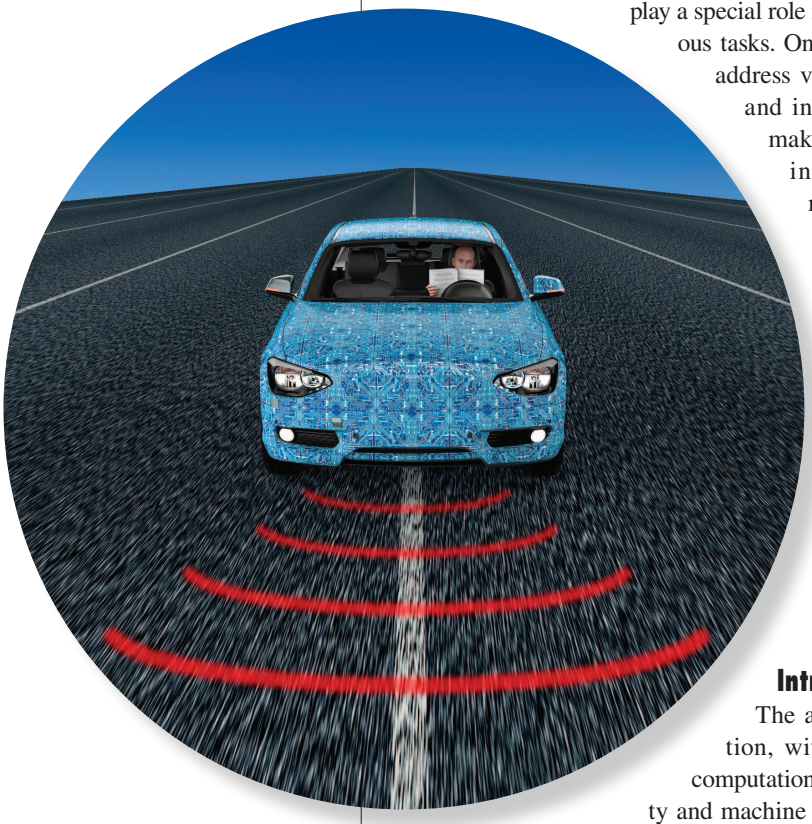
Fuliang Weng, Pongtep Angkititrakul,
Elizabeth E. Shriberg, Larry Heck,
Stanley Peters, and John H.L. Hansen

Conversational In-Vehicle Dialog Systems

The past, present, and future



SKY IMAGE LICENSED BY GRAPHIC STOCK
©ISTOCKPHOTO.COM/POSTERIORI

Automotive technology rapidly advances with increasing connectivity and automation. These advancements aim to assist safe driving and improve user travel experience. Before the realization of a full automation, in-vehicle dialog systems may reduce the driver distraction from many services available through connectivity. Even when a full automation is realized, in-vehicle dialog systems still play a special role in assisting vehicle occupants to perform various tasks. On the other hand, in-vehicle use cases need to address very different user conditions, environments, and industry requirements than other uses. This makes the development of effective and efficient in-vehicle dialog systems challenging; it requires multidisciplinary expertise in automatic speech recognition, spoken language understanding, dialog management (DM), natural language generation, and application management, as well as field system and safety testing. In this article, we review research and development (R&D) activities for in-vehicle dialog systems from both academic and industrial perspectives, examine findings, discuss key challenges, and share our visions for voice-enabled interaction and intelligent assistance for smart vehicles over the next decade.

Introduction

The automotive industry is undergoing a revolution, with progress in technologies ranging from computational power and sensors to Internet connectivity and machine intelligence. Expectations for smart vehicles are also changing, due to an increasing need for frequent engagement with work and family, and also due to technological progress of products in other areas, such as mobile and wearable devices.

