Review Article

Development and Evaluation of Automotive Speech Interfaces: Useful Information from the Human Factors and the Related Literature

Victor Ei-Wen Lo1,2 and Paul A. Green1

1 Driver Interface Group, University of Michigan Transportation Research Institute, 2901 Baxter Road, Ann Arbor, MI 48109-2150, USA
2 Department of Environmental Health Sciences, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA

Correspondence should be addressed to Victor Ei-Wen Lo; loe@umich.edu

Received 11 October 2012; Revised 10 January 2013; Accepted 17 January 2013

Academic Editor: Motoyuki Akamatsu

Copyright © 2013 V. E.-W. Lo and P. A. Green. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Drivers often use infotainment systems in motor vehicles, such as systems for navigation, music, and phones. However, operating visual-manual interfaces for these systems can distract drivers. Speech interfaces may be less distracting. To help designing easy-to-use speech interfaces, this paper identifies key speech interfaces (e.g., CHAT, Linguatronic, SYNC, Siri, and Google Voice), their features, and what was learned from evaluating them and other systems. Also included is information on key technical standards (e.g., ISO 9921, ITU P.800) and relevant design guidelines. This paper also describes relevant design and evaluation methods (e.g., Wizard of Oz) and how to make driving studies replicable (e.g., by referencing SAE J2944). Throughout the paper, there is discussion of linguistic terms (e.g., turn-taking) and principles (e.g., Grice’s Conversational Maxims) that provide a basis for describing user-device interactions and errors in evaluations.

1. Introduction

In recent years, automotive and consumer-product manufacturers have incorporated speech interfaces into their products. Published data on the number of vehicles sold with speech interfaces is not readily available, though the numbers appear to be substantial. Speech interfaces are of interest because visual-manual alternatives are distracting, causing drivers to look away from the road, and increasing crash risk. Stutts et al. [1] reported that adjusting and controlling entertainment systems and climate-control systems and using cell phones accounted for 19% of all crashes related to distraction. The fact that the use of entertainment systems is ranked the second among major causes of these crashes arises the argument that speech interfaces should be used for music selection. Tsimhoni et al. [2] reported that 82% less time was needed for drivers to enter an address using a speech interface as opposed to using a keyboard, indicating that a speech interface is preferred for that task. However, using a speech interface still requires cognitive demand, which can interfere with the primary driving task. For example, Lee et al. [3] showed that drivers’ reaction time increased by 180 ms when using a complex speech-controlled email system (three levels of menus with four-to-seven options for each menu) in comparison with a simpler alternative (three levels of menus with two options per menu).

Given these advantages, suppliers and automanufacturers have put significant effort into developing speech interfaces for cars. They still have a long way to go. The influential automotive.com website notes the following [13]:

In the 2012…, the biggest issue found in today’s vehicles are the audio, infotainment, and navigation system’s lack of being able to recognize voice commands. This issue was the source of more problems than engine or transmission issues. … Over the four years that the survey questions people on voice recognition systems, problems have skyrocketed 137 percent.
Consumer Reports [14] said the following:

*I was feeling pretty good when I spotted that little Microsoft badge on the center console. Now I would be able to access all of those cool SYNC features, right? Wrong.*

When I tried to activate text to speech, I was greeted with a dreadful “Not Supported” display. Did I do something wrong? After all, my phone was equipped seemingly with every feature known to man. . . . But most importantly, it was powered by Microsoft just like the SYNC system on this 2011 Mustang.

Needing guidance, I went to Ford’s SYNC website. . . . I was able to download a 12-page PDF document that listed supported phones. (There is an interactive Sync compatibility guide here, as well.) While I had naively assumed that my high-tech Microsoft phone would work with all the features of “SYNC powered by Microsoft,” the document verified that this was not the case. . . . Text to speech would only work with a small handful of “dumbphones” that aren’t very popular anymore. Anyone remember the Motorola Razr? That phone was pretty cool a couple of years ago.

