Error by Type and User Correction Strategies for the Music Selection Task

Error Type (Frequency)	Example	User Correction Strategy	Frequency
Misrecognized the user's utterance (7)	S: Play album This Way. M: Playing album Life for Rent	Repeat the utterance	2
		Repeat the utterance slower	2
		Change the pronunciation	1
		Change the selection method	1
		Did not notice the error and say the command word "next track"	1
Cannot recognize the user's utterance (2)	S: Next Track. Machine response was playing the same song from previous attempt.	Repeat the utterance	2

Human errors related to information provided by the subjects and user correction strategies are shown in Table 4-53. The errors related to information provided by subjects account for 51% (74/145) of the total human errors. Subject uttered the information before the time that the MP3 player could accept and process the signal (barge-in), which resulted in error feedback from machine. Sixty-one percent of the errors (14/23) occurred while driving, and 87% of the time (20/23) subjects repeated the prompt to correct the error. Another category of error was when a subject either said the utterance too late or did not provide information (time out). Fifty-six percent (10/18) of the errors occurred while driving, and half of the time that subjects repeat the utterance to correct the errors. Another common mistake made by the subject was when they did not press the button to activate the ASR before saying the utterance and, surprisingly, this kind of error occurred in 74% of the trials (10/14) when selecting the music was the primary task (no driving). The subjects stammered (play play uh album) when saying the information 12 times, and 67% of those trials occurred while driving. Subjects still used the same method and provided the information to correct the errors.

Table 4-53. Types and Frequency of Human Errors Relevant to Information Provided and User Correction Strategy for the Music Selection Task

Error Type (Frequency)	Example	User Correction Strategy	Frequency
Barge-in (23)		Repeat the prompt	20
		Change selection method	1
		Rephrase the prompt	1
		No response	1
Time out (18)		Use the same selection method and repeat the prompt	9
		Provide correct information	3
		Say invalid command (go back)	2
		to correct	2
		Repeat the prompt	
		Change selection method to Artist	1
		Use the same selection method and provide correct information	1
Did not press the button to activate the ASR (14)		User press the button to activate the ASR	13
		Repeat the prompt but still did not press the button to activate the ASR	1
Stammered the prompt or command (12)	"Play Play uh Album"	Use the same selection method and provide incorrect information	3
		Use the same selection method	2

		and provide correct information (Machine correct the error automatically and provide the correct information)	 7
Provided incorrect information (4)	"Play artist Sean **Paul**," instead of "Play artist "Sean **Pen**"	Use the same selection method and provide correct information (Machine correct the error automatically 1and provide the correct information)	1 3
Provided incomplete information (2)	"Play album **Still** Not Getting Any," instead of "Not Getting Any"	Change the selection method to Artist (Machine correct the error automatically 1and provide the correct information)	1 1
Repeated the prompt when the machine still searched (1)		(Machine recognize the prompt and provide correct feedback)	1

The frequency of different types of human errors relevant to each selection method and command and subjects' correction strategies is shown in Table 4-54. The errors relate to selection method and command accounts 39% (57/145) of the total human errors. Although there were no mistakes made by the subjects and machine for those 23 trials where subject changed the selection method, it was still counted as an error. Most of the time, the scenarios occurred when the subject used artist as the method to select the music and learned that there were too many songs by the same artist in the database. Therefore, they changed the selection method to album.

Table 4-54. Frequency of Types of Human Errors Relevant to Selection Method and Command and User Correction Strategy for Music Selection Task

Error Type (Frequency)	Example	User Correction Strategy	Frequency
Changed selection method (23)		Change selection method from artist to album Change selection method from album to artist	20 3
Used invalid selection method (15)	"Play song Can You Do the Work"	Change selection method to artist Change selection method to album Repeat the prompt (use the same	9 4

		method)	2
Forgot to say command word (8)	“Play **artist** Juliana Theory,” instead of “Play Juliana Theory”	Say the command word Repeat the prompt Repeat the prompt slower Change selection method (Machine correct the error automatically and provide the correct information)	4 1 1 1 1
Said invalid format of command (4)	“Play songs by Sean Paul,” instead of “Play songs by Artist Sean Paul”	Change selection method Say the correct format of command	3 1
Used wrong command (4)	Play **album** Lost and Found,” instead of “Play **artist** List and Found”	Repeat the prompt Say the correct command word	2 2
Said invalid command (3)	“Previous Track,” instead of “Go Back”	Change selection method Use the same method and repeat the prompt	2 1

Also, there were 15 trials (6 subjects - 1 young, 5 middle-age; 4 males, 2 females) in which subjects directly used song title to select music. Eleven of those 15 trials, involving using the song title, occurred while driving. Subjects often used the wrong command. For example, the information displayed was “Album: Lost and Found” and the subject say “Play Artist Lost and Found.”

There were 14 errors relevant to the subject’s knowledge (Table 4-55). Eleven of those 14 errors occurred when the subjects ended up with selecting the wrong music, and 9 out the 11 trials occurred while driving. Also 10 of the 11 trials ended with selecting incorrect music were when the subjects requested to select a specific song. To ensure the MP3 player was playing the correct song, subjects can listen to the music playing if they were familiar with the specific song, or they can look at the screen and check the information provided by the MP3 player. However, the MP3 player was placed in the center of steering wheel and that was not the same location of the manufacturer’s built-in system. This may affect the driving performance if drivers needed to look down to check the information. Also the fonts used for feedback shown on the MP3's screen were small and this too could affect the driving performance. Therefore, the subjects assumed the music played by the MP3 player was correct.

Table 4-55. Frequency of Other Types of Human Errors and User Correction Strategy for the Music Selection Task

Error Type (Frequency)	User Correction Strategy	Frequency
Select incorrect music (11)		11
Did not know how to go back to playing music screen when the iPhone screen change to home screen (2)	Ask the experimenter for help	2
Switched to manual selection (1)		1

4.4 Conclusions

To develop a simulation model to predict driving performance while using in-vehicle speech-controlled destination entry and music selection systems, one needs to know step by step how subjects would enter a destination or select music, the precise duration of each step, the type and frequency of each error, and the correction strategy for each error, all topics examined in this experiment. The task completion rate in this study was 97% for destination entry tasks and 99% for music selection tasks, respectively. The task completion time for those trials without any errors was 77 s for destination entry tasks and 19 s for music selection tasks. When errors occurred, the mean task completion time significantly increased at least 1.7 times.

Entering an address using a command-based speech system is a complex task and involves at least 15 turns for both subjects and a system to complete the task without any errors. When errors occurred, the total number of turns to complete the destination entry task almost doubled.

More specifically, the goal of this chapter is to address five questions. The first two questions were:

1. How long do drivers need to think of and utter commands and phrases?

2. How long does it take to say those commands or phrases?

The detailed steps and the equations or distributions used to predict the thinking and utterance times are described in the Table 4-56. The predictive parameters, if any, and the distribution that fit best, if there were no predictive parameters, varied between subtasks. For single word responses, utterance time was usually predicted by the number of syllables to be uttered. For multiple word responses, the number of words was a predictor. As a first cut, rough rule of thumb, allowing for approximately 0.5 s per word and 0.2 s per syllable seems reasonable.

Table 4-56. Subtask Time Predictions for the Destination Entry Task While Driving

Subtask	Predicted Equations or Distribution (times in seconds)
S: Thinking and processing time	Address related to personal: T = Lognormal (1.48, 0.57) Address not related to personal: T = Lognormal (1.31, 0.61)
S: Command utterance time	S_CMD T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
S: Think and utter the state name	S_Think_State T = Lognormal (-0.5, 0.49) S_State T = - 0.212 + 0.242*NSyllable + 0.098*Age – 0.140*Gender
S: Think and utter the city name	S_Think_City T = Normal (0.91, 0.78) S_City T = -0.103 + 0.172*NSyllable + 0.089* Age + 0.149*NWord
S: Think and utter the street name	S_Think_Street T = Lognormal (0.03, 0.58) S_Street T = 0.041 + 0.459*NWord + 0.188*Age – 0.103*Gender
S: Think and utter the house number	S_Think_House 1/T = Weibull (1.31, 2.42) S_House T = -0.805 + 0.681*NWord + 0.201*Age – 0.143*Workload – 0.197*Gender
S: Spell the word	Mean number of characters = Poisson (7.91) S_Spell T = -1.826 + 0.635*NCharacter + 1.09*NWord
S: Barge-in	Barge-in T = Normal (0.19, 0.11)
S: Time out	Time-Out T = Normal (7.12, 1.8)
M: Command prompt time	M_CMD T = 0.347 + 0.284*NWord
M: Process and prompt the state name	M_Proc_State T = 2.185 + 0.89*MultipleChoice M_State T = 0.117 + 0.163*NSyllable + 0.117*NWord
M: Process and prompt the city name	M_Proc_City T = 2.241 + 0.95*MultipleChoice M_City T = 0.211+ 0.212*NSyllable
M: Process and prompt the street name	M_Proc_Street T = 2.772 + 0.782*MultipleChoice M_Street T = 0.448+ 0.194*NSyllable

M: Process and prompt the house number	M_Proc_HouseNum T = Weibull (0.37, 2.44, 2.71) M_HouseNum T = 0.169+ 0.553*NWord
M: Route Processing time	M_Route Processing time T = 3.182 + 7.936 * Out_State

On the other hand, selecting music using speech was much easier than entering a destination and required only four total turns for both subjects and a system to complete the task without any errors. When errors occurred, the total turns to complete the music selection task almost doubled. The detailed steps and the equations or distributions used to predict the thinking and utterance times are shown in Table 4-57. As with the destination entry task, the predictive parameters were again number of syllables, number of words, and gender, but the combination that provided the best predictions, if there was one at all, varied with the subtask.

Table 4-57. Subtask Time Predictions for the Music Selection Task while Driving

Subtask	Predicted Equations or Distribution (times in seconds)
S: Thinking and processing time	S_Think T = 5.323 + 2.131*Age
S: Command and information utterance time	S_Utt T = 0.341 + 0.137*NSyllable + 0.165*Word - 0.221*Workload + 0.146*Age
S: Thinking_2	S_Think_2 T = Lognormal (1.863, 0.551)
S: Next Track	S_NextTrack T = 0.429 + 0.131*Age + 0.092*Gender
S: Number to of time say next track	For *Play Album*: 1.55 ± 0.83 For *Play Artist*: 1.56 ± 0.84
M: Process 1	M_Proc1 T = 1.711 – 0.49*Correct_MS – 0.079*Age
M: Feedback and music playing	Chime T = 0.4 s Pause T = Normal (0.267, 0.024) M_Utt T = 0.559 + 0.091* NSyllable + 0.084 * NWord Music playing is song specific
M: Process 2	M_Proc2 T = Weibull (0.69, 0.148) Chime T = 0.4 s Music playing is song specific

With regard to errors, there were two questions to address:

1. What are the types of errors that drivers make, and how often do they occur?
2. For each system response to an error, what is the driver's correction strategy?

Using *street address* (88%) at the first attempt to enter the address was the most frequent method for destination entry tasks, followed by *address book* (7%). When errors occurred, subjects preferred to use the *street address* method to correct the error. For music selection task, the probability of using *album* or *artist* to select a specific song was similar, 45.1 % and 45.8%. However, for 4% of the trials, subjects tried to use *song title* to select the specific song, which is not accepted by the iPhone 4S. When subjects failed to select the correct music on the first attempt, 37% of the time subjects used the same method for the second attempt. There was a 42% chance that subjects would switch from *play album* to *play artist*. When subjects used *play artist* on the first attempt, 59% of the time subjects switched to *play album* method.

Errors can be categorized into three groups: (1) information relevant, (2) system commands and entry method relevant, and (3) subjects' knowledge related to the system. Grice's conversation maxims and the error typology proposed by the Veronis [54], as well as the turn-taking and common grounding, can partially explain the errors that occurred during the experiment.

There were 1,088 errors distributed among 323 trials for the destination entry task. When given incorrect information, the mean number of errors was 3 times greater than the number of errors that occurred with correct information. Time-out and barge-in account for one-third of the human errors and the correction strategy *repeat the prompt/repeat the prompt slower* was the command method most used to correct the errors.

Although the music selection task was relatively easier to do than the destination entry task, 154 errors occurred. Barge-in and time out were still the two major errors for the music selection task. Repeat the prompt was the most frequently used method by subjects to correct these two types of errors.

Finally, with regard to driving performance, there was one question to address:

How is performance of the speech task affected by the level of driving task workload?

Entering destinations while driving significantly affected the drivers' performance for 7 of the 10 commonly used driving performance measures.

Driving workload affected the destination entry tasks when the given information was correct and no errors occurred. Task completion time while the vehicle was parked was 6 s and 7 s greater than the time while driving in the low and high workload conditions, and because of learning (as they occurred after the parked condition). Task completion time while driving is expected to be greater than while parked. There was no difference in task completion time during the low- and high-workload conditions. As was shown in Table 4-55, driving workload did not significantly affect thinking and utterance times.

Driving workload, however, did affect task performance in the music selection tasks. The task completion time for subjects driving in the low-workload condition was greater than when the vehicle was parked or while driving in the high-workload condition. If anything, one would expected task completion time to increase with workload.

Good design of a user interface in a vehicle could potentially minimize the driver distraction as indicated by these measures. To predict speech performance before the system is developed or revised rather than evaluation of a completed system costs less time and money. The present study provides important data that can be used to build a simulation model to predict the speech interface performance, a topic addressed in the next chapter.

CHAPTER 5

A Simulation Model to Predict the User Task Completion Time and Errors

5.1 Introduction

Simulation is the representation of a real world or artificial process over time [89, 90]. Simulations can be physical, where information is presented in real time with which a person interacts with, such as driving simulators, flight simulators, patient and surgical simulators, or nuclear power plant simulators [91-94]. Those simulators are used to train operators, often for situations that are risky, or on systems that are more expensive than the simulation to operate, or to collect data on how people function when using such devices. When people think of simulation, this is the use of simulation that often comes to mind and it was this type of simulation that was utilized in the research described in chapter 4.

Simulations can also be purely computational, such as those used to model financial systems, the weather, the spread of disease, and disaster preparedness, though there may be a visual interface for users such as the JACK biomechanical model [95-99]. See also

[100-102]. The underlying mathematics can be either discrete or continuous [103, 104]. An autopilot is a good example of continuous control. Discrete control involves a series of events, such as having a person to do a task, that has well defined start and end times, but may take time to complete. The simulation in this chapter is discrete control.

Within the human performance literature, which served as a basis for this chapter, there is a long tradition for using computational simulation to address problems [105, 106]. Computational simulation has been the most successful for complex problems, in particular where there are many tasks for the user or users to perform and the task durations are not a single value, but have distributions. Usually, each task can lead to one or more following tasks and which task follows is probabilistic.

An example of this kind of problem of interest to the U.S. Army is how many soldiers are needed to man a tank and how crew size affects the time to fire. In tank warfare, usually only the tank that fires first survives. A typical scenario involves a tank driving down a road (what the Army technically calls a road march) with the hatches open and the crew's heads outside the tank searching for the enemy [107].

Within the human factors domain, discrete event simulation models have been used very effectively to predict the performance of complex military systems [108, 109]. Historically, the human factors community uses either MicroSaint (more recently MicroSaint Sharp) or IMPRINT for those simulations, primary because MicroSaint and IMPRINT have features that are particularly useful for human performance simulation [110-112]. MicroSaint Sharp is a product of Micro Analysis and Design, now a subsidiary of Alion Science and Technologies. MicroSaint Sharp costs several thousand dollars per copy. IMPRINT, developed for the Army, is free to those in the U.S., and is used by the Army and for Army-funded projects. There are, however, many other alternative applications, with Promodel, Witness, Simula, Arena, and for historical reasons GPSS (http://en.wikipedia.org/wiki/List_of_discrete_event_simulation_software) being the best known candidates[113].

The tasks of using an in-vehicle infotainment system to enter destinations and select songs have many of the same fundamental elements as the tank problem. Figure 5-1 shows the example of entering a street address as a destination. The first subtask of which is to enter the name of the state. After the system asks for the state name, the driver could respond by saying “Michigan.” The car might recognize the entry as Minnesota (so the driver needs to correct the car), correctly identify Michigan (so they go to next task, saying a city name), or fail to detect the speech and ask the driver to say again (so they say “Michigan” again).

In the network, the performance of each subtask (e.g. completion times, path followed) is by rules for the subnetwork programmed by the simulation developer, including conditions, probabilities, and constraints) The complicating factor in many simulations is that task sequences can loop back and repeat, in this example, to repeat an unrecognized utterance or correct an error.

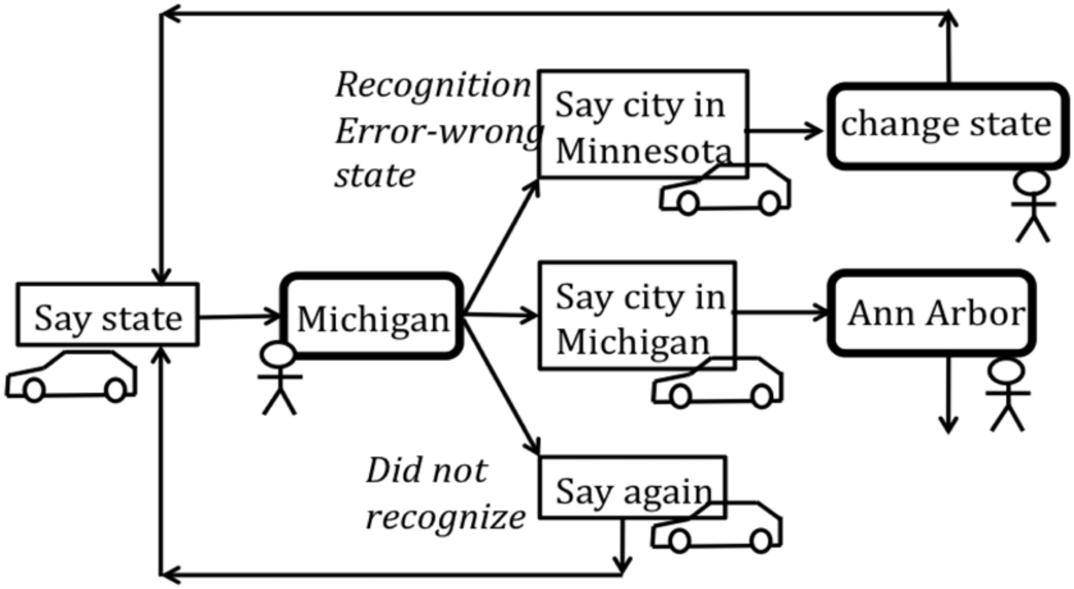


Figure 5-1. The task network for entering a state using speech

To predict system performance, data are needed on the probability of each branch in the network and the time distribution for each utterance (along with delays to recognize what is said, not shown), all for real drivers doing real tasks. Also needed are data on how drivers deal with errors. All of this information was collected in the prior chapter.

The purpose of this research was to provide a model structure (the tasks users perform and their sequence for various error contingencies) and data (either distributions of task time or prediction of them from various tasks variables, as well as estimates of error probabilities) for use with existing simulation software packages to predict user performance with speech interfaces in motor vehicles.

5.2 Method and Assumption

5.2.1 Assumptions

1. The survey data provides reasonable estimates on the frequency of destinations selected and the methods used to select them. The survey also provides related information for songs.
2. In real driving, the entry methods used and their frequency for destinations and songs will be the same as those in the driving simulator experiment.
3. The destination entry method selected depends upon if the destination is a residential or non-residential address.

The results from the simulator study showed that the destination entry method for residential address (home, friend's/relative's address) was different from the entry method of non-residential address. Clearly, the frequency using the method *Destination by Address Book* for residential address was greater than the frequency of non-residential address (0). Also, the *POI* method will not be used to select a residential address.

4. The model is driven by the number of attempts to enter the destination.

The number of attempts is defined as the number of times when the subjects change the entry method from the current method used to other selection method or the subjects use the command *go back* to reach the top level of the current entry method with the options of selecting other entry method or using the same method. As the speech interface used in previous experiment was a command-based navigation system, the cases will be too complicated to categorize if the definition of each attempt was based on the number of times that correction commands are uttered, such as *go back* or *correct*.