This article will address human–machine interaction (HMI) and advanced intelligent assistance during the use of smart vehicles with a focus on voice-based technology. Speech is a primary means for human-to-human communication, capable of conveying rich content, colorful emotion, and human intelligence. Therefore, it is the most suitable

Digital Object Identifier 10.1109/MSP.2016.2599201
Date of publication: 4 November 2016

driver–vehicle modality, allowing drivers to keep their eyes on the road and their hands on the wheel. A strong user need for a true voice-enabled intelligent assistance will soon make it possible to perform many daily tasks while driving, which previously could not be achieved safely. During the transition from highly manual vehicle operation to highly automated operation, moving from simple dialog systems to sophisticated driving assistance is a major trend, and both challenges and opportunities arise.

Review of past major activities

The results from well-designed driver data collections reveal useful insights into HMI. Innovations from R&D projects as well as related assistance systems provide a good foundation for future developments. An overview for past major activities is given next.

In-vehicle dialog data collection

In the past two decades, vehicles equipped with both human and vehicle sensors have been deployed to collect realistic data on drivers, vehicles, and driving environments for the development of human-centered technologies. Among the in-vehicle data, speech corpora were collected for developing in-vehicle automatic speech recognition (ASR) and spoken dialog systems. Speech and audio data collections in cars are necessary to capture specific driver speech patterns and in-vehicle noise characteristics. For ASR, collected data can also be used to expand natural language coverage in recognition grammars. Video and data from a variety of sensors (e.g., haptic input, brake and gas pedal pressure, steering wheel angle, vehicle position, etc.) were synchronously collected to capture multimodal interactions between the driver and vehicle, as well as driver's states and cognitive loads. Started in 1998, the SPEECHDAT-CAR effort was the first international program to form a multilanguage speech corpus for automotive applications with 23 research groups from nine European countries and one U.S. site. A number of additional large-scale automotive speech data collection were completed [1]–[9]. Some of them are fully or partially open to the public for research purposes; others are privately owned by companies for product development. To study naturalistic conversational speech during driving, most data collection experiments deployed a Wizard of Oz (WoZ) technique [10], typically involving a human navigator (wizard) sitting in a separate room to simulate a dialog system. Another common methodology is to collect actual driving data with event data recorders. In general, these systems only record specific vehicle information over a short period of time during a crash and usually do not include audio and video data. However, these data are very useful for investigating accident causes and designing interfaces.

Speech is a primary means for human-to-human communication, capable of conveying rich content, colorful emotion, and human intelligence.

A common finding among most data collection efforts validated very significant speaker variability across different driving conditions and maneuver operations. CU-Move focused on in-car noise conditions across different cabin structures. It found that having the windows open has more effect on the recognition accuracy than increasing the vehicle speed. AVICAR showed that combining both audio and video information significantly improves the speech recognition over the audio-only information under noisy conditions but is less beneficial in quiet conditions. SPEECHDAT-CAR reported that about 50% of collected speech data in vehicle is from speaker noise such as breathing, and coughing, while mispronunciation and incomprehensible speech could contribute up to 5% of data. From the corpus of the Center for Integrated Acoustic Information Research, it found that the number of fillers, hesitations, and slips per utterance unit was 0.31, 0.06, and 0.03, respectively. The Conversational Helper for Automated Tasks (CHAT) data collection shows that nearly 30% of proper names were partial names [9], and disfluent and distracted speech was prevalent [10].