One consumer, in commenting about the Chrysler UConnect system said the following [15]:

*I have a problem with Uconnect telephone. I input my voice tags but when I then say “Call Mary” the system either defaults to my ‘phone book folder or I get 4–6 names on the screen and am asked to “select a line”. I should just say “call Mary home” then I should here my voice calling “Mary home is that correct”. Can you assist?*

Thus, it is appropriate to ask what is known now about the design and evaluation of speech interfaces for cars and how they can be improved. Most engineered systems rely on models, equations, and data to predict system performance and evaluate system alternatives early in development. They do not exist for speech interfaces. Thus, for speech interfaces, the emphasis has been on usability testing, often conducted when development is nearly complete and changes are costly to make.

To be more specific, this paper summarizes the state of the art relating to speech interface design in general, as well as a particular simulation model, namely, one to predict user performance when interacting with a speech interface for destination entry and music selection. The model to be developed will allow for exploration of multiple alternative architectures, recognition rates, and command sets, matters that are very expensive to explore experimentally.

The questions addressed in this paper are as follows.

1. What are some examples of automotive speech interfaces?
2. Who uses speech interfaces, for what, and how often?
3. What are the key research results of the user performance using speech interfaces compared with the user performance using visual-manual interfaces?
4. How should speech interfaces be designed? What are the key design standards and references, design principles, and results from research?
5. How should speech interfaces be assessed and what should be measured?

To answer these questions, both the human factors literature and the linguistics terms and principles are summarized.

2. What Are Some Examples of Automotive Speech Interfaces?

In the USA, current speech interfaces include Ford SYNC, Chrysler UConnect, GM MyLink, Hyundai Genesis, and Toyota navigation with Entune. The commonly supported applications are navigation (e.g., destination entry, route guidance, and traffic information) and music selection (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).

To help the reader understand these applications, some examples follow. The extent to which the user and machine are verbose varies between examples.

(Navigation)

*M: Command please.*

*U: Find address.*

*M: Find address. The city please. (Default State is Michigan) U: Michigan City.*

*M: Please select the respective line or start spelling. U: M I C H I G A N.*

*M: Spell the name in blocks of letters or enter your destination again with change state or enter city. U: Change state.*

*M: Please enter the state name. U: Indiana.*

*M: Please select the respective line or start spelling. U: Line one.*


*M: Wabash Street. And house number or if you do not know that please say show map or start guidance. U: Six zero one.*

Table 1: Examples of well-known automotive speech interfaces and the applications they support.

<table>
<thead>
<tr>
<th>System</th>
<th>Research/product</th>
<th>Navigation</th>
<th>Restaurant finder</th>
<th>Music selection</th>
<th>Audio + CD</th>
<th>Car information</th>
<th>Traffic information</th>
<th>Cellular phone</th>
<th>Text message</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAT [4, 5]</td>
<td>Research</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linguatronic [8]</td>
<td>Product</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SENECA [9]</td>
<td>Research</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYNC [10]</td>
<td>Product</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volkswagen [12]</td>
<td>Product</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Restaurant Finder) [4]

U: Can you find me a Chinese restaurant in Palo Alto with good service?
M: There are 11 Chinese restaurants in Palo Alto that have good service. Here are the first few: Little Garden Restaurant China Delight Mandarin Gourmet Restaurant. Should I continue?
U: Hmm...Tell me more about, Garden Restaurant?
M: Little Garden Restaurant is an expensive Chinese Restaurant. Dress Code is casual, service is good and rating is three stars. It is on El Camino Real in Palo Alto.
U: Switch to navigation.
M: Take me to little Garden Restaurant, but stop by Bosch Page Mill Office.

(Music Selection)

M: Please say a command.
U: USB.
M: USB. Is it correct?
U: Yes.
M: USB. Please say a command.
U: PLAY TRACK “Like a Prayer.”
M: Play track “Like a prayer.”

(Phone call)

M: Main menu.
U: Phone.
M: Phone. Please say a command.
U: Call XXX XXX XXXX.
M: Calling XXX XXX XXXX.

(Radio)

U: AM one thousand six hundred.
   [It did not work if the user said "AM Sixteen Hundred" for this interface.]
M: AM one thousand six hundred.

These systems were developed based on ideas from a number of predecessor systems (Tables 1 and 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 most of the attention.