5. The probability of a method selected is independent of the selection for the previous attempt.

Similar to the reason for previous assumption, creating a simulation model that included conditional probabilities was beyond the scope of this initial effort.

6. Subjects will give up (balk) after five attempts.

The assumption refers to the number of attempts after changing the entry method. Although balking only occurred for five trials and only for some tasks and methods, the mean number of attempts before balking was five times. Although this is based upon limited data, it is the best (and only) available data.

7. The task completion frequency for each attempt for each method can be estimated from the results of the simulator experiment.

To represent the system used in the simulator experiment, the task completion frequency for each attempt of method was based on the results from the previous experiment.

8. The subtask time can be estimated from the results of the simulator experiment.

The estimated durations uttering the commands or phrases for subjects and systems were based on the results from the simulator study shown in Table 4-55 for navigation and Table 4-56 for music selection.

9. The probability of subtasks will occur is same as to those in the simulator experiment.

The probability of each sequential network was estimated based on the results from the simulator study.

10. The probability of each type of error occurs will be the same as those in the simulator experiment.

The probability error of human error occurrence was calculated by dividing the number of total human errors to the total turns from the subjects (3255) in the simulator studies and shown in Table 5-1. Also among the types of errors, user timeout and barge-in should occur for each subject's utterance after the system asks the information or command. Also, not all the types of errors will be in the simulation, to avoid any overly complicated subnetworks. In particular, rare errors ($p < 0.005$) will be ignored.

Table 5-1. Probability of Human Error Occurrences for Destination Entry Tasks Resulted Form the Simulator Study.

Category	Error Type	Frequency	Probability
Information Relevant	Time out	218	0.0670
	Barge-in	121	0.0372
	Stammer the prompt / command	38	0.0117
	Provide incorrect information	264	0.0811
	Provide incomplete information	71	0.0218
	Provide invalid information	28	0.0086
	Provide invalid format of information	75	0.0230
	Forgot to change the state name	10	0.0031
	Pick the wrong choice	1	0.0003
	Miss the correct information	1	0.0003
	Ask experimenter for confirmation	1	0.0003
	Say unnecessary words	6	0.0018
	Murmur	4	0.0012
	Fail to find the correct information	10	0.0031
Command and Entry Method Relevant	Cannot find the information by specific method	33	0.0009
	Change Entry method	3	0.0009
	Cannot determine the entry method	3	0.0025
	Forget to say command word	8	0.0025
	Say incorrect command	55	0.0169
	Say invalid command	25	0.0077
Subject's Knowledg e Relevant	Did not know the system reach the first step of the entry method	13	0.0040
	Did not know the ASR function has been deactivated	3	0.0009
	Give-up	8	0.0025

Table 5-2 shows the probability of each type of human error for music selection tasks over the total of 401 turns. Unlike the navigation system, the probability that subjects said invalid commands was low. However, the model still includes some types of errors as the subnetwork is relative simple and easier to build.

Table 5-2. Probability of Human Error Occurrences for Music Selection Tasks Resulted form the Simulator Study

Error Type (Frequency)	Frequency	Probability
Changed selection method	23	0.0573
Used invalid selection method	15	0.0374
Forgot to say command word	8	0.0200
Said invalid format of command	4	0.0100
Used wrong command	4	0.0100
Said invalid command	3	0.0075

5.2.2 Software

The software – IMPRINT (The IMproved Performance Research INtegration Tool) readily accessible to UMTRI Driver Interface Group, was used to build the simulation. The advantage of using this software is that physical and cognitive workload estimated for drivers can be obtained from the results. Therefore, information can be used in the evaluation of the driver performance when operating the navigation and music selection systems. On the other hand, the software requires that each task needs to be described in considerable detail, a very time consuming activity. The amount of detail required is much greater than available for traditional task analysis.

5.2.3 Model Creation

To create a discrete event simulation in IMPRINT, one first creates a high level flow chart of the tasks to be performed in the IMPRINT Pro graphical editor. For example, for music selection, this involves creating start and end nodes, and graphics to represent each of the three alternative subtasks (artist, album, song). As a reminder, these represent three different methods to find a song.

Next, for each task/subtask, one creates a more detailed description, in IMPRINT jargon, a subnetwork. So for example, to find a song by finding an album, one would create a step-by-step description of how that occurs, graphically representing each step as a node. One might think of this as being akin to writing a subroutine, which often calls other subroutines. This hierarchical structure makes the model easier to understand.

At some point, one needs to populate the model. For each node in the network, one needs to identify the conditions that cause it to begin and end/its duration. This is done using scripts written in C++. In addition, one needs to specify successive nodes, if there are any. If there are, then the completion of a node can cause the simulation to end, or lead several following tasks to start all at the same time, or one of several tasks, where each task has a probability.

After the model is populated, one runs it a number of times to build up data on network completion times and the probability various paths are taken. Those data are then examined to determine what does not make sense, and the simulation is revised and re-run. For example, if several tasks with similar probabilities were following a task, but the same following task occurred over hundreds of runs, one would suspect a programming error. The kinds of programming errors that commonly occur, mismatched or misplaced parentheses, missing semicolons, misspelled variable names, etc., all occur here.

After one has confidence in the model, then one may run it several hundred times to obtain the desired statistical distributions. Keep in mind that each run uses a different random seed to generate the outcome, and that it takes many runs to develop statistically representative results. Depending on the complexity of the discrete event simulations, a single run can take anywhere from milliseconds to a few seconds, and for huge models much longer. However, IMPRINT has a feature where the analyst sets the number of runs, so requesting 100 runs is just a matter of typing 100 in an entry box.

5.2.4 Validation

In the driving simulator study, there were 48 subjects recruited with equal number in age (3) and gender (2) groups. The data from half of subjects (24) was analyzed to obtain to build the simulation. The data from another half of the subjects (24) was used to validate the model by comparing the tasks completion time and error frequency.

5.3 Results and Discussion

5.3.1 Model for Navigation System

1. Overall Structure of the navigation systems.

The network diagrams are shown in Figure 5-2 to 5-8, the detail description of each subtask and node is in Appendix A. As described in the assumptions, the top level of the destination entry model was divided into two categories of the addresses – residential or non-residential address. The probability used in this simulation was based on the probability of trials in the simulator experiment, which the probability was 0.25 for residential address and 0.75 for non-residential address

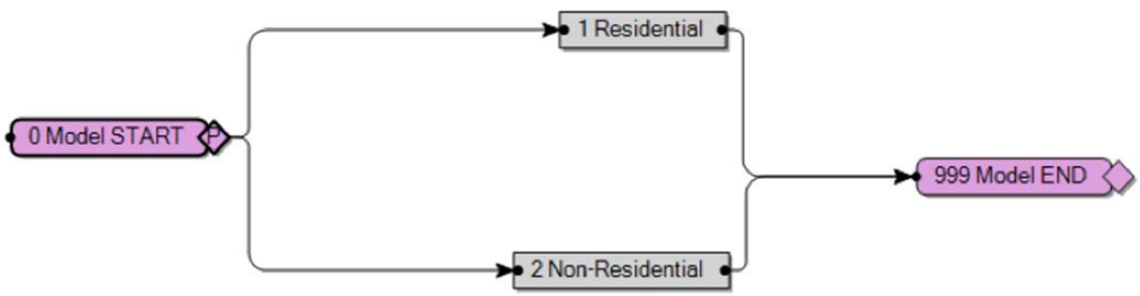


Figure 5-2. Top Level of the Simulation for Destination Entry Tasks.

2. The second level of the model is driven by the method used for destination entry task. The flow for residential and non-residential addresses is shown in Figures 5-3 and 5-4.

Subnetworks are presented in two blocks – one that counts the number of attempts using each method only. The second is that each particular method is independent to the previous attempt. The model using conditional probabilities was beyond the scope and is not implemented in this version of the model. As described earlier, the maximum number of attempts in this model is limited to five times. After the fifth attempt without success, the subjects will balk.

The methods used by subjects are *address book*, *street address*, and *previous destination* for the residential address and *street address*, *previous destination*, and *point of interest* for non-residential address. No network was built for the *address book* method for non-residential address because no subjects used this method in the driving simulator experiment. In the future, the users can add the link and change the probability of each subnetwork to simulate the use of the address book method for the non-residential addresses.

When successfully completing the task, simulation goes to the end node, ending a run. When the attempt failed to find the destination, the flow goes to the try-count node to determine the flow to the next run and repeats again until either the task succeeds or balks.

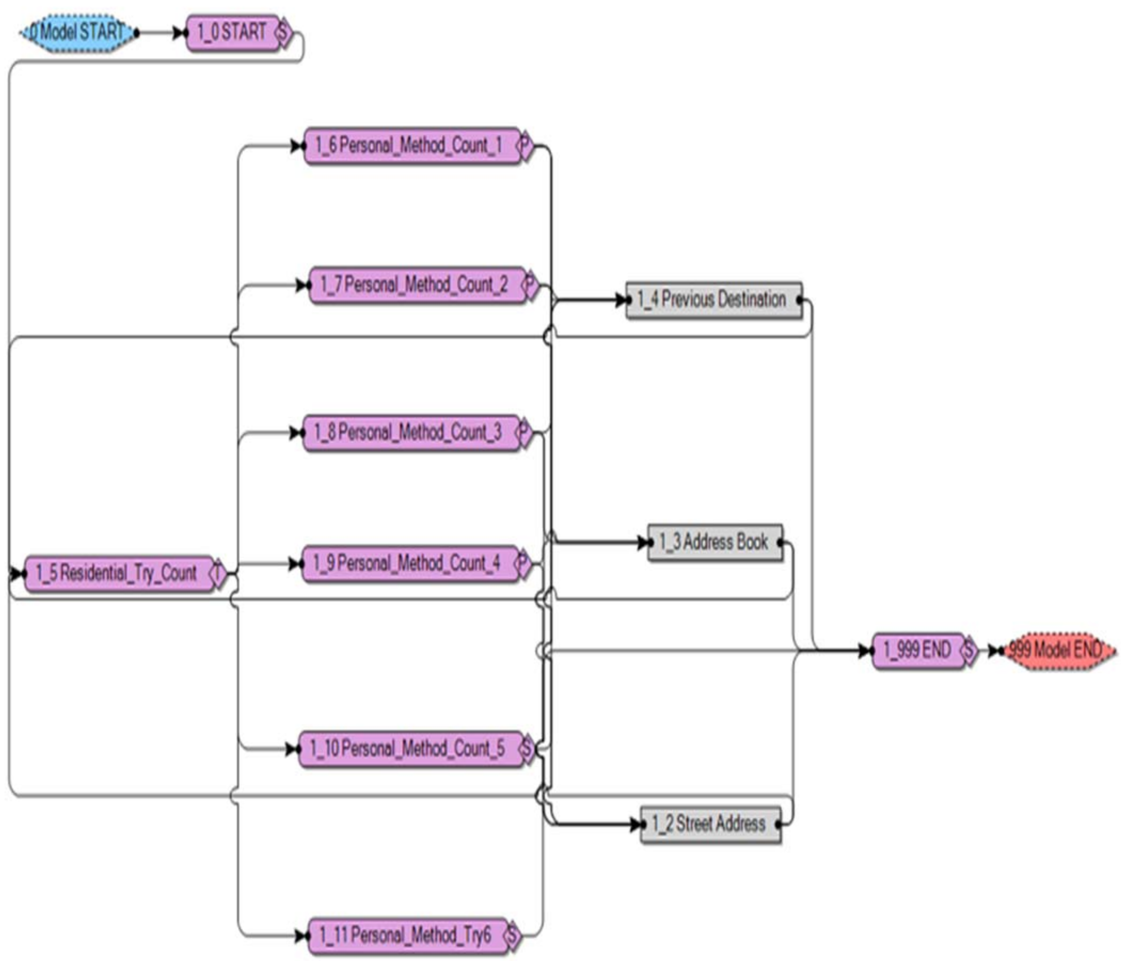


Figure 5-3. Sub Network for Residential Address

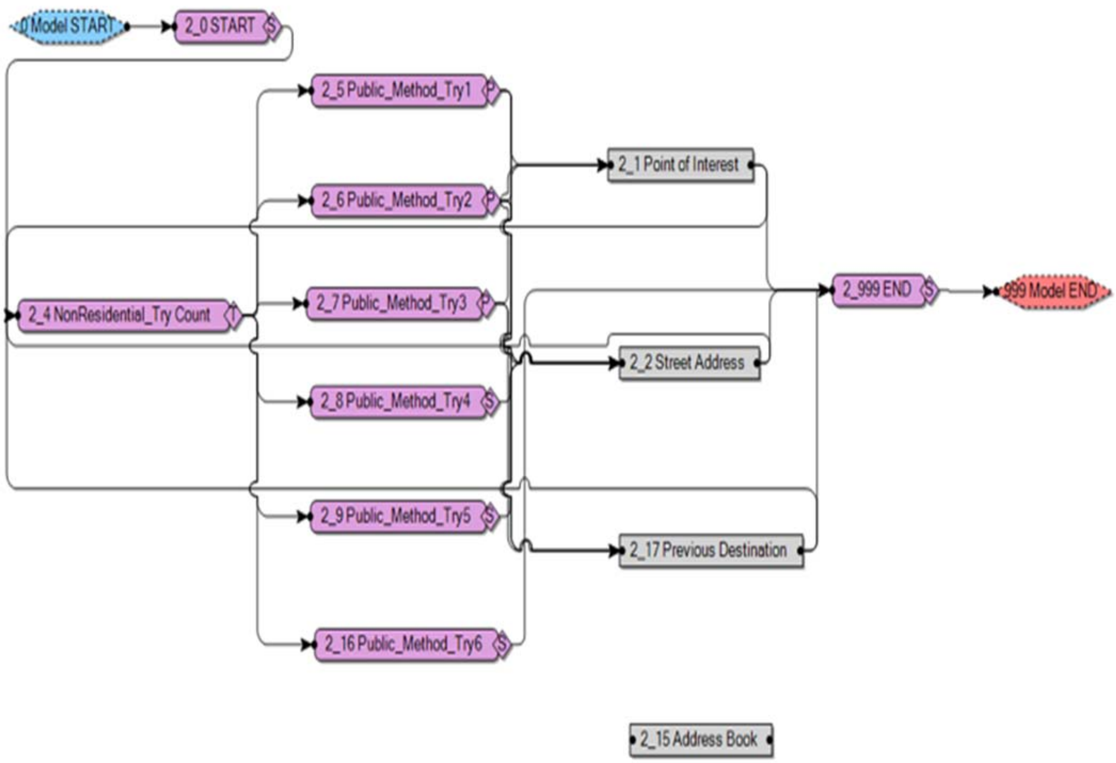


Figure 5-4. Subnetwork for Non-Residential Address

Subnetwork for Destination Entry Method – Street Address

All methods begin with the system asking for the command to determine which entry method subjects can use. Figure 5-5 shows the detailed net sub network using *street address*. After selecting the method, system requests the city information, in which case the state name is assumed to be the state name used from previous destination entry attempt. If subjects cannot remember the predetermined state, and did not change the state, errors can occur. Sometimes subjects choose to enter the state name to avoid errors.

The next step is entering the city name. With some probability, the system can correctly identify subject's utterance and provide direct feedback. Otherwise, depending on what the subject uttered and what was understood, the system can provide one of several alternative feedback messages that provide options to the subject from which they must select. These extra steps increase the number of turns required in completing the task.

Next, the street information is entered. The simulator experiment showed that subjects may only provide partial information, in particular, forgetting to say the street suffix (e.g., Road) or the directional prefix (e.g., North) with road name. This increases the error probability and can result in several extra turns to complete the entry tasks.

However, the only format that the system can recognize is if subjects say the house number digit-by-digit (one five zero zero). This is different from the way that people normally say house numbers (fifteen hundred). Using entry methods other than digit-by-digit results in errors.

Subnetwork for Destination Entry Method – Previous Destination

The subnetwork is shown in Figure 5-6. The subtasks include uttering commands, selecting the desired number from the lists, and starting guidance, relatively easy to do with the *street address* method.

Subnetwork for Destination Entry Method – Address Book

Similar to the method of *previous destination*, the subtasks for using *address book* include uttering commands (including the pre-determined word – destination by address book), selecting the pre-saved name lists, choosing the desired number form the lists, and starting guidance (Figure 5-7).