Key findings from past R&D projects for in-vehicle use cases

Over the past decade, a number of publicly funded projects specifically address in-vehicle use cases, e.g., Virtual Intelligent Codriver (VICO), Tools for Ambient Linguistic Knowledge (TALK), and CHAT [9]. In the automotive industry, it is well known that driver distraction is a major source of fatal traffic accidents; thus, minimizing driver distraction in automotive HMI has been a key goal. Driving-related activities have been classified as 1) the primary tasks of vehicle maneuvering, which require one's eyes on the road for navigation and obstacle avoidance, one's hands on the wheel for steering, and one's feet on the pedals for moving or stopping the vehicle and 2) secondary tasks, such as maintaining occupant comfort and accessing infotainment. While these primary tasks have been stable, increasingly, the secondary tasks are becoming richer and more diverse, especially as Internet connectivity and availability of information and services have become common. Manual controls for such sophisticated secondary tasks such as buttons and knobs are difficult for drivers to safely operate and are inadequate for complex tasks. As speech technology has matured, it has become an increasingly natural choice due to its convenience and robust capabilities. The CHAT-Auto addresses conversational dialog systems-related challenges, such as voice-operated multitasking with imperfect speech input and communicating a large amount of information content to drivers with limited cognitive capacity. It further demonstrated the feasibility of speech technology for representative vehicle use cases such as navigation by destination entry, point-of-interest (POI) finding, and music search. The European Union (EU) VICO project initiated voice-enabled assistance system prototyping for navigation-related tasks such as POIs and address input. The EU TALK project covered on adaptive multimodal and multilingual dialog systems with a main focus on reusability by allowing for the core dialog system to be separated from specific

applications, languages, or modalities. It also intends to show the capability of learning dialog strategies to improve communication with drivers when an ASR engine make errors. A statistical framework of a partially observable Markov decision process was used to address uncertainty in speech recognition results [11]. Despite different foci of these research projects, they share many important building blocks in an advanced speech dialog system, as shown in Figure 1.

Influential automotive infotainment HMI products on the market include BMW iDrive, Mercedes COMAND, Audi MMI, Ford Sync, GM CUE, Lexus Enform, and Acura AcuraLink. In general, these HMI products are used to control multiple in-car functionalities including navigation, entertainment, multimedia, telephony, vehicle dynamics, and so on. Existing HMI input technologies fall into two categories: haptic based and voice based. Of haptic-based input methods, the two most common are control knobs and touch screens. German cars, e.g., BMW iDrive and Audi MMI, typically use control knobs. While these control knobs are simple to operate, it often takes multiple steps to access a function deep within a menu. Thus, these interfaces impose cognitive demands on drivers, requiring them to memorize where menu items are located and navigate through the menu while driving. In contrast, Japanese cars tend to use touch screens. Navigating touch-screen interfaces can be more intuitive and efficient, but they impose a spatial limitation on drivers, as they must reach out to different areas of touch screen to access predesignated buttons in a shallower but still hierarchical menu structure. As both knobs and touch screens require hand-eye coordination, haptic interfaces pose inherent safety risks. Research shows that glancing away from the road for two seconds or longer may increase the risk of an accident from four to 24 times [12]. Such research evidence has led some original equipment manufacturers to prohibit the use of some of interfaces with hierarchical menu structures during driving. The voice-based systems have been shown to reduce driver look-away time and lessen spatial and cognitive demands on the driver. It is expected that recent advancement in speech recognition technology; other sensing technologies such as gesture on touch pad, on steering wheel surface, and in air; as well as deeper modality fusion technologies would greatly liberate the designers so that many new devices and services can be incorporated without many physical constraints in the cockpit [13], [14].

In the past two decades, vehicles equipped with both human and vehicle sensors have been deployed to collect realistic data on drivers, vehicles, and driving environments for the development of human-centered technologies.

Related intelligent assistant technologies

Likewise, many efforts were devoted to the field of general-purpose voice-enabled intelligent personal assistants (IPAs). The U.S. Defense Advanced Research Projects Agency (DARPA) funded a number of key IPA-related programs. The DARPA Communicator project was a major early effort in developing robust multimodal speech-enabled dialog systems with advanced conversational capabilities for engaging human users in mixed-initiative interactions. From 2003 to 2008, DARPA funded the Cognitive Agent that Learns and Organizes (CALO) project. CALO attempted to build an integrated system capable of true artificial intelligence's (AI's) key features such as the ability to learn and adapt in adverse situations, and comfortably interact with humans [15]. The CALO project had