The CHAT system uses an event-based, message-oriented system for the architecture with core modules of Natural Language Understanding (NLU), Dialogue Manager (DM), Content Optimization (CO), Knowledge Management (KM), and Natural Language Generation (NLG). CHAT uses the Nuance 8.5 speech recognition engine with class-based n-grams and dynamic grammars, and Nuance Vocalizer as the Text-to-Speech engine. There are three main applications—navigation, MP3 music player, and restaurant finder—to represent important applications in a vehicle [4, 5]. The example for restaurant finder shown earlier is a CHAT dialog.

CU-Move system is an in-vehicle, naturally spoken dialogue system, which can get real-time navigation and route-planning information [6]. The dialogue system is based on the MIT Galaxy-II Hub architecture with base components from CU-Communication system, which is mixed initiative and event driven. This system automatically retrieves the driving direction through Internet with route provider. The dialogue system uses the CMU Sphinx-II speech recognizer for speech recognition and Phoenix Parser for semantic parsing.

A prototype of a conversation system was implemented on the Ford Model U Concept Vehicle and was first shown in 2003 [7]. This system is used for controlling several noncritical automobile operations using speech recognition and a touch screen. The speech recognizer used in this system was speech2Go with adapted acoustic model and other enhancements to improve the speech accuracy. The dialogue manager was a multimodal version of the ETUDE, described by a recursive transition network. Supported applications were climate control, telephone, navigation, entertainment, and system preferences.

Linguatronic is a speech-based command and control system for telephone, navigation, radio, tape, CD, and other applications. The recognizer used in this device was speaker-independent [8].
SENCA SLDS consists of five units: COMMAND head unit connected via an optical Domestic Digital Bus to the Global System for Mobile Communication module, the CD Changer, and Digital Signal Processing module [9]. The system is a command-based speech control of entertainment (radio and CD), navigation, and cellular phones. The speech recognition technology of SENECA SLDS is based on the standard Linguatronic system using the following methods to match the user speech: spell matcher, Java Speech Grammar Format, voice enrollments (user-trained words), and text enrolments. For the dialogue processing, the ENCA SLDS uses a menu-based Command & Control dialogue strategy, including top-down access for main function and side access for subfunction.

SYNC is a fully integrated, voice-activated in-vehicle communication and entertainment system [10] for Ford, Lincoln, and Mercury vehicles in North America. Using commands in multiple languages, such as English, French or Spanish, drivers can operate navigation, portable digital music players, and Bluetooth-enabled mobile phones. The example for music selection shown earlier is a SYNC dialog.

VICO was a research project that concerned a natural-language dialogue prototype [11]. As the interface did not exist, researchers used the Wizard of Oz method to collect the human–computer interaction data. Here, a human operator, the wizard, was simulated system components—speech recognition, natural language understanding, dialogue modeling, and response generation. The goal of this project was to develop a natural language interface allowing drivers to get time, travel (navigation, tourist attraction, and hotel reservation), car, and traffic information safely while driving.

Volkswagen also developed its own in-vehicle speech system [12]. Detailed information about the architecture and methods used to design the system are not available. Supported applications include navigation and cellular phones.

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.