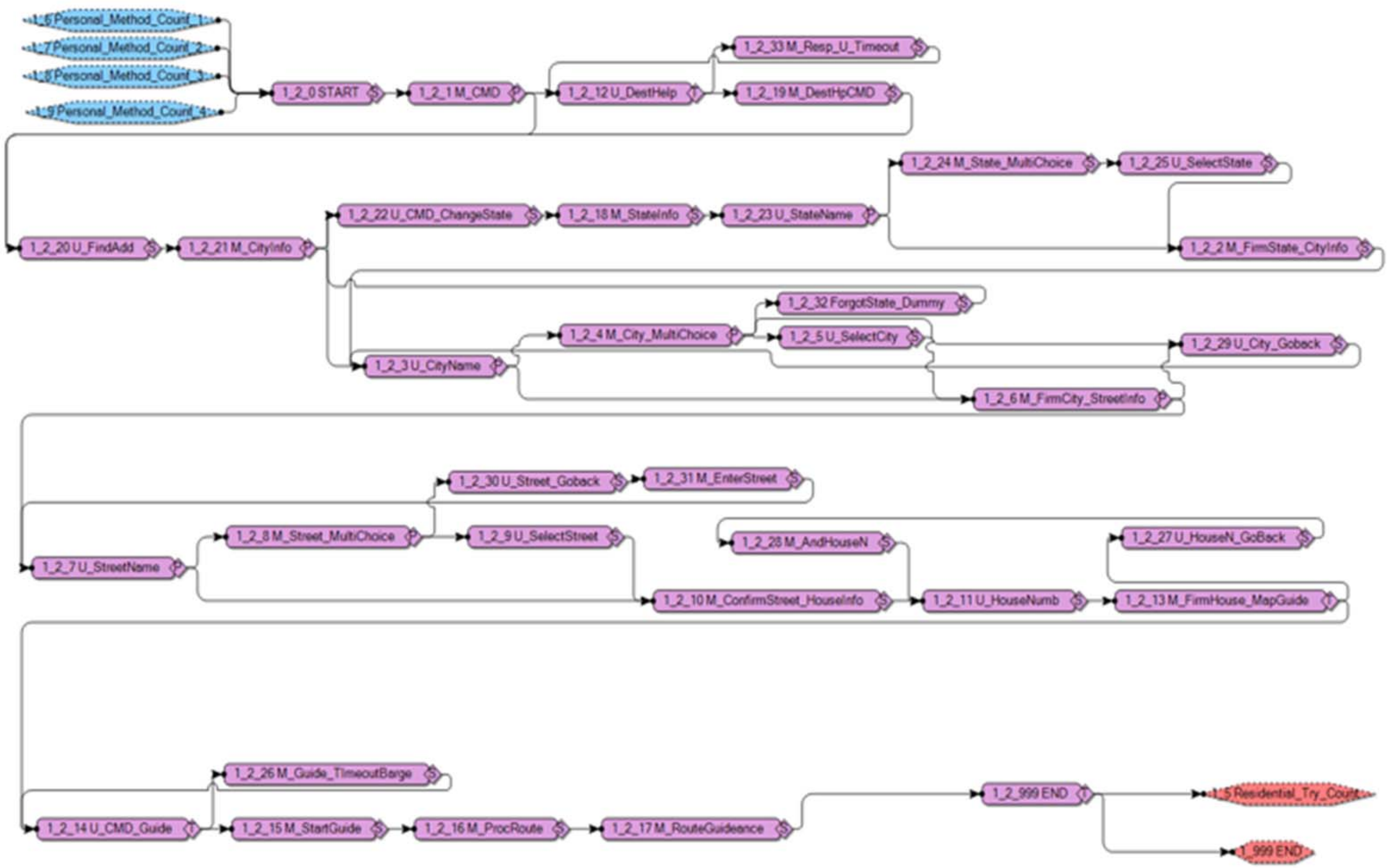


Figure 5-5. Subnetwork for Street Address

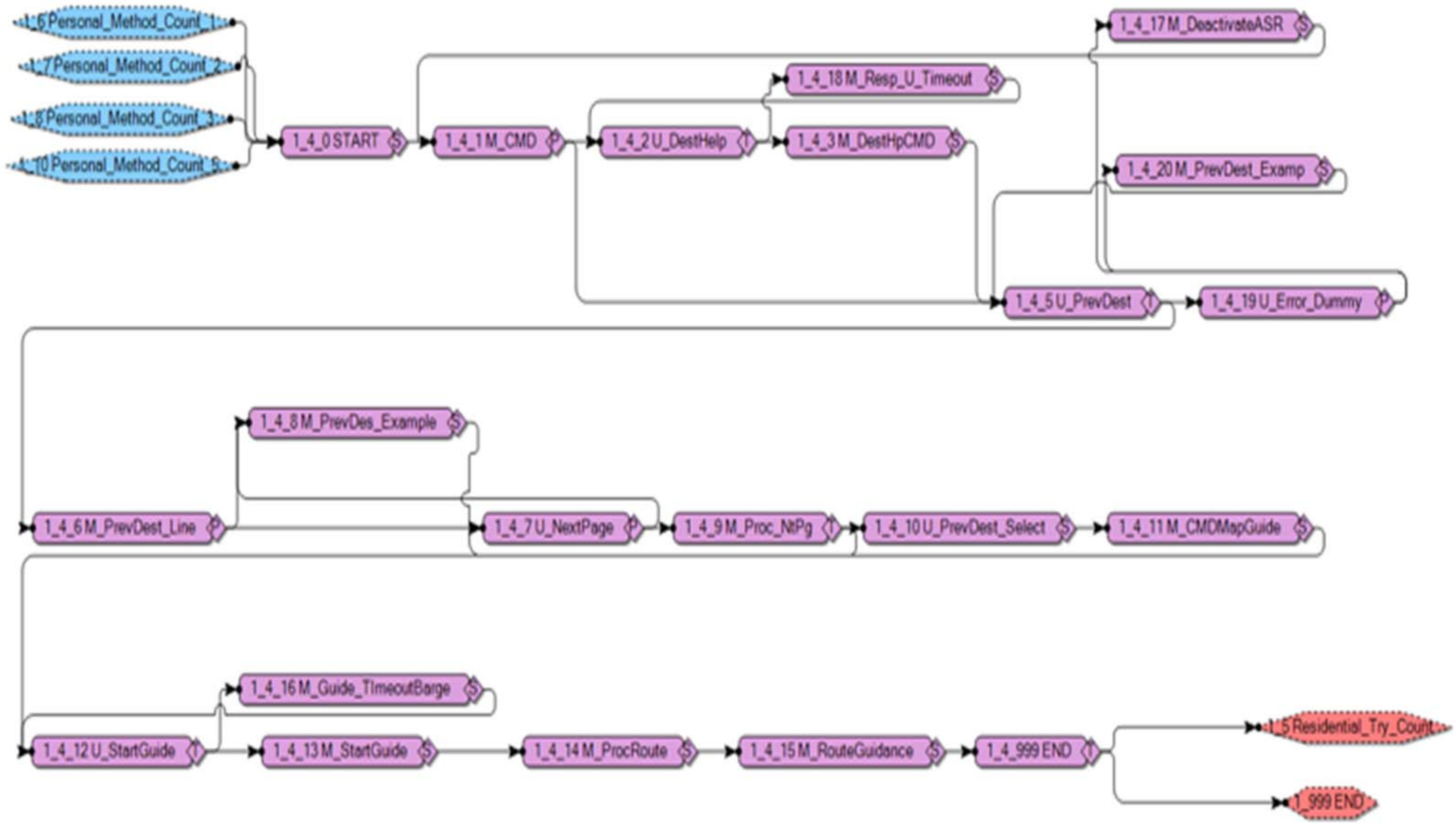


Figure 5-6. Subnetwork for Previous Destination

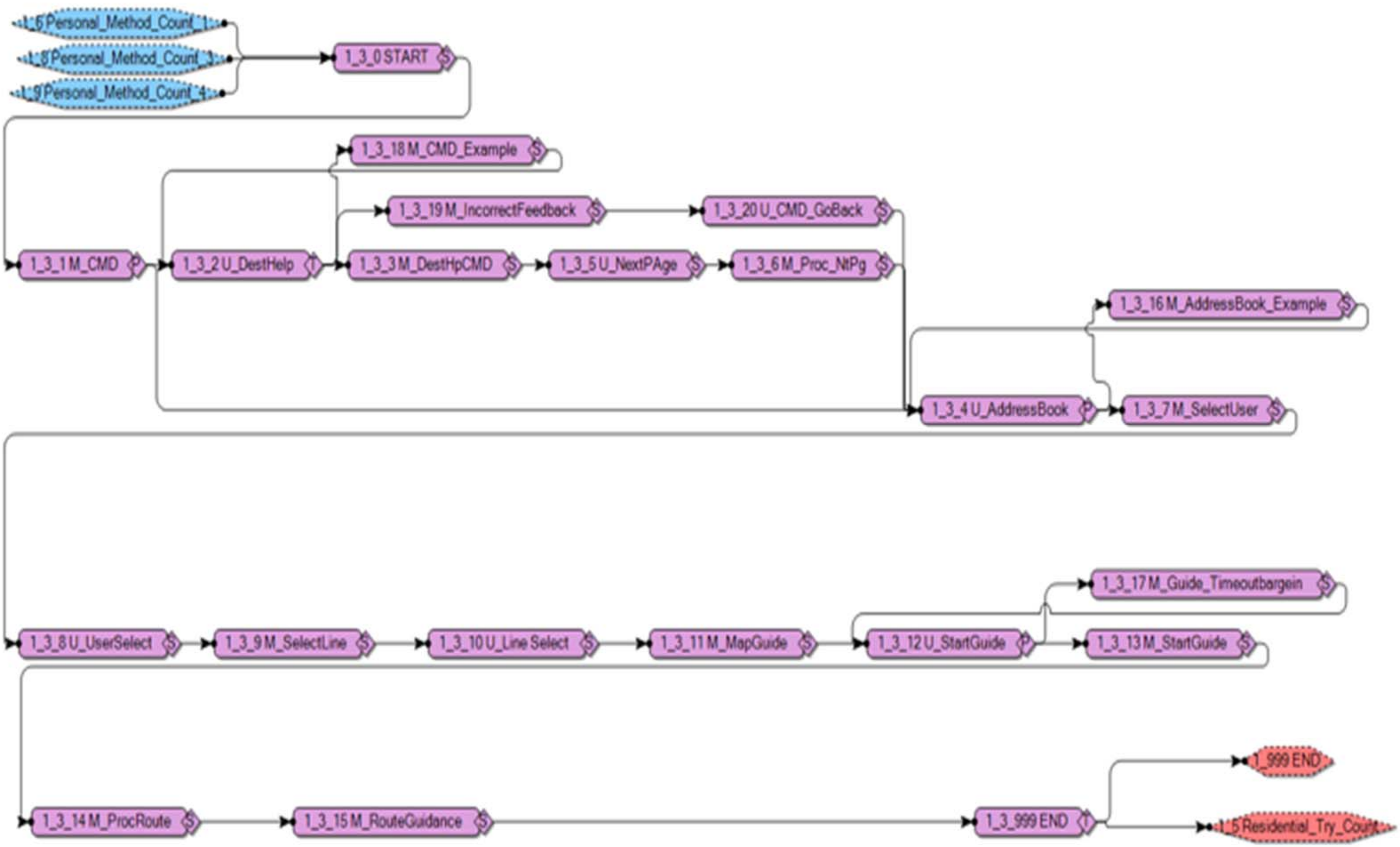


Figure 5-7. Subnetwork for Destination by Address Book

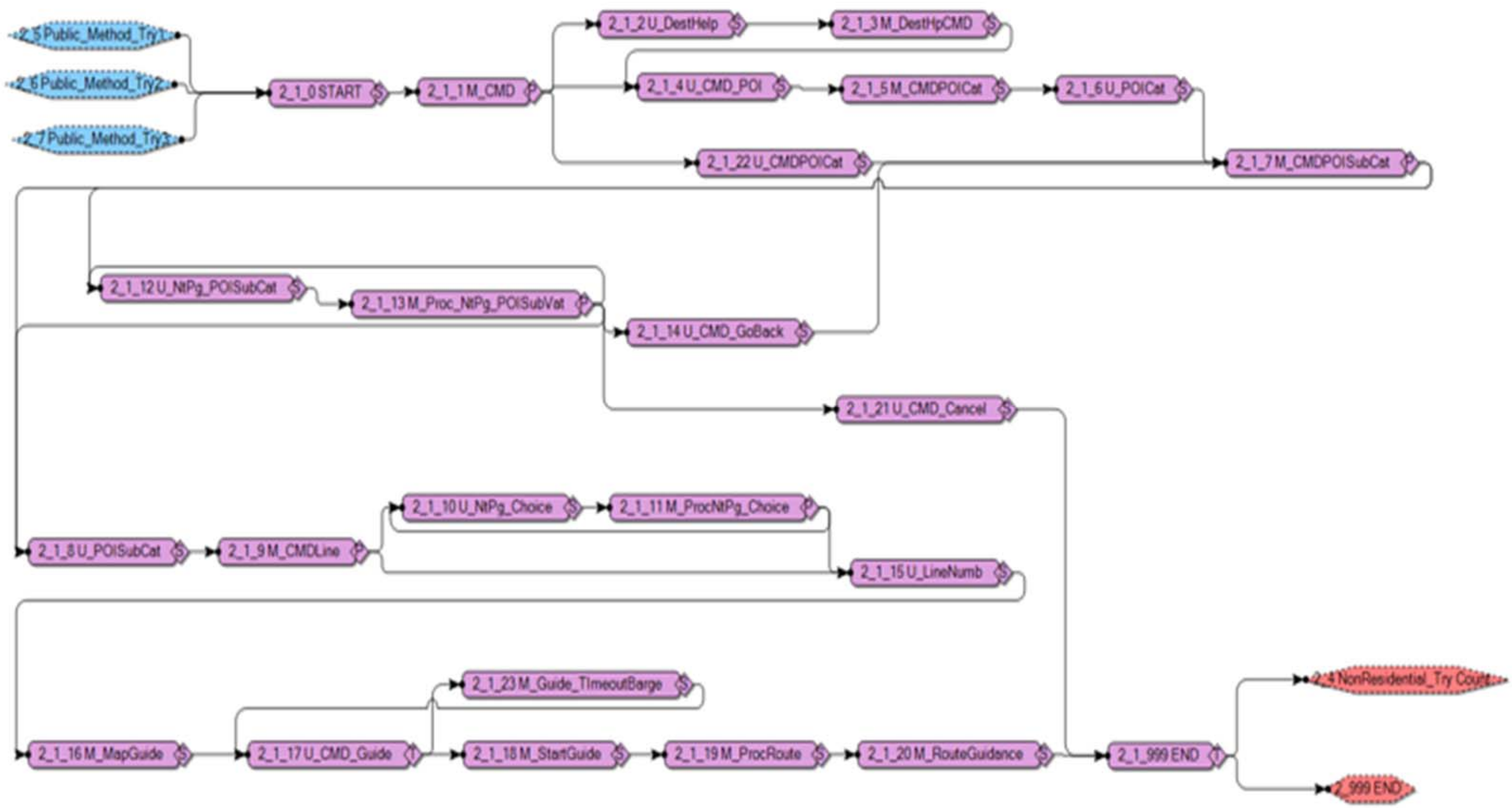


Figure 5-8. Subnetwork for Point of Interest

Subnetwork for Destination Entry Method – Point of Interest

The last method includes in the model is using *Point of Interest* and the subtasks include uttering commands (including the pre-determined word – find nearest POI), selecting the POI categories, and sub categories, and then choosing the desired number from the lists and starting guidance (Figure 5-8). The most difficult part for subjects is the subjects must have the knowledge on the relationship between the desired destination and the category and sub-category. For example, there is no category or subcategory titled "university." To select a university as POI, one must first select "Community" for main POI category, and then "Higher Education."

3. Variables and parameters sued to constructed the model

The variables used to construct the model are listed in Table 5-3. All the probabilities were based on the resulted from the simulation studies. For example, the probability of residential and nor residential address were calculated by counting the number of in- and out-state entries that occurred in scenario 3, 5, and 11 in simulator experiment. Also, the probabilities of the task success frequency were calculated by counting the number of trials with each entry method ending with correctly finding the destination and divided by the number of the total trials by each method.

Table 5-3. Variables for Simulation Model for the Destination Entry Task while Driving

Variables	Predicted Equations or Distribution (times in seconds)
Probability of residential and non-residential address	Residential address: 0.25; Non-residential address: 0.75
Probability of entry method used by each attempt	
Probability of each subtasks	
Error probability of each error type	
S: Thinking and processing time	Address related to personal: T = Lognormal (1.48, 0.57) Address not related to personal: T = Lognormal (1.31, 0.61)
S: Command utterance time	S_CMD T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
S: Think and utter the state name	S_Think_State T = Lognormal (-0.5, 0.49) S_State T = -0.212 + 0.242*NSyllable + 0.098*Age – 0.140*Gender
S: Think and utter the city name	S_Think_City T = Normal (0.91, 0.78) S_City T = -0.103 + 0.172*NSyllable + 0.089* Age + 0.149*NWord
S: Think and utter the street name	S_Think_Street T = Lognormal (0.03, 0.58) S_Street T = 0.041 + 0.459*NWord + 0.188*Age – 0.103*Gender
S: Think and utter the house number	S_Think_House 1/T = Weibull (1.31, 2.42) S_House T = -0.805 + 0.681*NWord + 0.201*Age – 0.143*Workload – 0.197*Gender
S: Spell the word	Mean number of characters = Poisson (7.91) S_Spell T = -1.826 + 0.635*NCharacter + 1.09*NWord
S: Barge-in	Barge-in T = Normal (0.19, 0.11)
S: Time out	Time-Out T = Normal (7.12, 1.8)
M: Command prompt time	M_CMD T = 0.347 + 0.284*NWord
M: Process and prompt the state name	M_Proc_State T = 2.185 + 0.89*MultipleChoice M_State T = 0.117 + 0.163*NSyllable + 0.117*NWord
M: Process and prompt the city name	M_Proc_City T = 2.241 + 0.95*MultipleChoice M_City T = 0.211+ 0.212*NSyllable
M: Process and prompt the street name	M_Proc_Street T = 2.772 + 0.782*MultipleChoice M_Street T = 0.448+ 0.194*NSyllable
M: Process and prompt the house number	M_Proc_HouseNum T = Weibull (0.37, 2.44, 2.71) M_HouseNum T = 0.169+ 0.553*NWord
M: Route Processing time	M_Route Processing time T = 3.182 + 7.936 * Out_State

5.3.2 Model for Music Selection Tasks

A simulation model was developed to predict the time for drivers to select music while driving. Three methods were explored: finding the song directly, using the artist, or using the album name to find the songs (Figure 5-9). Keep in mind that a task is defined as a goal and a method to achieve that goal. As these three methods are different, they are also three different types of tasks be predicted - searching for a specific album or for a specific artist, and searching for a specific song. A detailed explanation follows of each of the three subtasks.

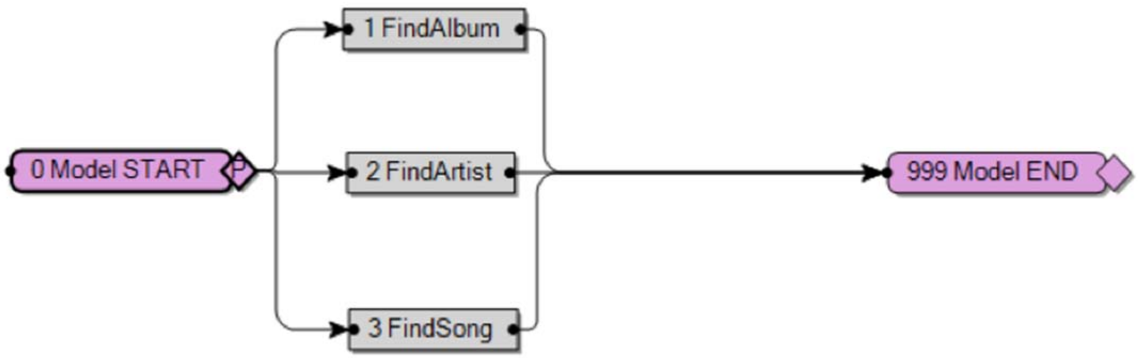


Figure 5-9. Overview of Finding a Song

Searching for a specific album

The subtask to search for a specific album includes 1) the driver learns the album name and activates the system, 2) the driver utters the command (*play album*) and the album name, 3) the system processes the information provided by a driver, and 4) the system provides the feedback (Figure 5-10). If errors occur, subjects needed to repeat the process to search again.

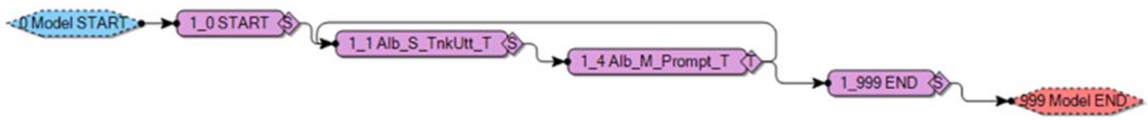


Figure 5-10. Overview of Model to Find a Specific Album

Searching for a specific artist

To search for a specific artist, the sequence is mostly the same, but the user utters "*play artist*" and the artist name (Figure 5-11).

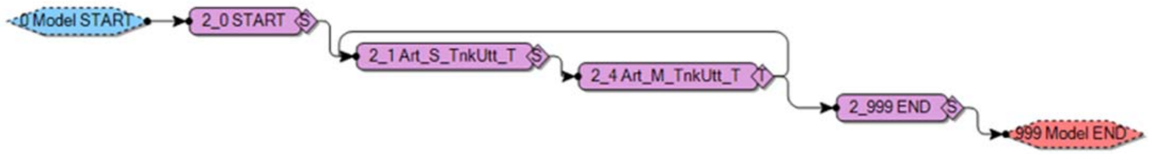


Figure 5-11. Overview of Model to Find a Specific Artist

Searching a specific song

The model to search for a song (Figure 5-12) is much more complex than searching for an artist or an album. In this task, subjects could only use the method of searching for an album/artist to narrow down the candidate song. Searching for a song directly is not supported due to a device limitation (iPhone 4s). Subjects may use more than one method when they failed using any one. Based on the data collected and shown in Chapter 4, subjects' speech entries could be categorized to these six cases.

1. Enter song name directly, try 1-3 times, and fail with errors as a result.
2. Enter song name directly, try 1-3 times, fail, enter artist name, change to next track if needed, succeed and finish this task.
3. Enter song name directly, try 1-3 times, fail, enter artist name, enter song name, fail again, enter song name, succeed and finish this task (this case was for only one subject).
4. Enter album name, change to next track if needed, succeed and finish this task, enter artist name, change to next track if needed, succeed and finish this task

Therefore, when subjects searched for a song by entering album/artist name in the beginning, they would not fail and may change the method. However, they could fail if they used the song title directly to search for it.

Table 5-4 shows the parameters for this task based on the results described in chapter 4. These variables include the probability of the searching method, task completion frequency, and duration of each subtask (Table 5-4).