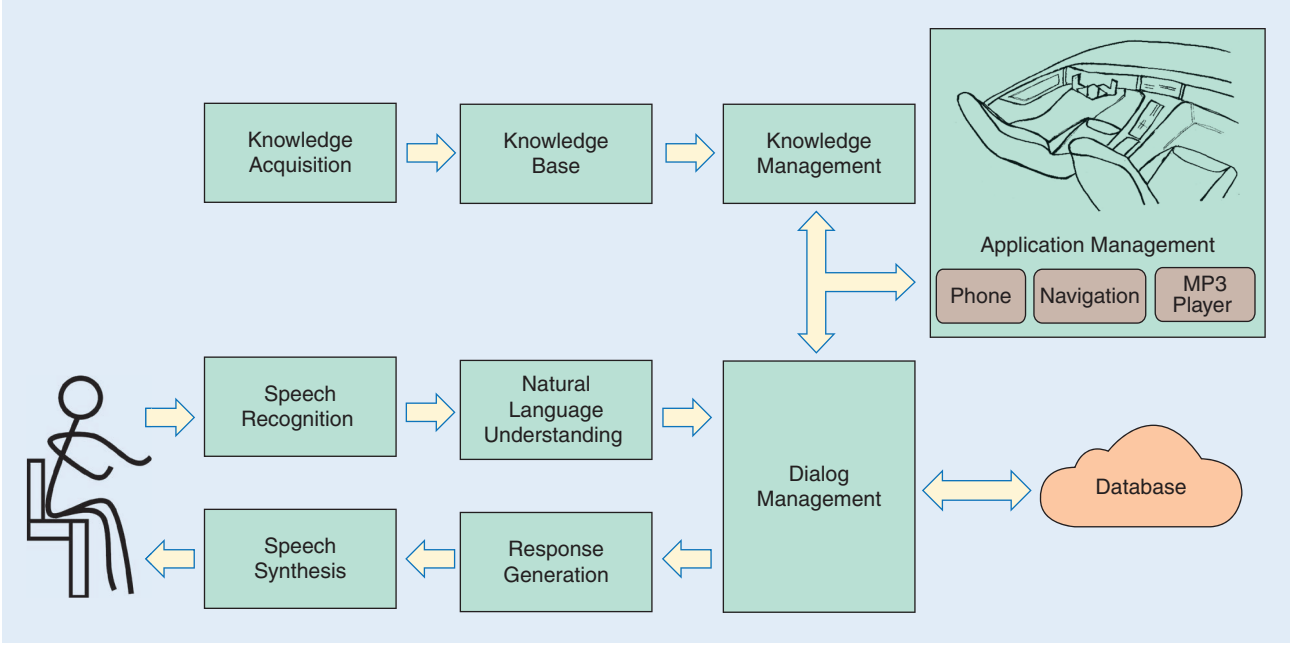


FIGURE 1. A generic block diagram for a typical in-vehicle spoken dialog system.