3. Who Uses Speech Interfaces, for What, and How Often?

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 might 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. [28] 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Full name</th>
<th>Developed by</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAT</td>
<td>Conversational Helper for Automotive Tasks</td>
<td>Center for the Study of Language and Information at Stanford</td>
</tr>
<tr>
<td></td>
<td></td>
<td>University, Research and Technology Center at Bosch, Electronics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Research Lab at Volkswagen of America, and Speech Technology and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Research Lab at SRI International [4, 5].</td>
</tr>
<tr>
<td>CU Move</td>
<td>Colorado University Move</td>
<td>University of Colorado speech group in 1999 [6].</td>
</tr>
<tr>
<td>Ford Model U</td>
<td></td>
<td>Ford [7].</td>
</tr>
<tr>
<td>Linguatronic</td>
<td></td>
<td>DaimlerChrysler Research and Technology in Ulm, Germany and TEMIC in 1996 [8].</td>
</tr>
<tr>
<td>SENECA</td>
<td>Speech control modules for Entertainment,</td>
<td>EU-project involving DaimlerChrysler, TEMIC Research and</td>
</tr>
<tr>
<td></td>
<td>Navigation, and communication Equipment in CArs</td>
<td>Department of Information Technology, University of Ulm [9].</td>
</tr>
<tr>
<td>SYNC</td>
<td></td>
<td>Ford in collaboration with Microsoft and Nuance [10].</td>
</tr>
<tr>
<td>VOIC</td>
<td>Virtual Intelligent Co-Driver</td>
<td>European project funded by five different partners: Robert Bosch</td>
</tr>
<tr>
<td>Volkswagen</td>
<td></td>
<td>GmbH, DaimlerChrysler AG, ITCirst, the University of Southern</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Volkswagen [12].</td>
</tr>
</tbody>
</table>
Table 3: Speech interface performance statistics from selected bench-top studies.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tasks</td>
<td>(1) NAV</td>
<td>(1) MP3</td>
<td>Phone for travel plan</td>
<td>NAV</td>
<td>Phone</td>
<td>NAV</td>
<td>NAV</td>
</tr>
<tr>
<td></td>
<td>(2) RF</td>
<td>(2) RF</td>
<td></td>
<td>phone</td>
<td>dialing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completion time (s)</td>
<td></td>
<td></td>
<td>260.3</td>
<td>Average: 63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completion rate</td>
<td>98%</td>
<td>MP3: 98%</td>
<td>73.6%</td>
<td>Average: 79%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turns^1</td>
<td>2.3^1</td>
<td></td>
<td>User: 19</td>
<td>Machine: 19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word recognition accuracy^2 (%)</td>
<td>NAV: 85.5%</td>
<td>MP3: 90%</td>
<td>RF: 94%</td>
<td>RF: 85%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word error rate (%)</td>
<td></td>
<td></td>
<td>26%</td>
<td>30–65%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User satisfaction rating^3</td>
<td>1.98</td>
<td>MP3: 2.24</td>
<td>RF: 2.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

^1Turn 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.

^2Word recognition accuracy (WA).

\[ WA = 100(1 - (Ws + Wi + Wd)/W) \%
\]

Ws: number of reference words which were substituted in output.
Wi: number of reference words which were inserted in output.
Wd: number of reference words which were deleted in output.

^3User satisfaction rating: 1 = strong agreement; 5 = strong disagreement.
Table 4: Driving performance statistics from selected studies (S: speech; M: manual, K: keyboard).

<table>
<thead>
<tr>
<th>Study</th>
<th>Method</th>
<th>Lane keeping</th>
<th>Brake reaction time</th>
<th>Peripheral detection time</th>
<th>Following distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forlines et al. [18]</td>
<td>Simulator</td>
<td>$S &lt; M$</td>
<td>No diff.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garay-Vega et al. [19]</td>
<td>Simulator</td>
<td>No diff.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gärtner et al. [20]</td>
<td>On road</td>
<td>$S &lt; M$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itoh et al. [21]</td>
<td>Simulator</td>
<td>$S &lt; M$</td>
<td>No diff.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maciej and Vollrath [22]</td>
<td>Simulator</td>
<td>$S &lt; M$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McCallum et al. [23]</td>
<td>Simulator</td>
<td>No diff.</td>
<td>No diff.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minker et al. [9]</td>
<td>On road</td>
<td>$S &lt; M$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranney et al. [24]</td>
<td>On road</td>
<td>$S &lt; M$</td>
<td></td>
<td></td>
<td>$S &lt; M$ (0.8 versus 0.87 s)</td>
</tr>
<tr>
<td>Shutko et al. [25]</td>
<td>Simulator</td>
<td>$S &lt; M$</td>
<td></td>
<td></td>
<td>$S &lt; M$ (Except incoming call)</td>
</tr>
<tr>
<td>Tsimhon et al. [26]</td>
<td>Simulator</td>
<td>$S &lt; K$</td>
<td></td>
<td></td>
<td>$S &lt; K$ (88 versus 167 m)</td>
</tr>
<tr>
<td>Villing et al. [27]</td>
<td>On road</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Task performance statistics from selected studies (S: speech; M: manual).