Table 5-4. Variables for Simulation Model for the Music Selection Task while Driving

Subtask	Predicted Equations or Distribution (times in seconds)
Probability of searching method	Selection song - *Song*: 8.33%, *Artist*: 46.88%, and *Artist*: 44.79%
Task completion frequency for each attempt	
S: Thinking and processing time	S_Think T = 5.323 + 2.131*Age
S: Command and information utterance time	S_Utt T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
S: Thinking_2	S_Think_2 T = Lognormal (1.863, 0.551)
S: Next Track	S_NextTrack T = 0.429 + 0.131*Age + 0.092*Gender
S: Number to say next track	For *Play Album*: 1.55 ± 0.83 For *Play Artist*: 1.56 ± 0.84
M: Process 1	M_Proc1 T = 1.711 – 0.49*Correct_MS – 0.079*Age
M: Feedback and music playing	Chime T = 0.4 s Pause T = Normal (0.267, 0.024) M_Utt T = 0.559 + 0.091* NSyllable + 0.084 * NWord Music playing is song specific
M: Process 2	M_Proc2 T = Weibull (0.69, 0.148) Chime T = 0.4 s Music playing is song specific

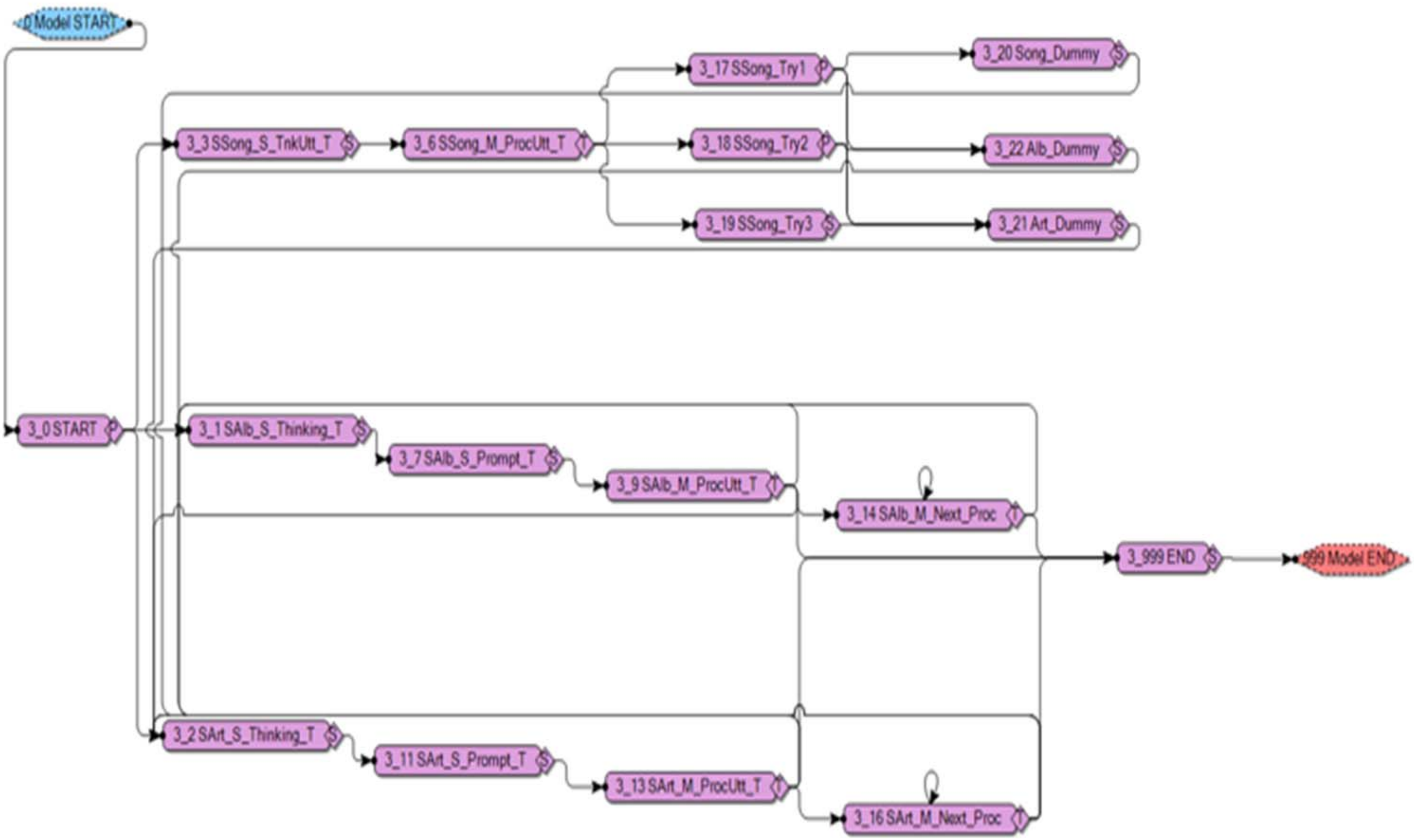


Figure 5-12. Overview of Model to Find a Specific Song

5.4 Simulation Validation

5.4.1 Destination Entry Tasks

Based on ten runs of 2000 repetitions each, the mean predicted task completion time was 99.04 ± 1.08s (Figure 5-13). The predicted task completion time from the simulation model was 24 s less than the mean task completion time (123.84 ± 78.07 s) resulting from the simulator experiment, $t_{(289, 0.05)} = 5.376$, $p < 0.001$), a difference of 24%. However, the variability of task completion time in the simulation was much less than that of the data from the experiment.

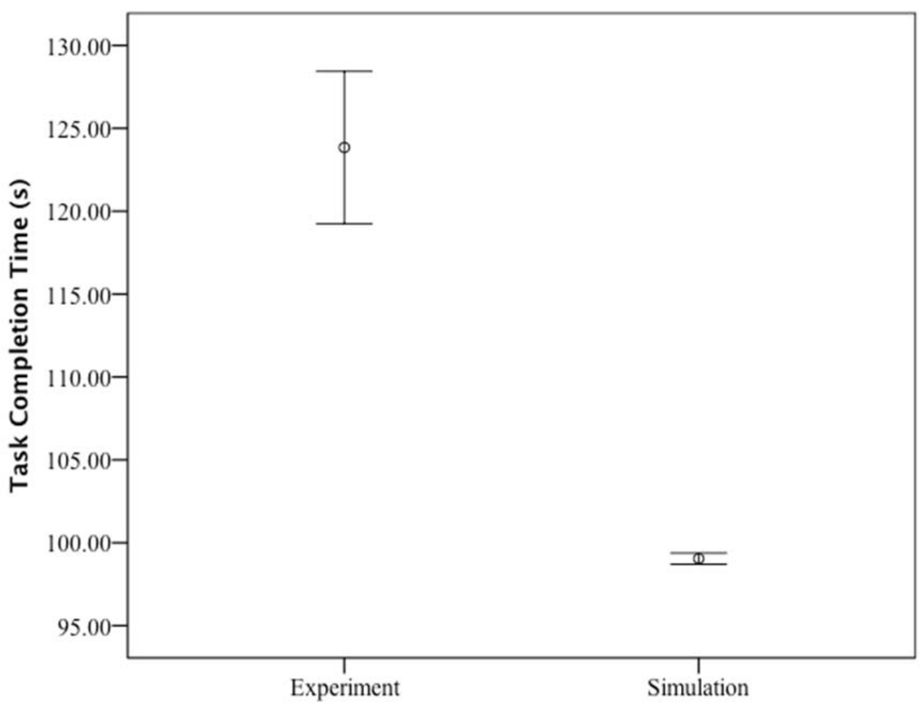


Figure 5-13. Mean and Standard Errors of the Task Completion Time of Destination Entry Task from the Experiment and Simulation.

The difference may be because the probability of each error type was calculated using the total human turns, instead of the turns relevant to each piece of unique information requested by the system. For example, the type of the invalid information should only occurred in the event when the system ask for the house number, or the event that system ask for city name only and subjects said city and state name together. Therefore, the probability for each type of error may be far less than the error occurrence resulted from the simulator study. Readjusting the probability for each type of error occurrence should improve the model.

Another reason for the experiment-simulation difference was that the four trials in which the drivers gave up were also included in the analysis. The tasks completion times from those four trials were quite large (133.91 s, 388.36 s, 195.52 s, and 482.6 s), increasing the mean and standard error from the experiment. Also, there were two trials in which the drivers tried more than 5 times to enter the destinations without giving up. The times for these two trials were also quite large (492.95 s and 464.45 s), again increasing the experimental estimated of the mean and standard error. Excluding those six trials, the mean task completion time from the simulator experiment was 118.82 ± 67.65 s, reducing the mean total time difference between the experiment and the simulation to 16 %.

Also the mean probability of each type of human errors used in the model is shown in Table 5-5. The only error estimate for which there was close agreement was the incorrect command error probability. The mean probability of the subject barge-in was predicted to occurrence was 0.36, which was almost triple than the results from simulator study.

Therefore, readjusting the error probability is necessary to reduce the difference between the simulation prediction and results from the experiment.

Table 5-5. Comparison Mean of Total Error Predicted from Simulation Model with the Task Completion from Simulator Experiment Validation Group.

	Barge-in	Time Out	Incomplete information	Incorrect Command
Model	0.355	0.517	0.014	0.046
Simulator Study	0.122	0.212	0.073	0.044

5.4.2 Music Selection Tasks

After ten runs with 2500 repetitions per run, the mean predicted task completion time was 19.41 s. The task completion times to search for a specific album, a specific artist, and a specific song were 11.76, 12.13, and 24.35 s, respectively (Table 5-6, Figure 5-14, and 5-15). The pooled mean task completion time of selecting music resulted from the validation group in simulator experiment was 24.25 s, which is significantly greater than the predicted time by simulation model ($t_{(159, 0.05)} = 3.218$, $p = 0.002$).

Table 5-6. Comparison of Task Completion Time (s) Predicted from Simulation Model with the Task Completion from Simulator Experiment Validation Group.

	Pooled	Album	Artist	Song
Model	19.41 ± 0.08	11.76 ± 0.11	12.13 ± 0.06	24.35 ± 0.14
Experiment	24.25 ± 18.49	14.71 ± 8.33	14.99 ± 12.09	30.11 ± 19.43

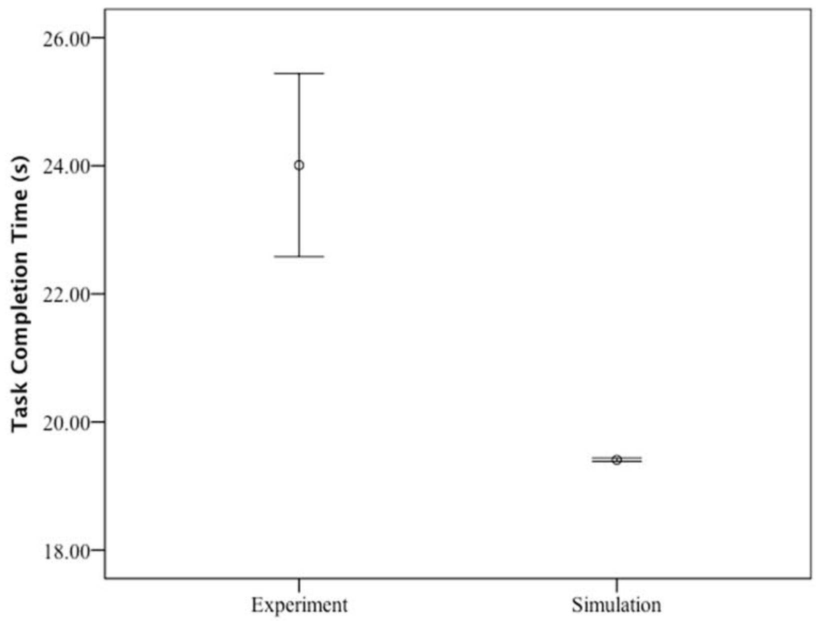


Figure 5-14. Mean and Standard Errors of the Task Completion Time of Music Selection Task from the Experiment and Simulation.

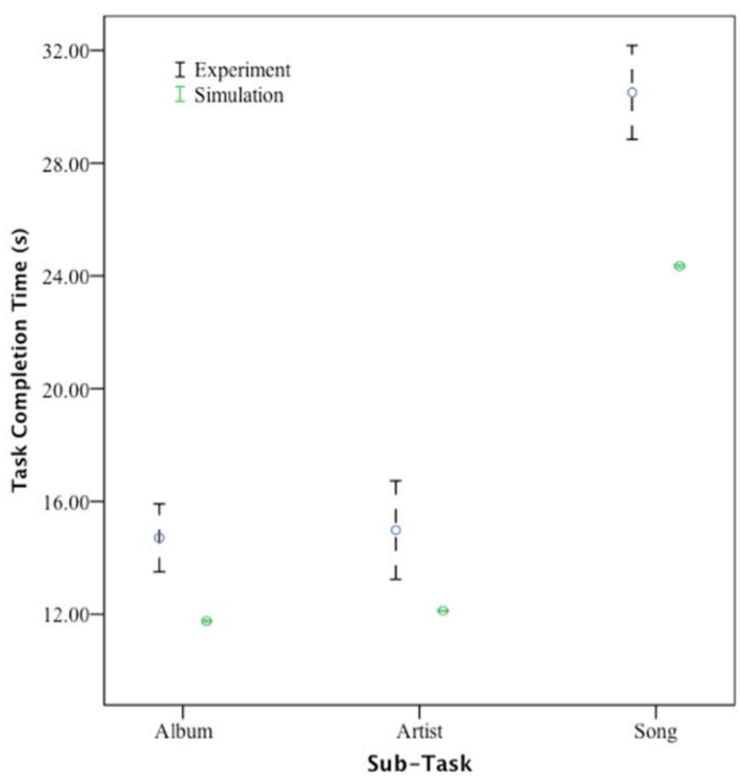


Figure 5-15. Mean and Standard Errors of the Task Completion Time of Music Selection Sub-Task from the Experiment and Simulation.

Although selecting a specific album or artist is an easy task, especially for less error-prone speech systems, which took two turns to complete the task, there was a 20% difference in the task completion time between the model prediction and the simulator experiment results. There was also a 20% difference in the task completion time between the model prediction and the simulator results while selecting a specific song. The reason for the difference is that the subtask time was not included in the model as the duration between the speech interface's feedback and the MP3 player starts to play the music is song specific. From the results in simulator study, this specific duration were ranged from 1 s to 11s. Also, when drivers want to listen to a specific song, the time for MP3 player to search for and begin to play the song should not distract the driver. Therefore, it is reasonable to not include the duration of this subtask into the model

Table 5-7 and Figure 5-16 shows the mean of the total errors, human errors, and machine errors from the simulation model and from simulator experiment. The results show that there was no difference of total errors between the model and simulator experiment, with only 0.1 time difference (($t_{(161, 0.05)} = -0.881$, $p = 0.379$). Also, there was no difference of human errors between the model prediction and simulator experiment ($t_{(161, 0.05)} = -0.326$, $p = 0.002$). However, the predicted machine error from the model was 2.5 times greater than the number of the simulation model ($t_{(159, 0.05)} = -2.242$, $p = 0.026$). The disagreement could be because the probability of a machine error resulting from the model-built group in simulator experiment was greater than the validation group, which resulted in a greater frequency than the model.

Table 5-7. Comparison of Total Errors, Human Errors, and Machine Errors Predicted from Model with the Errors from Experiment

	Total	Human Error	Machine Error
Model	0.36 ± 0.01	0.31 ± 0.01	0.05
Experiment	0.31 ± 0.74	0.29 ± 0.71	0.02 ± 0.18

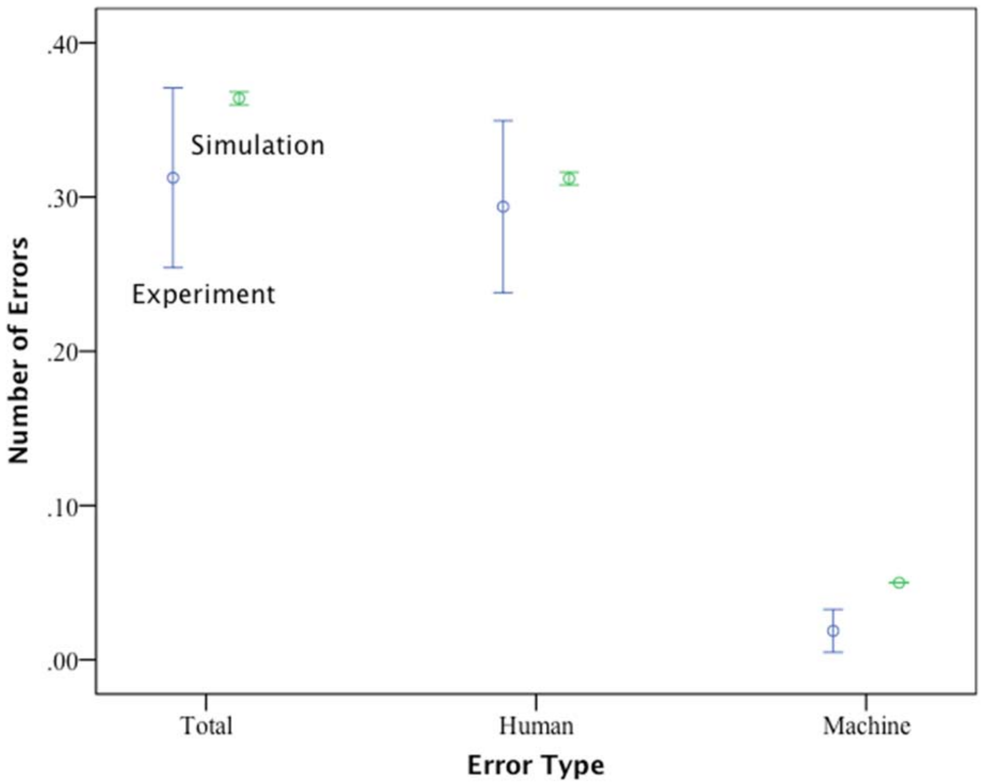


Figure 5-16. Mean and Standard Errors of the number of Errors by Error Types for Music Selection Task from the Experiment and Simulation.

5.5 Conclusion

The goal of this study was to provide a model structure (the tasks users perform and their sequence for various error contingencies) and data (either distributions of task time or prediction of them from various tasks variables, as well as estimates of error probabilities) for use with existing simulation software packages to predict user task performance with speech interfaces in motor vehicles.

For destination entry tasks, there was a 24% difference in the task completion time compared with the results in the simulator experiment (99 s vs. 123 s). The major reason for the difference has to do with the probability of each type of human error calculated in the simulator study. Adjustments to those error probabilities should lead to simulation task time estimates that more closely approximate the experimental data.

For music selection, there was a 20% difference of the task completion time comparing with the results in the simulator experiment. The major reason was that the time for the MP3 player to play the song was not included in the simulation. Total errors and human errors from the model agreed with the results from the experiment. However, the machine errors predicted by the model were greater than those from experiment, which required adjustment. The duration of subtasks as well as the frequency of selection method, task completion frequency, and number of times required to say next track are the variables used to construct the model. The task completion time and number of errors can be predicted by this model and were validated by the results from simulator experiment.

The U.S. DOT has just released revised design and assessment guidelines to help minimize in-vehicle distractions. Those guidelines concern visual-manual interfaces. As was pointed out in Chapter 1, using a speech interface may potentially decrease the eyes-off-road time, and research has demonstrated that increasing the duration of eyes-off-road time will increase the crash risk. However, as was noted earlier, even speech interfaces can have a visual display. To design a safe and easy to use speech interface, the results from this research provide pioneering data, equations, and a discrete event simulation model that can predict task times and errors when using speech interfaces. Using the information in this dissertation, the usability team can easily predict the task performance of a speech interface well before a first physical prototype is available, saving time, reducing development cost, and most importantly, enhancing the safety of the motoring public.

CHAPTER 6

Discussion, Conclusions, and Future Work

The goal of this research was to understand the interaction between the drivers and a speech interface and to provide a model structure (the tasks users perform and their sequence for various error contingencies) and data (either distributions of task time or predictions of them from various tasks variables, as well as estimates of error probabilities) for use with existing simulation software packages to predict user task performance with speech interfaces in motor vehicles. The research necessary to accomplish that goal has been described and the model has been completed. Associated with each of the activities of this dissertation have been conclusions, summarized here.

6.1 Summary of Findings

6.1.1 Review of Human Factors Related Literature for Automotive Speech Interfaces

1. There have been a number of research and development efforts to produce speech interfaces for motor vehicles, especially passenger cars.

2. Real world data on speech interface use is extremely limited, with the most important study being Winter's research on typical utterance patterns (stereotypes) [26].

3. As summarized in Barón and Green [28], along with the review in this dissertation of more recent studies, speech interfaces can have many advantages over visual-manual interfaces, with the most common finding (eight studies out of ten in which it was examined), that speech interfaces lead to better lane keeping. Also found, are improved peripheral detection time, and shorter brake reaction times, but these findings are based on one or two studies.