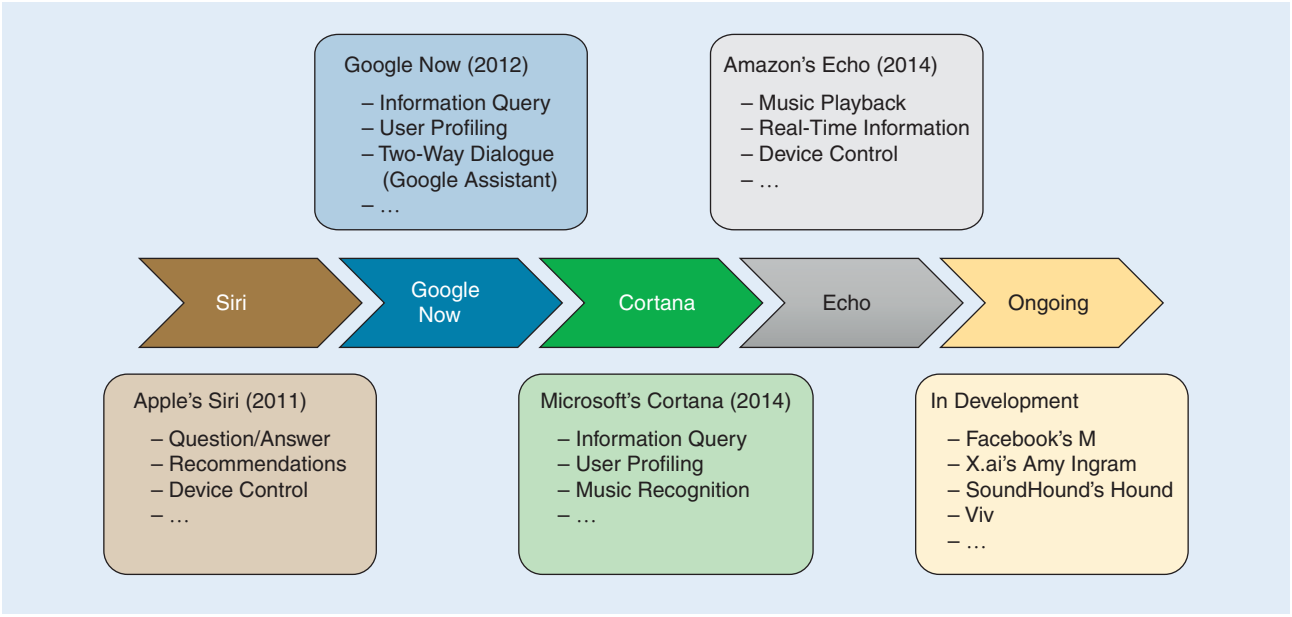


FIGURE 2. The development of voice-enabled IPAs.

a number of spin-offs, most notably SRI International's Siri intelligent software assistant.

Nuance speech technologies are behind many successful products in different markets. Due to limitations in speech understanding accuracy and coverage, early Nuance speech dialog systems in call-center applications strictly followed a visual menu design, rather than typical human interaction patterns, leading to low user satisfaction in many cases.

After the acquisition of Siri in 2011, Apple's launch of the IPA on the iPhone marked a turning point in the mass acceptance of speech technologies. With the massive computational power available through the cloud, more applications and AI technologies started to be integrated into dialog systems; see Figure 2. Consequently, IPAs were developed that allow users to operate devices, access information, and manage personal tasks in a much richer way. Notable IPAs include Apple Siri, Google Assistant, Microsoft Cortana, Amazon Echo, IBM Watson, and Baidu. In addition, the major players in the field are building up developer platforms around these IPAs to enable their own AI ecosystems. However, these systems typically do not have much dialog capability and, in most cases, focus on single-turn question-answers (Q&A) and simple actions. As a contrast, text-based chatbots from Facebook, Google, and others make use of dialog technologies in automating services via Messenger, bypassing the dependency on speech technology.

The rapid increase of high-quality cloud-based IPAs has been partly attributed to recent advances in deep learning technologies—especially deep neural networks.

The rapid increase of high-quality cloud-based IPAs has been partly attributed to recent advances in deep learning technologies—especially deep neural networks (DNNs) [16]. With the exception of speaker recognition in the late 1990s [17], deep learning methods have only recently surpassed hidden Markov with Gaussian mixture model (GMM-HMM)-based approaches in speech recognition performance. With powerful computing resources and advances in learning algorithms beyond early approaches [17], [18], hidden Markov models with DNN (DNN-HMM)-based approaches are able to reduce error rates in speech recognition by 50% [18]. Various studies have been conducted on noise robustness with DNN. Some focus on using environmental information via feature enhancement [19]–[22]. Others include different conditions, such as model noise, reverberation, or speaker variation, in an end-to-end speech recognition training with a recurrent neural network [23]. The latter contrasts with traditional approaches of separating ASR optimization into front-end signal processing and back-end model training. Effective front-end approaches have been developed to address acoustic echo, background noises, and multiple sound sources with microphone array processing [24]–[26]. The improved recognition accuracy significantly benefits the usability and adoption of the general IPAs.