<table>
<thead>
<tr>
<th>Study</th>
<th>Task completion time</th>
<th>Speech recognizer rate</th>
<th>Task completion rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carter and Graham [17]</td>
<td>$S &gt; M$</td>
<td>92.7%</td>
<td></td>
</tr>
<tr>
<td>Forlines et al. [18]</td>
<td>$S &lt; M$ (18.0 versus 25.2 s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$S &gt; M$</td>
<td></td>
<td>79.4% (recognition error rate: 20.6%)</td>
</tr>
<tr>
<td>Gärtner et al. [20]</td>
<td>Simple: 24.6 versus 12.8 s</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Complex: 74.4 versus 58.7 s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minker et al. [9]</td>
<td>$S &lt; M$ (63 versus 84 sec)</td>
<td></td>
<td>$S &lt; M$ (79 versus 90%)</td>
</tr>
<tr>
<td>Ranney et al. [24]</td>
<td>No difference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shutko et al. [25]</td>
<td>$S &lt; M$ (except dialing phone)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Villing et al. [27]</td>
<td>$S &gt; M$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The paper also contains considerable details 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. [29] 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. [30] present related research on in-vehicle corpora, which may be a useful resource to address on who, when, and how often the driver used the speech interfaces.

4. What Are the Key Research Results of the User Performance Using Speech Interfaces Compared with the User Performance Using Visual-Manual Interfaces?

There have been a number of studies on this topic. Readers interested in the research should read Barón and Green [31] and then read more recent studies.

Using Barón and Green [31] as a starting point, studies of the effects of speech interfaces on driving are summarized in four tables. Table 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 [16], 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 on the task complexity [11, 12].

Table 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., lower standard deviation of lane position).

Table 5 shows that task completion times for speech interfaces were sometimes shorter than that for visual-manual interfaces and sometimes longer, 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 was an important factor that...
### Table 6: Subjective rating and driving behavior statistics from selected studies (S: speech; M: manual).

<table>
<thead>
<tr>
<th>Study</th>
<th>Subjective rating</th>
<th>Driving behavior—glances</th>
<th>Glance duration</th>
<th>Number of glances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carter and Graham [17]</td>
<td>S &lt; M</td>
<td>Radio or CD control:</td>
<td>M: 1 s; S: 1/3 s</td>
<td></td>
</tr>
<tr>
<td>Faerber and Meier-Arendt [32]</td>
<td>S is preferred</td>
<td>M: 3; S: 1/1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gärtnert al. [20]</td>
<td>No difference</td>
<td></td>
<td></td>
<td>Complex tasks: S &lt; M</td>
</tr>
<tr>
<td>Itoh et al. [21]</td>
<td>S &lt; M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maciej and Vollrath [22]</td>
<td>S &lt; M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mccallum et al. [23]</td>
<td>S &lt; M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minker et al. [9]</td>
<td>S is preferred</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shutko et al. [25]</td>
<td>S &lt; M</td>
<td></td>
<td>(Except receiving an incoming call)</td>
<td></td>
</tr>
</tbody>
</table>

affected the task performance. Kun et al. [33] reported that low recognition accuracy (44%) can lead to greater steering angle variance. Gellatly and Dingus [34] reported that driving performance (peak lateral acceleration and peak longitudinal acceleration) was not statistically affected until the 60% recognition accuracy level was reached. Gellatly and Dingus [34] 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 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.

### 5. How Should Speech Interfaces Be Designed?

**What Are the Key Design Standards and References, Design Principles, and Results from Research?**

5.1. Relevant Design and Evaluation Standards. For speech interfaces, the classic design guidelines are those of Schumacher et al. [35], and the one set that is not very well known, but extremely useful, is the Intuity guidelines [36]. Najjar et al. [37] described user-interface design guidelines for speech recognition applications. Hua and Ng [38] 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), ISO 19358: 2002 (Ergonomics—Construction and application of tests for speech technology), ISO/IEC 2382-29: 1999 (Artificial intelligence—Speech recognition and synthesis), and ISO 8253-3: 2012 (Acoustics—Audiometric tests methods—Part 3: Speech Audiometry), focus on the evaluation of the whole system and its components [39–42]. 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) [43].