4. In terms of task performance, whether speech interfaces take more or less time than visual-manual interfaces varies with the study and the speech interface. In recently completed UMTRI studies, the issues are whether subjects know what to say, the quality of the speech recognizer, and the match between the subject's understanding of when the system listens and when it actually listens.

5. Finally, in terms of distraction related measures, speech interfaces led to fewer off-road glances.

6. There are a number of SAE, ISO, and ITU standards that relate to speech interfaces. However, the most significant standard, which is of concern to all automotive manufacturers and suppliers, are the U.S. DOT visual-manual guidelines, which specify test procedures for in-vehicle interfaces involving human subjects. The "final" guidelines were released on April 23, 2013.

7. Speech interfaces should be guided by principles from linguistics, over which there are many. Probably the most useful are Grice's Conversational Maxims [52].

8. The speech literature also provides many ways to classify errors, with the most important distinctions being between lexical, syntactic, and semantic levels.

9. Of the various ideas about how speech interfaces should be evaluated, the most significant comes from the PARADISE project [114]. The key idea is that there are multiple aspects to an evaluation, and they may need to be considered separately.

6.1.2 Method Used by Drivers to Enter the Destination for Navigation and Music Selection for MP3 Players

The purpose of the study described in Chapter 3 was to investigate for what purposes and how drivers used navigation devices and MP3 players. Thirty licensed drivers (16 F, 14 M; 28 ± 10 years) from southeast Michigan (typical drivers) and eleven licensed drivers (1 F, 10 M; 39 ± 10 years) from the Nissan Technical Center in Farmington Hills, Michigan (auto experts) were recruited.

1. Unexpectedly, subjects predominantly reported they used navigation systems to reach familiar destinations (typical drivers: 61%; auto experts: 89%).

2. *History* was self-reported to be a very common entry method (typical drivers: 30%; auto experts: 24%), which conflicted with data retrieved from navigation systems (both groups: <1%).

3. Visiting "Friends' Houses" (19%), "Home" (17%), "Shopping" (15%), and "Community" (13%) were the four top-ranked POI categories on the *Favorite* lists for typical drivers. "Home" (33%) was the most frequent POI category on the *Favorite* lists for auto experts.

4. Based on the history list, common trip purposes included *Shopping* and *Visiting Friends' Houses*.

5. The mean numbers of songs on their MP3 players were 420 for typical drivers and 1,200 for auto experts.

6. Of most interest are the three top-ranked methods the subjects used to search music, including "Song title," "Artist Name," and "Playlists."

The results of this work provide the test scenarios for the next experiments, and as well as relevant variables and parameters for the simulation model to be built. .

6.1.3 Driver Interaction with Navigation and Music Selection System Using Speech in a Driving Simulator

The purpose of the experiment described in Chapter 4 was to investigate how drivers interact with a speech interface for navigation and music selection. Specifically, the focus was on the duration of the subtasks for each system as well as the frequency of errors and error correction strategies.

1. The task completion frequency in this study was 97% for destination entry tasks and 99% for music selection tasks, respectively. The task completion time for those trials without any errors was 77 s for destination entry tasks and 19 s for music selection tasks.

2. The detailed steps and the equations or distributions used to predict the thinking and utterance times are described in the Table 4-55. For destination entry tasks, allowing approximately 0.5 s per word and 0.2 s per syllable for each utterance seems reasonable. As with the destination entry task, the predictive parameters for music selection utterances were again number of syllables, number of words, and gender, but the

combination that provided the best predictions, if there was one at all, varied with the subtask.

3. For the first attempt, the most frequent method to enter an address was the *street address* method (88), followed by the *address book* method (7%). When errors occurred, subjects preferred to use the street address method to correct the error.

4. There were 1,088 errors distributed among three groups: (1) information relevant, (2) system commands and entry method relevant, and (3) subjects' knowledge related to the system for destination entry tasks. Time-out and barge-in accounted one-third of the human errors. The correction strategy that repeating the utterance/repeating the utterance slower was the most common method used to correct the errors. For music selection task, 154 errors occurred. Barge-in and time out were also the two major errors for the music selection task. Repeat the utterance was the most frequently used method by the subjects to correct these two types of errors. Driving workload affected the destination entry tasks when the information given was correct and no errors occurred. The task completion time while the vehicle was parked was 6 s and 7 s longer than the time while driving in low and high workload conditions and the difference may be due to learning effects. This was not expected.

6.1.4 Simulation Development and Validation

The purpose of this research was to provide a model structure and data for use with an existing discrete-event simulation software package (IMPRINT) to predict drivers' task performance with speech interfaces in motor vehicles.

For destination entry tasks, there was a 20% difference in the task completion time compared with the results in simulator experiment (99 s vs. 123 s). The major reason for the difference has to do with how the probability of each type of human error was calculated from the simulator experiment results. Adjustments to those error probabilities should lead to simulation task time estimates that more closely approximate the experimental data.

For music selection, the predicted model agreed with the results from the simulator study for both task completion time and errors. The duration of subtasks, as well as the frequency of selection method, task completion frequency, and number of times required to say *next track*, are the variables used to construct the model. The task completion time and number of errors can be predicted using this model and was validated by the results from the simulator experiment.

6.2 Discussion of the Findings

6.2.1 Entry Method for Destination Entry and Music Selection Tasks

Results from the survey study showed that *street address* (46%), *POI* (31%), and *favorites/address book* (18%) was the most frequently used method to enter the destination (Table 3-3). Although results from the simulator study agreed the three methods were ranked top three, there was major disagreement on the frequency of the use of each method, *street address* (88%), *POI* (5%), and *favorites/address book* (7%). There are several reasons to explain this difference. First, the subjects who participated in the simulator experiment may not have been familiar with the speech interface (only 8% of subjects had experience using speech) and did not know some of the information had been pre-saved in the database, although subjects should learn this pre-saved information during the practice trials. Therefore, using *street address* to enter the destination at the first attempt was the easiest way to complete the task. Second, subjects may not remember the specific commands for the system, especially elderly subjects. Third, the command to use address book as entry method (*Destination by Address Book*) was listed on the second page of the three-page command lists. For convenience, subjects may not say "*Next Page*" to advance the command lists, which increases the total number of turns required to complete the task.

Detailed information on entry methods emerged from this research. From the survey, *Going home/visiting friends or relatives* was the most frequent destination recorded in the history lists. Sixty-three percent of the destination entries involved the *street address* method, 33% involved the *favorites/address book* method and 1% of the time drivers used the *history/previous destination* method. From the simulator experiment, there were three trials related to subject's home and relatives/friends home address. The percentages using *street address*, *favorites/address book*, *history/previous destination* methods for the first attempt were 67%, 28%, and 5%, respectively. The results from both experiments were close. However, none of the subjects used *Go Home* as the entry method during the simulator experiment, which is not what manufacturers/designers or the author expected. Again, this may be because subjects may have had limited experience with speech interfaces and were not familiar with the structure of the system.

When the destination was for shopping, there was a major disagreement between the survey and simulator experiment results. From the survey, the frequency of methods used recorded from subjects was *POI* (55%) and *street address* (37%). However, in the simulator study *POI* served as the entry method 6% of the time whereas *street address* accounted 94%. The reason for these differences is unknown.

The three most commonly used methods reported by subjects in the survey for music selection were *song title*, *artist name* and *playlists*. Due to a limitation of the device used in the simulator experiment, using song title or playlists to select a specific song was not available using speech. Therefore, no comparison can be made.

6.2.2 Trials with Incorrect Information

Providing incorrect information (the wrong city) was intended to simulate the situation where drivers were not familiar with the geographic area in which they were travelling and misunderstood the location of destination, which was the most difficult part of this experiment. Of particular interest was how drivers corrected these errors. This resulted in subjects giving up during the experiment in 7 of the 144 trials. Of those completed trials, *street address* was still the most frequent entry method subjects used to attempt the correction. Subjects saying "*Go Back*" to previous step to correct the information occurred 50% of the time, and "*Change City /Enter City*" 2.1% of the time. See Table 6-1.

Table 6-1. Frequency of Method Used by Subjects at the Second Attempt for Destination Entry Task While the Given Information was Incorrect

First Entry Method	Second Entry Method				
	Address	POI	Address Book	Show Map	Proceed without Entering House Number
Address	119	6	2	2	4
POI	8	3	0	0	0
Total	127	9	2	2	4

6.2.3 Subtasks and Prediction Estimate

Using command-based speech interface for destination entry task requires at least 15 total turns to complete the task without any errors. Tables 4-55 and 4-56 show the predicted equations or distributions with their parameters for each subtask for destination entry and music selection tasks. Although only commands of up to four words were examined, the estimated utterance time for one syllable was 0.2 s, which was close to the estimated time of 0.17 s from John's study [87], estimated from customer speech over the phone. Further, the results showed elderly subjects required more time than young and middle-aged subjects.

The time for subjects to think of the state, city, and house number followed similar distributions (Table 4-55). When these subtasks divided to elements, the times predicted from John's study were close to the mean time from the simulator experiment.

Barge-in was a major source of human errors. The time that subjects say commands or phrases before the system can accept the signal followed a normal distribution with a mean of 0.19 s. Although true for human-human conversation, machines cannot track back to earlier parts of the conversation and continue the dialogue. Also, naïve subjects did not realize they could press the voice button again to interrupt the system’s prompt and say a command or phrase.

6.2.4 Errors and Correction Strategies

The error types proposed in this study were not mutually exclusive. For example, the error that subject did not enter the house number digit by digit can be treated as an information relevant error or subject lack of knowledge error. Conversation maxims and

Veronis' typology of errors for nature language dialogue may explain part of the types of errors [52, 54].

A major source of human errors for the street name subtask was subjects failed to provide street suffixes (e.g., road) and /or direction of the road name (e.g., north). This type of error suggests that the design of a speech interface may violate the Grice's conversation maxims for human-to-human conversation. This type of error can result in multiple feedback messages from the system or incorrect information, both of which require subjects to take extra turns to complete the subtask and increase the task completion time. The more frequent correction strategies were spelling the word (31%) when the machine feedback was incorrect or providing requested information (selecting the correct information from the lists; 30%) when the correct information was on the lists.

Drivers may say numbers in many ways, such as one eight hundred for 1800, four thousand for 4000, or seventeen seventeen for 1717. However, the speech interface used in the simulator experiment cannot accept all of these variations, only digit by digit. Thus, the design of this system did not conform to expected human behavior. Although from a designer standpoint this design may reduce the chance of a recognition error occurring, it resulted in the increase of human error in this study. Also, subjects may think the given information was incorrect and change the city information, resulting in the increased task completion time. The major correction strategies used by subjects were rephrase (36%), repeat the utterance (17%), and say the command *go back* to rephrase the utterance (15%).

6.2.5 Simulation Predictions

The difference in the task completion time predicted by the simulation model for destination entry task model was 20% from the time resulted in the simulator study. The estimated duration for a subject to utter a syllable was close to the results in prior research. The major reason for the difference has to do with the probability of each type of human error calculated in simulator experiment. There were only three types of common human errors that occurred - barge-in, time out, and stammer the phrases or commands. The remaining types of errors were unique to subtasks. For example, subjects provided invalid format of information only occurred when the system asked for the house number, or when the system asked for only the city name, but subjects said both the city and state name. Some command-related errors, such as *destination by address book*, or *previous destination* should not occur when the system asks for city, state, street, and house number. Therefore, the probability used to construct the model may not reflect the true values in the simulator study.

Also, it was assumed that the probability of a method was selected for each attempt was not affected by the method selected for the previous attempt. This assumption may conflict with the simulator experiment results. There were too many conditional probabilities to consider beyond the scope of this dissertation. Furthermore, there is a concern that using conditional probabilities would make the simulation model too complicated and too system-specific.

Another issue that may affect the results was the *address book* method was not in the model for the non-residential address. The result was conflicted with the survey study that the *address book* method was ranked third as the method used to enter destination. In the simulator experiment, non-residential addresses of trials were not saved to the favorite lists. Whether this factor affects the simulation model is unknown.

The model for music selection predicts the task completion time quite well. The model did not include the duration between the speech interface's feedback and when a MP3 player starts to play the music. The duration is song specific. The more steps subjects need to perform, the higher number of errors may occur. Also, searching the specific song using song title is not feasible by the device used in this study. This also conflicts with the results in the survey study that using song title to select music was the most frequently used method by subjects.

However, one should be careful that the model may not predict performance with some currently manufactured systems. On the plus side, systems such as Ford SYNC, requires drivers to provide the information step by step, a method similar to the navigation system.

6.2.6 Task Completion Time from the Simulator Experiment and Model Prediction

From the simulator experiment and model prediction, the task completion times for destination entry task were 77 s (without any error) and 99 s (w/ or w/o errors). Comparing with the SAE Recommended Practice J2364, the task completion time from the simulator experiment and model prediction is five times greater than the guideline. However, the readers should keep in mind that SAE J2364 is intended for **visual-manual interfaces**, not for speech interfaces. Also, the driver's eyes require looking at the visual-manual interface when operating these systems. Research has shown that the probability of a crash increases when the eyes-off-road duration is greater than 2 seconds [4].

On the other hand, the driver can keep their eyes on the road with occasional glances at the system screen when operating the speech interface. Interesting, even though the navigation interfaces utilized speech, there were instances where drivers need to glance at a display to make selection (by line number). Table 4-10 shows the duration of each commands or phrases prompted by the speech system. Interestingly, machine prompts consumed almost two-thirds of the total task completion time for destination entry tasks. For example, entering a destination outside the current state using the *street address* method required at least 12 turns and about 48 s when there were no errors, excluding the machine processing time. Therefore, focusing on the absolute task completion time may not be appropriate as the total duration for driver's input is far less.

The long task completion time for the destination entry task may discourage drivers from using a speech interface for this task. Designers should focus on how to reduce the number of turns to complete the destination entry tasks and the machine prompt time.

6.3 Comments on this Research

6.3.1 Strengths

1. This research was based on the methods commonly used to select destinations and select songs.
2. The song database was created based on what drivers actually have.

3. The experiment to examine driver use of speech interfaces was conducted in a reasonably contemporary driving simulator and workload was varied. This allows for careful experimental control of the test conditions.

4. The utterance timing was extremely reliable. Data from analysts was double- and triple- checked.

5. The speech interface used in the simulator experiment was a real system

6. The model structure, task times, and task errors were based on data from real drivers in the simulation experiment.

7. The model not only considers correct entries, but also how drivers deal with errors, both of their own making and due to system imperfections.

8. The model predicts the task completion time for music selection as well as the destination entry tasks.

6.3.2 Weaknesses

1. Only the navigation system from Mobis/Hyundai was examined. Although the model structure for other systems may differ, predicted times for some subtasks still can be used.

2. Learning played an important role in the simulator experiment, with the performance measure changing as a function of practice and driving conditions. More practice trials are needed.

3. The research was conducted in a driving simulator, where conditions were well controlled, not on a real road where there are more variations.

6.4 Recommendations for Future Work

6.4.1 Explore Drivers with the Experience in Using Speech Interfaces.

Subjects recruited for survey and simulator studies were convenience samples. Subjects were recruited on a first come first serve basis until the target number was reached for each age-gender group. This approach is typically used for human factors studies. However, only 8% of these subjects recruited for both studies had experience with using the speech interface of a navigation system. Had there been significant funding to recruit and pay subjects, a more experienced group of subjects could have been recruited. The simulation model built based on the results from these two experiments is most appropriate for drivers who are naïve to speech interfaces. Also, naïve drivers may not have the knowledge on the structure of the speech system as well or the pre-determined commands results, which may alter the frequency of each type of error. Studies with more experienced speech interface users will hopefully enable a better understanding of the human-speech system interactions.

6.4.2 Investigate The Possibility of Developing Keystroke-Level Like Models from Empirical Data for Command-Based Speech Interface.

The sum of the times for the subtasks for subjects to think of the state, city, and house number from the simulator experiment were close to the data from John [87]. Future investigations should consider other subtasks and they should be included in the simulation model to broaden the tasks covered. These data should be incorporated into a future draft of SAE Recommended Practice J2365.

6.4.3 Expand The Simulation Model Built for One Typical Command-Based Speech System to General Command-Based Interface.

As the auto manufacturers were unwilling to provide a speech system for these experiments, this simulation model is limited to represent the Mobis/Hyundai system for navigation and iPhone with Siri application for music selection. Although, the method to find a specific song by *song title* was included, the probability to successfully select the correct song was 0, which is not true in the Ford SYNC system. Exploring other in-vehicle speech systems will allow modeling tasks for other automotive products.

6.4.4 Explore The Systems to Natural Language Speech System.

Without any errors, taking 15 total turns with mean time of 77 s to complete a destination entry task is long and may not always be acceptable by drivers. Although command based speech interfaces in motor vehicles will be in production for some time in the future, future generation systems will use natural language speech. They have the potential to further reduce the task completion time. The variables and equations provided in this research coupling an experiment on how drivers interact with the natural language system can be used for simulation to predict the driver performance.

6.4.5 Conduct an On Road Experiment to Validate the Predicted Subtasks Time from The Simulator Study.

As suggested earlier, an on the road experiment is a logical next step. It is unknown who will fund it or when, but such an experiment could cost $200,000 to $300,000 assuming an instrumented test vehicle is available and a manufacturer is willing to provide a speech interface. Given this cost, the most likely sponsor is NSF, not an auto manufacturer or supplier.

Also, a surprising challenge is obtaining a speech interface. The auto manufacturers were not willing to provide a speech interface for this dissertation, but Mobis, a supplier, was in part because of the extraordinarily good relations the University has with Mobis. This challenge was a combination of the manufacturers not having an established effort to evaluate speech interfaces outside their organization, a lack of standalone systems that were available, and a concern that the research would find problems with their systems and publicize the problems. The argument of getting hundreds of thousands of dollars of research for free was not sufficiently compelling.

6.4.6 Validate The Error Categories, Type of Errors and Correction Strategies.

It is hard to classify the type of error and correction strategies. For example, subjects may say the command *go back* and then repeat the same information. Should this be counted in using the command *go back* or *repeat* as correction method? Is there any linguistics theory or principle that can classify this situation due to a system limitation?

6.4.7 Revise The Probability of Types of Human Error Calculation.

Although there was 20% difference between the predicted task completion time and the simulator experiment results, the probability of human error should be revised to provide more precise predictions. As a consequence, the variables used in the simulation construction will be more useful to the designers.

6.4.8 Analyze The Workload Prediction Function Offered by IMPRINT.

To provide more information to the designers or usability evaluation professionals, the predictive model with the workload prediction function may show the peak workload when drivers are involved in performing a secondary task. By focusing on the worst cases for each subtask, the ease of use of the interface can be improved and distraction minimized.

6.4.9 Develop More Efficient Methods to Reduce Speech Data

The duration of each utterance and each response time was determined by listening to each interaction, looking at the intensity waveform on the screen, and manually determining when each utterance began an ended. This process required a virtual army of data analysts (or at least a squad) and took a significant amount of time to complete. Computer tools to do this more quickly and economically are desired. Aside from funding and equipment, this is the major roadblock to future research.

6.5 Conclusions

In this research, the methods frequently used by drivers for destination entry and music selection tasks, the most frequently visited POI categories, records saved as their favorites in their navigation devices, the interactions between the drivers and the interface for destination entry and music selection tasks, the equations and distributions with their parameters to predict the driver's utterance and the machine's prompt, the types of errors with the probability of occurrence from the driver and the system, correction strategies with their probability used by drivers, and the models to predicted driver's task performance for the destination entry and music selection tasks have been identified and developed. Some of the results have been published in peer-reviewed journal articles [17, 25]. Other results will be published to peer-reviewed journal articles as well. These articles should assist automotive manufacturers and suppliers in understanding the key findings in this research and applying to both improvements in existing systems and designs of new systems.

The simulation models for destination entry and music selection tasks proposed in this research were developed using the software package (IMPRINT), which is restricted to those researchers who work for the U.S. military or who have contracts and/or projects with the U.S. military. It may not ne the best choice for the automotive industry. As mentioned earlier, these models require modifications (to improve the error probability estimates) and further validation (on the road) to improve the predictions of task performance. After those enhancements, the model could be implemented in other popular or easy accessed simulation applications used by academics or industry, such as ProModel (used by Industrial and Operations Engineering at the University of Michigan) or Micro Saint SHARP (commercial package) and posted on the website at the University of Michigan Transportation Research Institute Driver Interface Group.

APPENDIX A

Details Explanation and Code for Simulation

This table provides task time for the task networks of destination entry tasks from Figure 5-2 to Figure 5-8.

Table A-1. Task Description and Task Time of the Model for Destination Entry Task

Task	Description	Task Time
0 Model START	Determine the path to residential or non-residential addresses.	
1 Residential	A function that contains sub-networks of tasks to enter the residential address.	
2 Non-Residential	A function that contains sub-networks of tasks to enter the non-residential address.	
1_5 Residential _Try_Count	Determine the number of attempts.	
1_6 Personal_Method_Count_1	Determine the entry method for the first attempt for a residential address.	
1_7 Personal_Method_Count_2	Determine the entry method for the second attempt for a residential address.	
1_8 Personal_Method_Count_3	Determine the entry method for the third attempt for a residential address.	
1_9 Personal_Method_Count_4	Determine the entry method for the fourth attempt for a residential address.	
1_10 Personal_Method_Count_5	Determine the entry method for the fifth attempt for a residential address.	
1_11 Personal_Method_Count_6	Subjects give up the task and end the simulation.	
1_2 Street Address	A function that contains sub-networks of tasks using the	

	Street Address method	
1_3 Address Book	A function that contains sub-networks of tasks using the *Address Book* method	
1_4 Previous Destination	A function that contains sub-networks of tasks using the *Previous Destination* method	
1_2_1 M_CMD	The system asks for commands from the user. M: Command Please.	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_2 U_DestHelp	The user utters the command for destination. U: Destination Help.	Time out. T = Normal (7.12, 1.8) Normal Utterance. T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_19 M_DestHpCMD	The system confirms the user's utterance and asks for commands of entry method. M: Destination help. Command please.	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_33 M_Resp_U_Timeout	The system provides examples to the user. M: For example, say find nearest POI or say help at any time.	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_20 U_FindAdd	The user utters a command for the *street address* method. U: Find address.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_21 M_CityInfo	The system confirms the user's utterance and asks for the city name. M: Find Address. The city please.	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_22 U_CMD_ChangeState	The user says the command *Change State* when the destination is not same as the default state.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_18 M_StateInfo	The system asks for the state name. M: Please enter the state name.	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_23 U_StateName	The user thinks of and utters the state name.	T = Lognormal (-0.5, 0.49) - 0.212 + 0.242*NSyllable + 0.098*Age – 0.140*Gender
1_2_24 M_State_MultiChoice	The system provides several possible states and asks the user to choose the correct one. M: Please select the respective line or start spelling.	T = 2.185 + 0.89*MultipleChoice + 0.347 + 0.284*NWord + Pause + Beep
1_2_25 U_SelectState	The user selects the correct state name.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_2 M_FirmState_CityInfo	The system confirms the state name and asks for the city name. M: XXX. The city please.	T = 2.185 + (0.117 + 0.163*NSyllable + 0.117*NWord)
1_2_3 U_CityName	The user thinks of and utters the city name.	T = (Normal (0.91, 0.78)) + (-0.103 + 0.172*NSyllable + 0.089* Age + 0.149*NWord)
1_2_4 M_City_MultiChoice	The system provides several possible cities and asks the user to choose the correct one. M:	T = (2.241 + 0.95*MultiChoice) + (0.211+ 0.212*NSyllable) + Pause + Beep

	Please select the respective line or start spelling.	
1_2_5 U_SelectCity	The user selects the correct city name	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_32 ForgotState_Dummy	The user forgets to change state and goes back to the node of uttering the command *Change State*.	
1_2_29 U_City_Goback	The user says the command *Go Back/Correct* to correct the error and reenters the city name	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_6 M_FirmCity_StreetInfo	The system processes and confirms the city name, and asks for the street name. M: XXX. The street please.	T = (2.441) + (0.448 + 0.194*NSyllable + 0.117*NWord) + (0.347 + 0.284*NWord) + Pause + Beep
1_2_7 U_StreetName	The user thinks of and utters the street name.	T = (Lognormal (0.03, 0.58)) + (0.041 + 0.459*NWord + 0.188*Age – 0.103*Gender)
1_2_8 M_Street_MultiChoice	The system processes and provides several possible streets and asks the user to choose the correct one. M: Please select the respective line or start spelling.	T = (2.772 + 0.782*MultipleChoice) + (0.448+ 0.194*NSyllable) + Pause + Beep
1_2_9 U_SelectStreet	The user selects the correct street name.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_30 U_Street_Goback	The user says the command *Go Back/Correct* to correct the error and reenter the street name.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_31 M_Street_EnterStreet	The system asks the user to enter the street name. M: Sorry, what is the street?	T = 0.347 + 0.284*NWord + Pause + Beep
1_2_10 M_ConfirmStreet_HouseInfo	The system processes and confirms the street name and asks for house number. M: XXX. And house number or if you don't know that, please say show map or start guidance.	T = 2.772 + (0.448 + 0.194*NSyllable + 0.3 + 0.347 + 0.284*NWord) + Pause + Beep
1_2_11 U_HouseNumb	The user thinks of and utters the house number.	T = (1/ Weibull (1.31, 2.42)) + (-0.805 + 0.681*NWord + 0.201*Age – 0.143*Workload – 0.197*Gender)
1_2_13 M_FirmHouse_MapGuide	The system processes and confirms house number and asks for choice of show map or start guidance. M: XXX. Show map or start guidance.	T = (Weibull (0.37, 2.44)) + (0.169 + 0.533*NWord + 0.3 + 0.347 + 0.284*NWord) + Pause + Beep
1_2_27 U_HouseN_Goback	The user says the command *Go Back/Correct* to correct the error of the house number.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_2_28 M_AndHouseN	The system processes and asks the user to enter the house number again. M: And house number or if you don't know that, please say show map or start guidance.	T = (0.169 + 0.533*NWord) + 0.3 + (0.347 + 0.284*NWord) + Pause + Beep
1_2_14 U_CMD_Guide	The user utters the command. U:	T = 0.136 + 0.133*NSyllable +

	Start guidance.	0.082*Age + 0.094*NWord
1_2_15 M_StartGuidance	The system confirms the user's utterance. M: Start guidance.	T = (0.347 + 0.284*NWord) + Pause + Chime
1_2_15 M_Guide_TimeoutBarge	The system provides guidance when the user barges in or times out. M: Say show map, start guidance, or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_2_16 M_ProcRoute	The system processes the route guidance.	T = 3.182 + 7.936*Out_State
1_2_17 M_RouteGuidance	The system provides the route guidance. M: Please proceed to the highlighted route and then the route guidance will start.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_2_999 End	End the simulation of the *Street Address* method and determine the next step.	
1_3_1 M_CMD	The system asks for commands from the user. M: Command Please.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_2 U_DestHelp	The user utters the command for the destination. U: Destination Help.	Time out. T = Normal (7.12, 1.8) Normal Utterance. T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_3 M_DestHpCMD	The system confirms the user's utterance and asks for a command for the entry method. M: Destination help. Command please.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_19 M_IncorrectFeedback	The system misrecognizes the user's prompt and provides incorrect feedback. M: Find nearest Korean restaurant, line please.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_18 M_CMD_Example	The system provides guidance when the user barges in or times out. M: You can say, for example, destination help or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_20 U_CMD_GoBack	The user utters the command *Go Back* to renter the command.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_5 U_NextPage	The user utters the command *Next Page* to find the command for address book method.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_18 M_Proc_NtPg	The system processes and advances to next page.	T = 0.142 + Beep
1_3_5 U_AddressBook	The user utters the command *Destination by Address Book* for the address book method.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_7 M_SelectUser	The system processes and asks the user to select the pre-saved records.	T = 0.4 + (0.347 + 0.284*NWord) + Pause + Beep
1_3_16 M_AddressBook_Example	The system provides guidance when the user barges in or times out. M: For example, say find nearest POI or say help at any	T = (0.347 + 0.284*NWord) + Pause + Beep

	time.	
1_3_8 U_UserSelect	The user selects the list of pre-saved lists of users' names.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_9 M_SelectLine	The system confirms and asks the user to select the pre-saved records. M: User X. Line please.	T = 2.41 + (0.347 + 0.284*NWord) + Pause + Beep
1_3_10 U_LineSelect	The user selects the list of pre-saved records of destination.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_9 M_MapGuidance	The system confirms and asks for a choice of show map or start guidance. M: XXX. Show map or start guidance..	T = 2.818 + (0.347 + 0.284*NWord) + Pause + (0.347 + 0.284*NWord) + Pause + Beep
1_3_12 U_StartGuide	The user utters the command. U: Start guidance.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_3_13 M_StartGuidance	The system confirms the user's utterance. M: Start guidance.	T = (0.347 + 0.284*NWord) + Pause + Chime
1_3_17 M_Guide_TimeoutBarge	The system provides guidance when the user barges in or times out. M: Say show map, start guidance, or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_14 M_ProcRoute	The system processes the route guidance.	T = 3.182 + 7.936*Out_State
1_3_15 M_RouteGuidance	The system provides the route guidance. M: Please proceed to the highlighted route and then the route guidance will start.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_3_999 End	End the simulation of the *Address Book* method and determine the next step.	
1_4_1 M_CMD	The system asks for commands from the user. M: Command Please.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_2 U_DestHelp	The user utters the command for destination. U: Destination Help.	Time out. T = Normal (7.12, 1.8) Normal Utterance. T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_4_3 M_DestHpCMD	The system confirms the user's utterance and asks for a command of entry method. M: Destination help. Command please.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_3 M_Resp_U_Timeout	The system responds when the user times out. M: You can say, for example, destination help or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_5 U_PrevDest	The user utters the command *Previous Destination* for the previous destination method.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_4_19 U_Error_Dummy	Determine the user's error types and the system's responses.	
1_4_20 M_PrevDest_Example	The system responds when the user times out. M: For example, say find nearest POI or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_21 M_DeactivateASR	The system deactivates the ASR	T = Normal (10.01, 0.5)

	and the user reactivates the ASR.	
1_4_6 M_PrevDest_Line	The system processes and asks the user to select the requested destination on the lists.	T = 0.313 + (0.347 + 0.284*NWord) + Pause + Beep
1_4_7 U_NextPage	The user utters the command *next page* to search for the requested destionation on the lists.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_4_8 M_PrevDes_Example	The system responds when the user times out. M: For example, say line two, next page, help, repeat, or back.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_9 M_Proc_NtPg	The system processes and advances to next page.	T = 0.142 + Beep
1_4_10 U_PrevDest_Select	The user selects the requested destination on the lists.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_4_11 M_CMDMapGuidance	The system confirms and asks for a choice of show map or start guidance. M: XXX. Show map or start guidance..	T = 2.818 + (0.347 + 0.284*NWord) + Pause + (0.347 + 0.284*NWord) + Pause + Beep
1_4_12 U_StartGuide	The user utters the command. U: Start guidance.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
1_4_13 M_StartGuidance	The system confirms the user's utterance. M: Start guidance.	T = (0.347 + 0.284*NWord) + Pause + Chime
1_4_16 M_Guide_TimeoutBarge	The system provides guidance when the user barges in or times out. M: Say show map, start guidance, or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_14 M_ProcRoute	The system processes the route guidance.	T = 3.182 + 7.936*Out_State
1_4_15 M_RouteGuidance	The system provides the route guidance. M: Please proceed to the highlighted route and then the route guidance will start.	T = (0.347 + 0.284*NWord) + Pause + Beep
1_4_999 End	End the simulation of the *Previous Destination* method and determine the next step.	
2_1_1 M_CMD	The system asks for a command from the user. M: Command Please.	T = (0.347 + 0.284*NWord) + Pause + Beep
2_1_2 U_DestHelp	The user utters the command for the destination. U: Destination Help.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_3 M_DestHpCMD	The system confirms the user's utterance and asks for commands for the entry method. M: Destination help. Command please.	T = (0.347 + 0.284*NWord) + Pause + Beep
2_1_4 U_CMD_POI	The user utters the command for the POI method. U: Find nearest POI.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_5 M_CMDPOICat	The system confirms the user's utterance and asks for the POI category. M: Please select a POI	T = (0.347 + 0.284*NWord) + Pause + Beep

	category.	
2_1_6 U_POICat	The user utters the POI category.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_7 M_CMDPOISubCat	The system confirms the user's utterance and asks for the POI subcategory. M: XXX. Please select a POI subcategory.	T = 2.826 + (0.347 + 0.284*NWord) + Pause + Beep
2_1_12 U_NtPg_POISubCat	The user utters the command *Next Page* to search for the requested destination on the POI subcategory.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_13 M_Proc_NtPg_POISubCat	The system processes the user's utterance and advances to next pages of the POI subcategory.	T = Normal (7.12, 1.8)
2_1_14 U_CMD_GoBack	The user utters the command *Go Back* to search for the requested destination on the POI subcategory.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_21 U_CMD_Cancel	The user utters the command *Cancel* to restart the trial.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_6 U_POISubCat	The user utters the POI subcategory.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_9 M_CMDLine	The system confirms the user's utterance and asks the user for the choice of POI subcategory. M: XXX. Line please.	T = (0.347 + 0.284*NWord) + Pause + Beep
2_1_10 U_NtPg_Choice	The user utters the command *Next Page* to search for the requested destination.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_11 M_ProcNtPg_Choice	The system processes the user's utterance and advances to next page.	T = 0.142 + Beep
2_1_10 U_LineNumb	The user selects the requested destination.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_16 M_MapGuidance	The system confirms and asks for a choice of showing map or starting guidance. M: XXX. Show map or start guidance..	T = 2.818 + (0.347 + 0.284*NWord) + Pause + (0.347 + 0.284*NWord) + Pause + Beep
2_1_17 U_CMD_Guide	The user utters the command. U: Start guidance.	T = 0.136 + 0.133*NSyllable + 0.082*Age + 0.094*NWord
2_1_18 M_StartGuidance	The system confirms the user's utterance. M: Start guidance.	T = (0.347 + 0.284*NWord) + Pause + Chime
2_1_23 M_Guide_TimeoutBarge	The system provides guidance when the user barges in or times out. M: Say show map, start guidance, or say help at any time.	T = (0.347 + 0.284*NWord) + Pause + Beep
2_1_19 M_ProcRoute	The system processes the route guidance.	T = 3.182 + 7.936*Out_State
2_1_20 M_RouteGuidance	The system provides the route guidance. M: Please proceed to the highlighted route and then the route guidance will start.	T = (0.347 + 0.284*NWord) + Pause + Beep
2_1_999 End	End the simulation of the *POI* method and determine the next step.	

The table provides transition probabilities for the task networks of destination entry tasks from Figure 5-2 to Figure 5-8.

Table A-2. The Transition of Probabilities of Task Sequences for Destination Entry Tasks.

Task	Proceeding Task	Following Task	Probability
0 Model STRAT		1_1 Residential	0.2500
		1_2 Non-Residential	0.7500
1_5 Residential_Try_Count	1_0 START	1_6 Personal_Method_Count_1	1st attempt
		1_7 Personal_Method_Count_2	2nd attempt
		1_8 Personal_Method_Count_3	3rd attempt
		1_9 Personal_Method_Count_4	4th attempt
		1_10 Personal_Method_Count_5	5th attempt
		1_11 Personal_Method_Count_6	6th attempt
1_6 Personal_Method_Count_1	1_5 Residential_Try_Count	1_2 Street Address	0.6250
		1_3 Address Book	0.3056
		1_4 Previous Destination	0.0694
1_7 Personal_Method_Count_2	1_5 Residential_Try_Count	1_2 Street Address	0.2941
		1_4 Previous Destination	0.7059
1_8 Personal_Method_Count_3	1_5 Residential_Try_Count	1_2 Street Address	0.6000
		1_3 Address Book	0.1000
		1_4 Previous Destination	0.3000
1_9 Personal_Method_Count_4	1_5 Residential_Try_Count	1_2 Street Address	0.5000
		1_3 Address Book	0.5000
1_10 Personal_Method_Count_5	1_5 Residential_Try_Count	1_2 Street Address	1.0000
1_11 Personal_Method_Count_6	1_5 Residential_Try_Count	1_999 End	1.0000
1_2_1 M_CMD	1_2_0 START	1_2_12 U_DestHelp	0.1035
		1_2_20 U_FindAdd	0.8965
1_2_12 U_DestHelp	1_2_1 M_CMD 1_2_33 M_Resp_U_Timeout		
		1_2_19 M_DestHpCMD	0.9261
		1_2_33 M_Resp_U_Timeout	0.0769
1_2_21 M_CityInfo	1_2_20 U_FindAdd	1_2_3 U_CityName	0.8393
		1_2_22 U_CMD_ChangeState	0.1607
1_2_23 U_StateName	1_2_18 M_StateInfo	1_2_24 M_State_MultiChoice	0.6900
		1_2_2 M_FirmState_CityInfo	0.3100

1_2_3 U_CityName	1_2_21 M_CityInfo 1_2_2 M_FirmState_CityInfo		
		1_2_4 M_City_MultiChoice	0.8730
		1_2_6 M_FirmCity_StreetInfo	0.1270
1_2_4 M_City_MultiChoice	1_2_3 U_CityName	1_2_32 ForgotState_Dummy	0.0050
		1_2_29_U_City_Goback	0.1000
		1_2_5 U_SelectCity	0.8950
1_2_29_U_City_Goback	1_2_4 M_City_MultiChoice 1_2_6 M_FirmCity_StreetInfo		
		1_2_3 U_CityName	
1_2_6 M_FirmCity_StreetInfo	1_2_3 U_CityName		
	1_2_5 U_SelectCity		
		1_2_29_U_City_Goback	0.1000
		1_2_7 U_StreetName	0.9000
1_2_7 U_StreetName	1_2_6 M_FirmCity_StreetInfo 1_2_31 M_EnterStreet		
		1_2_8 M_Street_MultiChoice	0.6531
		1_2_10 M_ConfirmStreet_HouseInfo	0.3469
1_2_8 M_Street_MultiChoice	1_2_7 U_StreetName	1_2_9 U_SelectStreet	0.9000
		1_2_30 U_Street_Goback	0.1000
1_2_10 M_ConfirmStreet_HouseInfo	1_2_7 U_StreetName		
	1_2_9 U_SelectStreet		
		1_2_11 U_HouseNumb	
1_2_11 U_HouseNumb	1_2_10 M_ConfirmStreet_HouseInfo 1_2_28 M_AndHouseN		
		1_2_13 M_FirmHouse_MapGuide	
1_2_13 M_FirmHouse_MapGuide	1_2_11 U_HouseNumb	1_2_27 U_HouseN_GoBack	0.1300
		1_2_14 U_CMD_Guide	0.8700
1_2_14 U_CMD_Guide	1_2_13 M_FirmHouse_MapGuide 1_2_26 M_Guide_TimeoutBarge		
		1_2_26 M_Guide_TimeoutBarge	0.1600
		1_2_15 M_StartGuide	0.8400
1_2_999 END	1_2_17 M_RouteGuidance	1_5 Residential_Try_Count	N=1, p = 0.2273 N=2, p =