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 [44] is relevant. In fact, ISO 15006 concerns nonspoken warnings.

There are standards in development. SAE J2988, Voice User Interface Principles and Guidelines [45], 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 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 are 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/.

5.2. Key Books. There are a number of books on speech interface design, with the primary references being Hopper’s classic [46], Balentine and Morgan [47], Cohen et al. [48], and Harris [49]. A more recent reference is Lewis [50].

5.3. 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.

5.3.1. 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 [51]. 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 do support barge-in, which requires the users to press the voice-activation button. However, less than 10% of subjects (unpublished data from the authors) knew and used this function.

5.3.2. Utterance Types (Speech Acts). Speech acts refer to the kinds of utterances made and their effect [53]. According to Akmajian et al. [54], there are four categories of speech acts.

(i) Utterance acts include uttering sounds, syllables, words, phrases, and sentences from a language including filler words ("ummi").

(ii) Illocutionary acts include asking, promising, answering, and reporting. Most of what is said in a typical conversation is this type of act.

(iii) Perlocutionary acts are utterances that produce an effect on the listener, such as inspiration and persuasion.

(iv) Propositional acts are acts in which the speaker refers to or predicts something.

Searle [55] classifies speech acts into five categories.

(i) Assertives commit the speaker to address something (suggesting, swearing, and concluding).

(ii) Directives get the listener to do something (asking, ordering, inviting).

(iii) Commissives commit the speaker to some future course of action (promising, planning).

(iv) Expressives express the psychological state of the speaker (thanking, apologizing, welcoming).

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

5.3.3. Intent and Common Understanding (Conversational Implicatures and Grounding). Sometimes speakers can communicate more than what is uttered. Grice [56] 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 7).

5.3.4. Which Kinds of Errors Can Occur? Skanztze [52] provides one of the best-known schemes for classifying errors (Table 8). Notice that Skanztze does so from the perspective of a device presenting an utterance and then processing a response from a user.

Véronis [57] 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 [58]. 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 the knowledge of rules (“the interface does not hear users’ input correctly”).

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 avoiding errors, and, when they occur, facilitating 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 [59–62].

Knowing how often these strategies occur suggests what needs to be supported by the interface. The SENECA project [9, 20] revealed that the most frequent errors for navigation tasks were spelling problems of various types, entering or
Grice’s conversational maxims (with examples added by the authors).

<table>
<thead>
<tr>
<th>Maxim</th>
<th>Example</th>
<th>Guidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxim of Quantity: be informative.</td>
<td>Machine: Please say the street name. User: 2901 Baxter Road (2901 is the house number)</td>
<td>(i) Make your contribution of information as is required, that is, for the current purpose of the conversation. (ii) Do not make your contribution more informative than is required.</td>
</tr>
<tr>
<td>Maxim of Quality: make your contribution one that is true:</td>
<td>U: “Toledo Zoo, Michigan” (but Toledo is in Ohio)</td>
<td>(i) Do not say what you believe to be false. (ii) Do not say that for which you lack evidence.</td>
</tr>
<tr>
<td>Maxim of Relevance: be relevant.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Maxim of Manner: be perspicuous</td>
<td>(i) M: Please say set as destination, dial or back. U: Dial. O no Don't dial, Back (User want to say “back”)</td>
<td>(i) Avoid obscurity of expression (ii) Avoid unnecessary ambiguity (iii) Be brief (avoid unnecessary prolixity) (iv) Be orderly.</td>
</tr>
<tr>
<td></td>
<td>(ii) M: Please say the POI category. U: Let's see. Recreation</td>
<td></td>
</tr>
</tbody>
</table>

Examples of errors in different modules of speech-controlled interfaces (adapted from Skanztze [52]).

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

Enhanced version of Véronis [57] error-classification scheme.