			0.4000
		1_999 END	N=1, p = 0.7727
			N=2, p = 0.6000
			N=3, p = 1
			N=4, p = 1
			N=5, P = 1
1_3_1 M_CMD	1_3_0 START	1_3_2 U_DestHelp	0.6774
		1_3_4 U_AddressBook	0.3226
1_3_2 U_DestHelp	1_3_1 M_CMD		
	1_3_18 M_CMD_Example		
		1_3_18 M_CMD_Example	0.0476
		1_3_19 M_IncorrectFeedback	0.0952
		1_3_3 M_DestHpCMD	0.8572
1_3_4 U_AddressBook	1_3_1 M_CMD		
	1_3_20 U_CMD_Goback		
	1_3_6 M_Proc_NtPg		
	1_3_16 M_AddressBook_Example		
		1_3_7 M_SelectUser	0.9642
		1_3_16 M_AddressBook_Example	0.0358
1_3_12 U_StartGuide	1_3_11 M_MapGuidance		
	1_3_17 M_Guide_TimeoutBargein		
		1_3_13 M_StartGuidance	0.9284
		1_3_17 M_Guide_TimeoutBargein	0.0716
1_3_999	1_3_15 M_RouteGuidance	1_5 Residential_Try_Count	N=1, p = 0.2273
			N=2, p = 0.4000
		1_999 END	N=1, p = 0.7727
			N=2, p = 0.6000
			N=3, p = 1
			N=4, p = 1
			N=5, P = 1
1_4_1 M_CMD	1_4_0 START		
	1_4_17 M_DeactivateASR		
		1_4_2 U_DestHelp	0.8462
		1_4_5 U_PrevDest	0.1538
1_4_2 U_DestHelp	1_4_1 M_CMD		
	1_4_18 M_CMD_Example		
		1_4_18 M_Resp_U_Timeout	0.0769
		1_4_3 M_DestHpCMD	0.9231
1_4_5 U_PrevDest	1_4_1 M_CMD		
	1_4_3 M_DestHpCMD		
	1_4_20 M_PrevDest_Example		
		1_4_6 M_PrevDest_Line	0.8571

		1_4_19 U_Error_Dummy	0.1429
1_4_19 U_Error_Dummy	1_4_5 U_PrevDest	1_4_20 M_PrevDest_Example	0.5000
		1_4_17 M_DeactivateASR	0.5000
1_4_6 M_PrevDest_Line	1_4_5 U_PrevDest	1_4_7 U_NextPage	0.9200
		1_4_8 M_PrevDes_Example	0.0800
1_4_7 U_NextPage	1_4_6 M_PrevDest_Line 1_4_8 M_PrevDes_Example		
		1_4_8 M_PrevDes_Example	0.0800
		1_4_9 M_Proc_NtPg	0.9200
1_4_9 M_Proc_NtPg	1_4_7 U_NextPage	1_4_10 U_PrevDest_Select	
		1_4_7 U_NextPage	
1_4_12 U_StartGuide	1_4_13 M_MapGuidance 1_4_16 M_Guide_TimeoutBargein		
		1_4_13 M_StartGuidance	0.7143
		1_4_16 M_Guide_TimeoutBargein	0.2857
1_4_999	1_4_15 M_RouteGuidance	1_5 Residential_Try_Count	N=1, p = 0.2000
		1_999 END	N=1, p = 0.8000 N=2, p = 1 N=3, p = 1 N=4, p = 1 N=5, P = 1
2_4 NonResidential_Try_Count	2_1 START	2_5 Public_Method_Count_1	1st attempt
		2_6 Public_Method_Count_2	2nd attempt
		2_7 Public_Method_Count_3	3rd attempt
		2_8 Public_Method_Count_4	4th attempt
		2_9 Public_Method_Count_5	5th attempt
		2_10 Public_Method_Count_6	6th attempt
2_5 Public_Method_Count_1	2_4 NonResidential_Try_Count	2_1 Point of Interest	0.0636
		2_2 Street Address	0.9338
		2_17 Previous Destination	0.0026
2_6 Public_Method_Count_23	2_4 NonResidential_Try_Count	2_1 Point of Interest	0.1429
		2_2 Street Address	0.8214
		2_17 Previous Destination	0.0357
2_7 Public_Method_Count_3	2_4 NonResidential_Try_Count	2_1 Point of Interest	0.2857
		2_2 Street Address	0.5714
		2_17 Previous Destination	0.1429
2_8 Public_Method_Count_4	2_4 NonResidential_Try_Count	1_2 Street Address	1.0000

2_9 Public_Method_Count_5	2_4 NonResidential_Try_Count	2_2 Street Address	1.0000
2_10 Public_Method_Count_6	2_4 NonResidential_Try_Count	2_999 End	1.0000
2_1_1 M_CMD	2_1_0 START	2_1_2 U_DestHelp 2_1_4 U_CMD_POI 2_1_22 U_CMDPOICat	0.0714 0.8572 0.0714
2_1_4 U_CMD_POI	2_1_1 M_CMD 2_1_3 M_DestHpCMD	2_1_5 M_CMD_POICat	
2_1_7 M_CMDPOISubCat	2_1_22 U_CMDPOICat 2_1_6 U_POICat 2_1_14 U_CMD_GoBack	2_1_12 U_NtPg_POISubCat 2_1_8 U_POISubCat	0.5000 0.5000
2_1_12 U_NtPg_POISubCat	2_1_13 M_Proc_NtPg_POISubVat 2_1_7 M_CMDPOISubCat	2_1_13 M_Proc_NtPg_POISubVat	
2_1_13 M_Proc_NtPg_POISubVat	2_1_12 U_NtPg_POISubCat	2_1_12 U_NtPg_POISubCat 2_1_8 U_POISubCat 2_1_14 U_CMD_GoBack 2_1_21 U_CMD_Cancel	0.0600 0.9000 0.0300 0.0100
2_1_14 U_CMD_GoBack	2_1_13 M_Proc_NtPg_POISubVat	2_1_7 M_CMDPOISubCat	
2_1_21 U_CMD_Cancel	2_1_13 M_Proc_NtPg_POISubVat	2_1_999 END	
2_1_8 U_POISubCat	2_1_7 M_CMDPOISubCat 2_1_13 M_Proc_NtPg_POISubVat	2_1_9 M_CMDLine	
2_1_9 M_CMDLine	2_1_8 U_POISubCat	2_1_10 U_NtPg_Choice 2_1_15 U_LineNumb	0.9500 0.0500
2_1_10 U_NtPg_Choice	2_1_9 M_CMDLine 2_1_11 ProcNtPg_Choice	2_1_11 ProcNtPg_Choice	
2_1_11 ProcNtPg_Choice	2_1_10 U_NtPg_Choice	2_1_10 U_NtPg_Choice 2_1_15 U_LineNumb	0.1000 0.9000
2_1_15 U_LineNumb	2_1_9 M_CMDLine 2_1_11 ProcNtPg_Choice	2_1_16 M_MapGuide	
2_1_17 U_CMD_Guide	2_1_16 M_MapGuide 2_1_23 M_Guide_TimeoutBarge	2_1_23 M_Guide_TimeoutBarge 2_1_18 M_StartGuide	0.1600 0.8400
2_1_23 M_Guide_TimeoutBarge	2_1_17 U_CMD_Guide	2_1_17 U_CMD_Guide	
2_1_999 END	2_1_21 U_CMD_Cancel 1_4_15 M_RouteGuidance		

2_4 Residential_Try_Count	N=1, p = 0.7778 N=2, p = 0.7500 N=3, p=0.7500
2_999 END	N=1, p = 0.2222 N=2, p = 0.2500 N=3, p = 0.2500 N=4, p = 1 N=5, P = 1

The table provides task time for the task networks for music selection tasks from Figure 5-9 to Figure 5-12.

Table A-3. Task Description and Task Time of the Model for Music Selection Task

Task	Description	Task Time
0 Model START	Determine the path of the subtask.	
1 FindAlbum	A function that contains sub-networks of tasks to search for a specific album.	
2 FindArtist	A function that contains sub-networks of tasks to search for a specific artist.	
3 FindSongt	A function that contains sub-networks of tasks to search for a specific song.	
1_1 Alb_S_TnkUtt_T	The user thinks of and utters the information to search for a specific album.	<u>Think time</u> 1st attempt T = 5.323 + 2.131*Age else T = Lognormal (1.863, 0.551) <u>Utterance Time</u> T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
1_4 Alb_M_Prompt_T	The system processes and provides the feedback to the user.	<u>Process time</u> T = 1.711 – 0.49*Correct_MS – 0.079*Age <u>Utterance Time</u> T = Chime + Pause + (0.559 + 0.091* NSyllable + 0.084 * NWord)
2_1 Art_S_TnkUtt_T	The user thinks of and utters a request to search for a specific artist.	<u>Think time</u> 1st attempt T = 5.323 + 2.131*Age else T = Lognormal (1.863, 0.551)

		Utterance Time T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
2_4 Alb_M_Prompt_T	The system processes and provides the feedback to the user.	Process time T = 1.711 – 0.49*Correct_MS – 0.079*Age Utterance Time T = Chime + Pause + (0.559 + 0.091* NSyllable + 0.084 * NWord)
3_0 START	Determine the music selection method.	
3_1 SAlb_S_Thinking_T	The user chooses the *Album* method to search a specific song.	1st attempt T = 5.323 + 2.131*Age else T = Lognormal (1.863, 0.551)
3_7 SAlb_S_Prompt_T	The user utters the command and album name.	T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
3_9 SAlb_M_ProcUtt_T	The system processes the user's utterance, provides the feedback to the user, and plays the music.	Process time T = 1.711 – 0.49*Correct_MS – 0.079*Age Utterance Time T = Chime + Pause + (0.559 + 0.091* NSyllable + 0.084 * NWord)
3_14 SAlb_M_Next_Proc	The number of times that the user needs to say the command *next track* and machine processes the user's utterance until finding the requested song.	Think time T = Lognormal (1.863, 0.551) User Utterance Time T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age Machine Process time T = Weibull (0.69, 0.15)
3_2 SArt_S_Thinking_T	The user chooses the *Artist* method to search a song.	1st attempt T = 5.323 + 2.131*Age else T = Lognormal (1.863, 0.551)
3_11 SAlb_S_Prompt_T	The user utters the command and artist name.	T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
3_13 SAlb_M_ProcUtt_T	The system processes the user's utterance, provides the feedback to the user, and plays the music.	Process time T = 1.711 – 0.49*Correct_MS – 0.079*Age Utterance Time T = Chime + Pause + (0.559 + 0.091* NSyllable + 0.084 * NWord)
3_16 SAlb_M_Next_Proc	The number of times that the user needs to say the command *next track* and machine processes the user's utterance until finding the requested song	Think time T = Lognormal (1.863, 0.551) User Utterance Time T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age

		<u>Machine Process time</u> T = Weibull (0.69, 0.15)
3_3 SSong_S_TnkUtt_T	The user thinks of and utters the command phrase and song title to search for a specific song using the *Song* method.	<u>Think time</u> 1^{st} attempt T = 5.323 + 2.131*Age else T = Lognormal (1.863, 0.551) <u>Utterance Time</u> T = 0.341 + 0.137 * NSyllable + 0.165 * Word - 0.221 * Workload + 0.146*Age
3_6 SSong_M_ProcUtt_T	The system processes the user's utterance and provides the feedback to the user.	<u>Process time</u> T = 1.711 – 0.49*Correct_MS – 0.079*Age <u>Utterance Time</u> T = Chime + Pause + (0.559 + 0.091* NSyllable + 0.084 * NWord)
3_17 SSong_Try 1	The 1^{st} attempt that the user searches a specific song using the *Song* method.	
3_18 SSong_Try 2	The 2^{nd} attempt that the user searches a specific song using the *Song* method.	
3_19 SSong_Try 3	The 3^{rd} attempt that the user searches a specific song using the *Song* method.	
3_20 Song_Dummy	The dummy node the user chooses the *Song* method to search a specific song.	
3_21 Art_Dummy	The dummy node the user chooses the *Artist* method to search a specific song.	
3_22 Alb_Dummy	The dummy node the user chooses the *Album* method to search a specific song.	

The table provides transition probabilities for the task networks of music selection tasks from Figure 5-9 to Figure 5-12.

Table A-4. The Transition of Probabilities of Task Sequences for Music Selection Tasks

Task	Proceeding Task	Following Task	Probability
0 Model STRAT		1FindAlbum 2 FindArtist 3 FindSong	0.2 0.2 0.6
1_1 Alb_S_TnkUtt_T	1_0 START 1_4 Alb_M_Prompt_T	1_4 Alb_M_Prompt_T	
1_4 Alb_M_Prompt_T	1_1 Alb_S_TnkUtt_T	1_1 Alb_S_TnkUtt_T	1^{st} attempt, p = 0.1500 2^{nd} attempt, p = 0.5000

		1_999 END	1st attempt, p = 0.8500 2nd attempt, p = 0.500 3rd attempt, p = 1.000
2_1 Art_S_TnkUtt_T	2_0 START 2_4 Art_M_Prompt_T	2_4 Alb_M_Prompt_T	
2_4 Art_M_Prompt_T	2_1 Art_S_TnkUtt_T	2_1 Art_S_TnkUtt_T	1st attempt, p = 0.1200 2nd attempt, p = 0.500
		2_999 END	1st attempt, p = 0.8800 2nd attempt, p = 0.500 3rd attempt, p = 1.000
3_0 START	0 Model START	3_1 SAlb_S_Thinking_T	0.4479
		3_2 SArt_S_Thinking_T	0.4688
		3_3 SSong_S_TnkUtt_T	0.0833
3_1 SAlb_S_Thinking_T	3_0 START 3_9 SAlb_M_ProcUtt_T 3_14 SAlb_M_Next_Proc 3_22 Alb_Dummy 3_13 SAlb_M_ProcUtt_T 3_16 SAlb_M_Next_Proc	3_7 SAlb_S_Prompt_T	
3_9 SAlb_M_ProcUtt_T	3_7 SAlb_S_Prompt_T	3_1 SAlb_S_Thinking_T	Tactical path
		3_14 SAlb_M_Next_Proc	Tactical path
		3_2 SArt_S_Thinking_T	Tactical path
		3_999 END	Tactical path
3_14 SAlb_M_Next_Proc	3_9 SAlb_M_ProcUtt_T	3_14 SAlb_M_Next_Proc	Tactical path
		3_1 SAlb_S_Thinking_T	Tactical path
		3_999 END	Tactical path
3_2 SArt_S_Thinking_T	3_0 START 3_9 SAlb_M_ProcUtt_T 3_13 SAlb_M_ProcUtt_T 3_16 SAlb_M_Next_Proc 3_21 Art_Dummy	3_11 SAlb_S_Prompt_T	
3_13 SAlb_M_ProcUtt_T	3_11 SAlb_S_Prompt_T	3_16 SAlb_M_Next_Proc	Tactical path
		3_2 SArt_S_Thinking_T	Tactical path
		3_1 SAlb_S_Thinking_T	Tactical path
		3_999 END	Tactical path
3_16	3_13	3_16	Tactical path

SAlb_M_Next_Proc	SAlb_M_ProcUtt_T	SAlb_M_Next_Proc	
		3_2 SArt_S_Thinking_T	Tactical path
		3_1 SAlb_S_Thinking_T	Tactical path
		3_3 SSong_S_TnkUtt_T	Tactical path
		3_999 END	Tactical path
3_3 SSong_S_TnkUtt_T	3_0 START		
	3_16 SAlb_M_Next_Proc		
	3_20 Song_Dummy		
		3_6 SSong_M_ProcUtt_T	1.0000
3_6 SSong_M_ProcUtt_T	3_3 SSong_S_TnkUtt_T	3_17 SSong_Try 1	1st attempt
		3_18 SSong_Try 2	2nd attempt
		3_19 SSong_Try 3	3rd attempt
3_17 SSong_Try 1	3_6 SSong_M_ProcUtt_T	3_20 Song_Dummy	0.2000
		3_21 Art_Dummy	0.6000
		3_22 Alb_Dummy	0.2000
3_18 SSong_Try 2	3_6 SSong_M_ProcUtt_T	3_21 Art_Dummy	0.5000
		3_22 Alb_Dummy	0.5000
3_19 SSong_Try 3	3_6 SSong_M_ProcUtt_T	3_21 Art_Dummy	1.0000
3_20 Song_Dummy	3_17 SSong_Try 1	3_3 SSong_S_TnkUtt_T	1.0000
3_21 Art_Dummy	3_17 SSong_Try 1		
	3_18 SSong_Try 2		
	3_19 SSong_Try 3		
		3_2 SArt_S_Thinking_T	1.0000
3_22 Alb_Dummy	3_17 SSong_Try 1		
	3_18 SSong_Try 2		
		3_1 SAlb_S_Thinking_T	1.0000

References

[1] M. Heron, "Deaths: Leading Causes for 2009," *National Viatl Statistics Reports,* vol. 61, pp. 1-95, 2012.

[2] U. S. Census Bureau, "Statistical Abstract of the United States: 2012 (131st Edition)," Department of Commerce. Washington D.C., 2011.

[3] J. C. Stutts, D. W. Reinfurt, L. Staplin, and E. A. Rodgman, "The Role of Driver Distraction in Traffic Crashes," AAA Foundation for Traffic Safety. Washington D.C., 2001.

[4] S. G. Klauter, Dingus, T.A., Neale, V.L., Sudweeks, J.D., and Ramsey, D.J., "The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data." (DOT HS 810 594), National Highway Traffic Safety Administration. Washington, D.C., 2006.

[5] C. Carter and R. Graham, "Experimental Comparison of Manual and Voice Controls for the Operation of In-Vehicle Systems," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA, 2000, pp. 286-289.

[6] C. Forlines, B. Schmidt-Nielsen, B. Raj, K. Wittenburg, and P. Wolf, "A Comparison Between Spoken Queries And Menu-Based Interfaces for In-Car Digital Music Selection," in *International Conference on Human-Computer Interaction (INTERACT)*. Rome, Italy, 2005, pp. 536-549.

[7] K. Itoh, Y. Miki, N. Yoshitsugu, N. Kubo, and S. Mashimo, "Evaluation of a Voice-Activated System Using a Driving Simulator," in *Society of Automotive Engineers(2004-01-0232)*. Warrendale, PA, 2004.

[8] J. Maciej and M. Vollrath, "Comparison of Manual vs. Speech-Based Interaction with In-Vehicle Information Systems," *Accident Analysis and Prevention,* vol. 41, pp. 924-930, Sep 2009.

[9] W. Minker, U. Haiber, P. Heisterkamp, and S. Scheible, "The SENECA Spoken Language Dialogue System," *Speech Communication,* vol. 43, pp. 89-102, Jun 2004.

[10] T. A. Ranney, J. L. Harbluk, and Y. I. Noy, "Effects of Voice Technology on Test Track Driving Performance: Implications for Driver Distraction," *Human Factors,* vol. 47, pp. 439-454, Sum 2005.

[11] J. Shutko, Mayer, K., Laansoo, E., and Tijerina L., "Driver Workload Effects of Cell Phone, Music Player, and Text Messaging Tasks with the Ford SYNC Voice Interface versus Handheld Visual-Manual Interfaces," in *Society of Automotive Engineerings (2009-01-0786)*. Warrendale, PA, 2009.

[12] O. Tsimhoni, D. Smith, and P. Green, "Address Entry While Driving: Speech Recognition versus a Touch-Screen Keyboard," *Human Factors,* vol. 46, pp. 600-610, Win 2004.

[13] M. C. McCallum, J. L. Campbell, J. B. Richman, J. L. Brown, and E. Wiese, "Speech Recognition and In-Vehicle Telematics Devices: Potential Reductions in Driver Distraction," *International Journal of Speech Technology,* vol. 7, pp. 25-33, 2004.

[14] J. D. Lee, B. Caven, S. Haake, and T. L. Brown, "Speech-Based Interaction With In-Vehicle Computers: The Effect of Speech-Based E-mail on Drivers' Attention to The Roadway," *Human Factors,* vol. 43, pp. 631-640, Win 2001.

[15] D. A. Salvucci, D. Markley, M. Zuber, D. P. Brumby, and Acm, "iPod Distraction: Effects of Portable Music-Player Use on Driver Performance," in *Conference on Human Factors in Computing Systems, Vol 1 and 2*. New York, 2007, pp. 243-250.

[16] J. C. Chang, A. Lien, B. Lathrop, and H. Hees, "Usability Evaluation of A Volkswagen Group In-Vehicle Speech System," in *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Essen, Germany, 2009.

[17] V. E.-W. Lo and P. Green, "Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and Related Literature," *International Journal of Vehicular Technology,* 2013 (Accepted).

[18] F. Weng, S. Varges, B. Raghunathan, F. Ratiu, H. Pon-Barry, B. Lathrop, Q. Zhang, H. Bratt, T. Scheideck, K. Xu, M. Purver, R. Mishra, A. Lien, M. Raya, S. Peters, Y. Meng, J. Russell, L. Cavedon, E. Shriberg, H. Schmidt, and R. Prieto, "CHAT: A Conversational Helper for Automotive Tasks," in *Proceddings of the 9th International Conference on Spoken Language Processing (Inter-Speech/ICSLP)*. Pittsburgh, PA, 2006, pp. 1061-1064.

[19] F. Weng, B. Yan, Z. Feng, F. Ratiu, M. Raya, B. Lathrop, A. Lien, A. Varges, M. Mishra, F. Lin, M. Purver, H. Bratt, Y. Meng, S. Peters, T. Scheideck, B. Raghunathan, and Z. Zhang, "CHAT To Your Destination," In Proceedings of 8th SIGdial Workshop on Discourse and Dialogue. Antwerp, Belgium, 2007.

[20] J. H. L. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward, ""CU-Move": Robust Speech Processing for In-Vehicle Speech System," in *ICSLP-2000: International Conference on Spoken Language Processing*. Beijing, China, 2000, pp. 524 - 527.

[21] R. Pieraccini, K. Dayanidhi, J. Bloom, J.-G. Dahan, M. Phillips, B. R. Goodman, and K. V. Prasad, "Multimodal Conversational Systems for Automobiles," *Commun. ACM,* vol. 47, pp. 47-49, 2004.

[22] P. Heisterkamp, "Linguatronic Product-Level Speech System for Mercedes-Benz Cars," *Proceedings of the First International Conference on Human Language Technology Research*. San Diego, CA, 2001.

[23] (January 29). *SYNC*. Available: http://www.ford.com/syncmyride/ - /home/

[24] P. Geutner, F. Steffens, and D. Manstetten, "Design of the VICO Spoken Dialogue System: Evaluation of User Expectations by Wizard-of-Oz Experiments," *Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain, 2002.

[25] V. E.-W. Lo, P. A. Green, and A. Franzblau, "Where Do People Drive? Navigation System Use by Typical Drivers and Auto Experts," *Journal of Navigation,* vol. 64, pp. 357-373, Apr 2011.

[26] U. Winter, T. J. Grost, and O. Tsimhoni, "Language Pattern Analysis for Automotive Natural Language Speech Applications," in *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Pittsburgh, PA, 2010.

[27] K. Takeda, J. H. L. Hansen, P. Boyraz, L. Malta, C. Miyajima, and H. Abut, "International Large-Scale Vehicle Corpora for Research on Driver Behavior on the Road," *IEEE Transactions on Intelligent Transportation Systems,* vol. 12, pp. 1609-1623, Dec 2011.

[28] A. Barón and P. A. Green, "Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review," University of Michigan Transportation Research Institute (UMTRI-2006-5). Ann Arbor, MI, 2006.

[29] B. Pellom, W. Ward, J. Hansen, R. Cole, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan, "University of Colorado Dialog Systems for Travel and Navigation," *Proceedings of the 1st International Conference on Human Language Technology Research*. San Diego, CA, 2001.

[30] L. Garay-Vega, A. K. Pradhan, G. Weinberg, B. Schmidt-Nielsen, B. Harsham, Y. Shen, G. Divekar, M. Romoser, M. Knodler, and D. L. Fisher, "Evaluation of Different Speech and Touch Interfaces to In-Vehicle Music Retrieval Systems," *Accident Analysis and Prevention,* vol. 42, pp. 913-920, May 2010.

[31] U. Gärtner, W. König, and T. Wittig, "Evaluation of Manual vs. Speech Input When Using a Driver Information System in Real Traffic.," in *Driving Assessment 2001: The First International Driving Symposium on Human Factors in Driving Assessment, Training and Vehicle Design*. Aspen, CO, 2001.

[32] O. Tsimhoni, D. Smith, and P. Green, "Destination Entry While Driving: Speech Recognition versus a Touch-Screen Keyboard," University of Michigan Transportation Research Institute (UMTRI-2001-24). Ann Arbor, MI, 2002.

[33] J. Villing, C. Holtelius, S. Larsson, A. Lindstrom, A. Seward, and N. Aberg, "Interruption, Resumption and Domain Switching in In-Vehicle Dialogue," in *6th International Conference on Natural Language Processing*. Gothenburg, Sweden, 2008, pp. 488-499.

[34] A. Kun, T. Paek, and Z. Medenica, "The Effect of Speech Interface Accuracy on Driving Performance," Interspeech 2007. Antwerp, Belgium, 2007.

[35] A. Gellatly and T. A. Dingus, "Speech Recognition and Automotive Applications: Using Speech to Perform In-Vehicle Tasks," in *Proceefings of the Human Factors 32nd Annual Meeting*. Santa Monica, CA, 1998, pp. 1247 - 1251.

[36] B. Faerber and G. Meier-Arendt, "Speech Control Systems for Handling of Route Guidance, Radio and Telephone in Cars: Results of a Field Experiment," in *Vision in Vehicle - VII*. Amsterdam, Netherlands, 1999, pp. 507 - 515.

[37] R. M. Schumacher, M. L. Hardzinski, and A. L. Schwartz, "Increasing the Usability of Interactive Voice Response Systems: Research and Guidelines for Phone-Based Interfaces," *Human Factors,* vol. 37, pp. 251-264, Jun 1995.

[38] *Intuity Conversant Voice Information System Version 5.0 Application Design Handbook* vol. Doc. 585-310-551. Denver, CO: AT&T Product Documentation Development, 1994.

[39] L. J. Najjar, J. J. Ockeman, and J. C. Thompson, "User Interface Design Guidelines for Speech Recognition Applications," IEEE VARIS 98 Workshop. Atlanta, GA, 1998.

[40] Z. Hua and W. L. Ng, "Speech Recognition Interface Design for In-Vehicle System," *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Pittsburgh, PA, 2010.

[41] ISO. ISO Standard 9921: Ergonomics – Assessment of Speech Communication. International Organization for Standardization: Geneva, Switzerland, 2003.
[42] ISO. ISO/TR Technical Report 19358: Ergonomics – Construction and Application of Tests for Speech Technology. International Organization for Standardization: Geneva, Switzerland, 2002.
[43] ISO. ISO/IEC Standard 2382-29: Information Technology – Vocabulary – Part 29: Artificial Intelligence – Speech Recognition and Synthesis. International Organization for Standardization: Geneva, Switzerland, 1999.
[44] ISO. ISO Standard 8253-3: Acoustics – Audiometric Test Methods – Part 3: Speech Audiometry. International Standards Organization: Geneva, Switzerland, 2012.
[45] ISO. ISO Standard 16982: Ergonomics of Human-System Interaction – Usability Methods Supporting Human-Centered Design. International Organization for Standardization: Geneva, Switzerland, 2002.
[46] ISO. ISO Standard 15006: Road Vehicles – Ergonomic Aspects of Transport Information and Control Systems – Specifications for In-Vehicle Auditory Presentation. International Organization for Standardization: Geneva, Switzerland, 2011.
[47] SAE International. SAE Recommended Practice J2988: Voice User Interface Principles and Guidelines (Draft). Society of Automotive Engineers: Warrendale, PA, 2012.
[48] S. C. Levinson, *Pragmatics*. Cambridge, NY: Cambridge University Press, 1983.
[49] J. L. Austin, *How to Do Things with Words*. Cambridge, MA: Harvard University Press, 1962.
[50] A. Akmajian, R. A. Demers, A. K. Farmer, and R. M. Harnish, *Linguistics: An Introduction to Language and Communication*, 5th ed. Cambridge, MA: MIT Press, 2001.
[51] J. R. Searle, "A Taxonomy of Illocutionary," in *Language, Mind and Knowledge*. vol. VII, K. Gunderson, Ed., ed: Minnesota Studies in the Philosophy of Science, 1975, pp. 344-369.
[52] H. P. Grice, "Logic and Conversation," in *Syntax and Semantics 3: Speech Acts*, P. Coole and J. L. Morgan, Eds., ed New York, N.Y.: Academic Press, 1975, pp. 41-58.
[53] G. Skanztze. (2002, Error Detection in Spoken Dialogue Systems. Available: http://www.speech.kth.se/~gabriel/courses/DialogueSystems5p/ErrorDetection.pdf
[54] J. Véronis, "Error in Natural Language Dialogue Between Man and Machine," *International Journal of Man-Machine Studies,* vol. 35, pp. 187-217, Aug 1991.
[55] N. Chomsky, *Aspects of Theory of Syntax*. Cambridge, MA: The MIT Press, 1965.
[56] M. L. Bourguet, "Towards a Taxonomy of Error-Handling Strategies in Recognition-Based Multi-Modal Human-Computer Interfaces," *Signal Processing,* vol. 86, pp. 3625-3643, Dec 2006.
[57] C.-M. Karat, C. Halverson, D. Horn, and J. Karat, "Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Pittsburgh, PA, 1999.

[58] K. Larson and D. Mowatt, "Speech Error Correction: The Story of the Alternates List," *International Journal of Speech Technology,* vol. 6, pp. 183-194, 2003.

[59] D. Litman, M. Swerts, and J. Hirschberg, "Characterizing and Predicting Corrections in Spoken Dialogue Systems," *Computational Linguistics,* vol. 32, pp. 417-438, Sep 2006.

[60] V. E.-W. Lo, S. M. Walls, and P. A. Green, "Simulation of iPod Music Selection by Drivers: Typical User Task Time and Patterns for Manual and Speech Interfaces," University of Michigan Transportation Research Institute (UMTRI-2007-9). Ann Arbor, MI, 2007.

[61] J. F. Kelley, "An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications," Proceedings of the SIGCHI conference on Human Factors in Computing Systems. Boston, MA, 1983.

[62] J. D. Gould, J. Conti, and T. Hovanyecz, "Composing Letters with a Simulated Listening Typewriter," *Communications of the AcCM,* vol. 26, pp. 295-308, 1983.

[63] P. A. Green and L. Wei-Hass, "The Wizard of Oz: A Tool for Repaid Development of User Interfaces," University of Michigan Transportation Research Institute (UMTRI-1985-27). Ann Arbor, MI, 1985.

[64] A. K. Sinah, S. R. Klemmer, J. Chen, J. A. Landay, and C. Chen, "Suede: Iterative, informal prototyping for speech interfaces," in *CHI 2001*. New York, NY, 2001.

[65] L. Dybkjær, N. O. Bernsen, and W. Minker, "Evaluation and Usability of Multimodal Spoken Language Dialogue Systems," *Speech Communication,* vol. 43, pp. 33-54, Jun 2004.

[66] M. Walker, C. Kamm, and D. Litman, "Towards Developing General Models of Usability with PARADISE," *Natural Language Engineering,* vol. 6, pp. 363-377, 2000.

[67] M. Hajdinjak and F. Mihelic, "The PARADISE Evaluation Framework: Issues and Findings," *Computational Linguistics,* vol. 32, pp. 263-272, Jun 2006.

[68] J. Schweitzer and P. A. Green, "Task Acceptability and Workload of Driving Urban Roads, Highways, and Expressway: Ratings from Video Clips," University of Michigan Transportation Research Institute (UMTRI-2006-6). Ann Arbor, MI, 2007.

[69] P. Green, B. T.-W. Lin, J. Schweitzer, H. Ho., and K. Stone, "Evaluation of A Method to Estimate Driving Workload in Real Time: Watching Clips versus Simulated Driving," University of Michigan Transportation Research Institute (UMTRI-2011-29). Ann Arbor, MI, 2011.

[70] P. Green, "Using Standards to Improve the Replicability and Applicability of Driver Interface Research," in *Automotive UI 2012*. Portsmouth, NH, 2012.

[71] M. R. Savino, "Standardized Names and Definitions for Driving Performance Measures," M.S., Dept. Mech. Eng., Tuffs University, Medford, MA, 2009.

[72] SAE International. SAE Recommended Practice J2944: Operational Definitions of Driving Performance Measures and Statistics (Draft). Society of Automotive Engineers: Warrendale, PA, 2012.

[73] G. Burnett, "'Turn Right at the Traffic Lights': The Requirement for Landmarks in Vehicle Navigation Systems," *Journal of Navigation,* vol. 53, pp. 499-510, Sep 2000.

[74] D. W. Eby and L. J. Molnar, "In-Vehicle Route Guidance Preferences of Driving Tourists," *ITS Journal,* vol. 6, pp. 261-279, 2001.

[75] A. J. May, T. Ross, and S. H. Bayer, "Drivers' Information Requirements When Navigating in An Urban Environment," *Journal of Navigation,* vol. 56, pp. 89-100, Jan 2003.

[76] W. C. Lee, M. C. Ma, and B. W. Cheng, "Field Comparison of Driving Performance Using a Portable Navigation System," *Journal of Navigation,* vol. 63, pp. 39-50, Jan 2010.

[77] P. S. Hu and T. R. Reuscher, "Summary of Travel Trends: 2001 National Household Travel Survey," U.S. Department of Transportation
Federal Highway Administration. Washington, D.C., 2004.

[78] D. R. McElhaney, C. Purvis, R. Tweedie, P. Fulton, A. Pisarski, and F. E. Jarema, "National Travel Surveys. FHWA Study Tour", 1994.

[79] Alliance of Automotive Manufacturers. Statement of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication System. Washington, D.C., 2006.

[80] SAE International. SAE Recommended Practice J2364: Navigation and Route Guidance Function Accessibility While Driving. Society of Automotive Engineers: Warrendale, PA, 2004.

[81] Department of Trasnport, "Transport Statistics Bulletin - National Travel Survey: 2008" Department of Trasnport. London: United Kingdom, 2009.

[82] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Anderson-Wesley, Inc., 1977.

[83] S. M. Walls, A. Barón, and P. A. Green, "Simulation of Destination Entry by Drivers: Typical User Task Time and Patterns for Manual and Speech Interfaces," University of Michigan Transportation Research Institute (UMTRI-2007-7). Ann Arbor, MI, 2007.

[84] D. Manstetten, W. Krautter, B. Grothkopp, F. Steffens, and P. Geutner, "Using a Driving Simulator to Perform a Wizard-of-Oz Eexerimrnt on Speech-Controlled Driver Information Systems," *1st Human-Centered Transportation Simulation Conference (HCTSC2001)*. Iowa City, Iowa, 2001.

[85] L. Hsieh, R. Young, and S. Seaman, "Development of the Enhanced Peripheral Detection Task: A Surrogate Test for Driver Distraction," *SAE International Journal of Passenger Cars - Electronic and Electrical Systems,* vol. 5, pp. 317-325.

[86] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language,* vol. 50, pp. 696-735, 1974.

[87] B. E. John, "Extensions of GOMS Analyses to Expert Performance Requiring Perception of Dynamic Visual and Auditory Information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Seattle, WA, 1990, pp. 107-116.

[88] J. Östlund, B. r. Peters, B. Thorslund, J. Engström, G. Markkula, A. Keinath, D. Horst, S. Juch, S. Mattes, and U. Foehl, "Driving Performance Assessment - Methods and Metrics," Information Society Technologies (IST), 2005.

[89] J. Banks, Ed., *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. New York, NY: Wiley, 1998, p.^pp. Pages.

[90] B. P. Zeigler, H. Praehofer, and T. G. Kim, *Theory of Modeling and Simulation*, 2nd ed. New York, NY: Academic Press, 2000.

[91] D. L. Fisher, J. Caird, and J. D. Lee, Eds., *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. Boca Raton, FL: CRC Press, 2011.

[92] A. J. Rehmann, R. D. Mitman, and M. C. Reynolds, "A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research," Crew System Ergonomics Information Analysis Center (CSERIAC). Wright-Patterson Air Force Base, OH DOT/FAA/CT-TN95/46, 1995.

[93] A. G. Gallagher and G. C. O'Sullivan, *Fundamentals of Surgical Simulation: Principles and Practice*. London, UK: Springer-Verlag, 2012.

[94] A. Yow, B. Walters, B. Plott, R. Laughery, and J. Persensky, "Predicting Nuclear Power-plant Operator Performance Using Discrete Event Simulation," *Cognitive, Technology, and Work,* vol. 7, pp. 29-35, 2005.

[95] W. C. Brainard and J. Tobin, "Pitfalls in Financial Model Building," *American Economic Review,* vol. 58, pp. 99-122, 1968.

[96] C. W. Richardson and D. A. Wright, "WGEN: A Model for Generating Daily Weather Variables.," U.S. Department of Agriculture, Agricultural Research Service. Washington D.C., ARS-8, 1984.

[97] M. L. Brandeau, "Modeling and Simulation in Public Health: A Little Help Can Go a Long Way," in *Proceedings of the 2008 Winter Simulation Conference*. Austin, TX, 2008.

[98] P. M. J. Christie and R. R. Levary, "The Use of Simulation in Planning the Transportation of Patients to Hospitals Following a Disaster," *Journal of Medical Systems,* vol. 22, pp. 289-300, 1998.

[99] U. Raschke, "The Jack Human Simulation Tool," in *Working Postures and Movements: Tools for Evaluation and Engineering*, N. J. Delleman, C. M. Haslegrave, and D. B. Chaffin, Eds., ed Boca Raton, FL: CRC Press, 2004, pp. 431-437.

[100] K. M. Thompson, D. E. Burmaster, and E. A. C. Crouch, "Monte Carlo Techniques for Quantitative Uncertainty Analysis in Public Health Risk Assessments," *Risk Analysis,* vol. 12, pp. 53-63, 1992.

[101] D. Fone, S. Hollinghurst, M. Temple, A. Round, N. Lester, A. Weightman, K. Roberts, E. Coyle, G. Bevan, and S. Palmer, "Systematic Review of the Use and Value of Computer Simulation Modelling in Population Health and Health Care Delivery," *Journal of Public Health Volume,* vol. 25, pp. 325-335, 2003.

[102] C. Keane, *Modeling Behavior in Complex Public Health Systems: Simulation and Games for Action and Evaluation*: Springer Publishing, 2013.

[103] R. E. Nance, "A History of Discrete Event Simulation Programming Langaugues," Virginia Polytechnic Institute and State University, Department of Computer Science. Blacksburg, VA TR 93-21, 1993.

[104] J. Banks, J. S. Carson, B. L. Nelson, and D. M. Nicol, *Discrete-Event System Simulation*, 5th ed. New York, NY: Prentice Hall, 2009.

[105] R. W. Pew and A. S. Mavor, Eds., *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington D.C.: National Academy Press, 1999, p.^pp. Pages.

[106] P. A. Hancock, D. A. Vincenzi, J. A. Wise, and M. Mouloua, *Human Factors in Simulation and Training*. Boca Raton, FL: Taylor and Francis, 2009.

[107] D. K. Mitchell, "Workload Analysis of the Crew of the Abrams V2 SEP: Phase I Baseline IMPRINT Model," U.S. Army Research Laboratory. Aberdeen Proving Ground, MD ARL-TR-5028, 2009.

[108] B. P. Hunn and O. H. Heuckeroth, "A Shadow Unmanned Aerial Vehicle (UAV) Improved Performance Research Integration Tool (IMPRINT) Model Supporting Future Combat Systems," U.S. Army Research Laboratory. Aberdeen Proving Ground, MD ARL-TR-3731, 2006.

[109] D. K. Mitchell and C. Samms, "Workload Warriors: Lessons Learned from a Decade of Mental Workload Prediction Using Human Performance Modeling," in *Proceedings of the Human Factors and Ergonomics Society 53th Annual Meeting*. San Antonio, TX, 2009.

[110] D. K. Mitchell and J. Y. C. Chen, "Impacting System Design with Human Performance Modeling and Experiment: Another Success Story," in *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. San Francisco, CA, 2006.

[111] R. Laughery, "Using Discrete-event Simulation to Model Human Performance in Complex Systems," in *Proceedings of the 1999 Winter Simulation Conference*. Phoenix, AZ, 1999.

[112] D. W. Schunk, W. K. Bloechle, and R. Laughery, "Micro Saint: Micro Saint Modeling and the Human Element," in *Proceedings of the 2002 Winter Simulation Conference*. San Diego, CA, 2002.

[113] L. M. S. Dias, G. A. B. Pereira, P. Vik, and J. A. Oliveira, "Discrete Simulation Tools Ranking-A Commercial Software Packages Comparison Based on Popularity," in *ISC 2011, Industrial Simulation Conference*. Venice, Italy, 2011.

[114] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics*. Morristown, NJ, 1997, pp. 271-280.