<table>
<thead>
<tr>
<th>System</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Competence</td>
</tr>
<tr>
<td>Lexical level (word)</td>
<td>Letter substitution Letter insertion Letter deletion</td>
</tr>
<tr>
<td>Syntactic level (sentence structure)</td>
<td>Missing rule</td>
</tr>
<tr>
<td>Semantic level (meaning)</td>
<td>Incomplete or contradictory knowledge representation Unexpected situation</td>
</tr>
</tbody>
</table>

Lo et al. [63] 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 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, that is, missing or inappropriate PTA activation.
Table 10: Variables used for evaluating entire systems or system modules.

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

and song title, album name, and/or artist name. Relationship errors were common because subjects were not familiar with the given songs/albums/artists.

6. How Should Speech Interfaces Be Assessed and What Should Be Measured?

6.1. What Methods Should Be Used? 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 [4, 5, 11, 16, 64–66], sometimes implemented using Suede [67]. 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 the interface to be tested and redesigned.

6.2. What Should Be Measured? Dybkjaer has written several papers on speech interface evaluation, the most thorough of which is Dybkjaer et al. [68]. That paper identified a number of variables that could be measured (Table 10), in part because there are many attributes to consider.

Walker et al. [69] proposed a framework of usability evaluation of spoken dialogue systems, known as PARADISE (PARAdigm for Dialogue System Evaluation). (See [70] for criticisms.) Equations were developed to predict dialog efficiency (which depends on mean elapsed time and the mean number of user moves) and 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):

\[ \kappa = \frac{(P(A) - P(E))}{1 - P(E)}, \]  

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 [71, 72]. 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 [73]. (See [74] for more information.) Using the definitions in SAE Recommended Practice [2944] [75] is recommended.

7. Summary

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

7.1. What Are Some Examples of Automotive Speech Interfaces? Common automotive examples include CHAT, CU Move, Ford Model U, Linguatronic, SENeca, SYNc, VOIC, and Volkswagen. Many of these examples began as collaborative projects that eventually became products. SYNc is the best known.

Also important are nonautomotive-specific interfaces that will see in-vehicle use, in particular, Apple Siri for the iPhone and Google Voice Actions for Android phones.

7.2. Who Uses Speech Interfaces, for What, and How Often? Unfortunately, published data on who uses speech interfaces and how real drivers in real vehicles use them is almost zero. There are several studies that examine how these systems are used in driving simulators, but those data do not address this question.

7.3. What Are the Key Research Results of the User Performance Using Speech Interfaces Compared with the User Performance Using Visual Manual Interfaces? To understand the underlying research, Barón and Green's study [31] is a recommended summary. Due to the difference of task complexity while testing, comparing alternative speech systems is not so easy.
However, when compared with visual-manual interfaces, speech interfaces led to consistently better lane keeping, shorter peripheral detection time, lower workload ratings, and shorter glance durations away from the road. Task completion time was sometimes greater and sometimes less, depending upon the study.

7.4. How Should Speech Interfaces Be Designed? What Are the Key Design Standards and References, Design Principles, and Results from Research? There are a large number of relevant technical standards to help guide speech interfaces. In terms of standards, various ISO standards (e.g., ISO 9921, ISO 19358, ISO 8253) focus on the assessment of the speech interaction, not on design. Speech-quality assessment is considered by ITU-P800. For design, key guidelines include [35–38]. A number of books also provide useful design guidance including [46–50].

Finally, the authors would recommend that any individual seriously engaged in speech-interface design should understand the linguistic terms and principles (turns, speech acts, grounding, etc.) as the literature provides several useful frameworks for classifying errors and information that provides clues as to how to reduce errors associated with using a speech interface.

7.5. How Should Speech Interfaces Be Assessed and What Should Be Measured? The Wizard of Oz method is commonly used in the early stages of interface development. In that method, an unseen experimenter behind the scenes simulates the behavior of a speech interface by recognizing what the user says or is speaking in response to what the user says, or both. Wizard of Oz simulations take much less time to implement than other methods.

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 and gap variability, are measured as eyes-off-the-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 its 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 [2944] [75] 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 paper provides some of the background information needed to create those models.

References

12 International Journal of Vehicular Technology


[73] P. Green, “Using standards to improve the replicability and applicability of driver interface research,” in Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular (AutomotiveUI ’12), Portsmouth, UK.