The automotive industry has taken two different approaches to address the arrival of cloud-based, voice-enabled assistance systems from major IT companies. The shallow integration approach leverages in-vehicle microphones, loudspeakers, control buttons, or a head unit screen to enable mobile devices to synchronize the systems' look and feel. This approach is represented by the MirrorLink standard from Car Connectivity Consortium, Apple's CarPlay, and Google's Android Auto. The deep integration approach requires the integration of the embedded and cloud systems at a component level. Recent experiments have shown that combining ASR results from embedded and cloud-based engines can reduce word error rate by up to almost 30% for

Table 1. Major technical challenges for the in-vehicle dialog system.

Dialog System/Components	Features	Challenges
ASR	Noise robustness	Robust ASR under different driving conditions with different types of noise environments, e.g., road, wind, or traffic noises.
	Cognitive state sensitivity	Learning the driver cognitive states from speech for better recognition: what his/her emotional state is, whether he/she is busy with maneuvering the vehicle, drowsy, listening to radio or music, or thinking about something else.
	Addressee detection	Discriminate between system-directed from human-directed (or ambient) speech without using push-to-talk button. Tracking user conversation with the system for multiple dialog turns (also DM).
IU	Spoken language understanding	Make sense from broken human speech with high level of hesitation, revision, or wrong word order or high speech recognition error rate.
	Multimodal understanding	Integration of input from other natural modes (e.g., gesture on surface or in air, eye gaze) and accurate understanding of user intent.
DM		Decision about what system states need to be communicated to the drivers and at what conditions in coordination of the requests from the user in the dynamic changing context. Careful management of driver emotion, and selected content to the driver for the condition.
Knowledge management		Domain knowledge to be selected based on dynamic context and personal preference via user modeling and used to facilitate the IU as well as recommendation and decision making.
Natural language generation		Context-dependent content generation and prioritization. Discourse-aware flexible expression generation under different driving conditions and requested item availability as the template-based approach will no longer be sufficient.
Application manager	Fast multidomain integration	Multiple nomadic devices may be brought into the vehicles. To specify devices and services so to voice-enable them, integrate devices dynamically (plug and play) or perform service composition.
	Multiagent integration	Coordination, integration, and enrichment of the output from multiple assistant agents brought in primarily through various smart mobile devices, such as Siri, Google Now, Cortana, or embedded systems.
System level	Testing	Systematic testing of whole-system behavior with respect to system latency, response appropriateness, and task-completion rate.
	Specification/validation	Develop a detailed specification with many different possible combinations of scenarios, and validate the resulting system according to system specification based on standards, e.g., ISO 26262.

limited in-vehicle domains [27]. These gains will likely carry over to other modules in the future.

While these general IPAs benefit from leveraging huge computing resources in the cloud, they are highly dependent on Internet connectivity. However, many remote locations, where drivers need navigation systems the most, do not have Internet connectivity. This crucial limitation provides a strong justification for the role of an embedded global positioning system and associated embedded speech interface.

In summary, menu-structured multimodal interaction is the current state-of-the-art technology for operating in-vehicle infotainment products. Embedded voice interface solutions are less popular, due to limited recognition accuracy and lack of natural language capability. Cloud-based IPAs provide a good alternative for in-vehicle uses but with only a shallow integration. While many forward-looking building blocks in dialog systems—such as disfluency detection and content management—have been investigated in research projects, the conversion rate into in-vehicle embedded products has been slow. Among the most well known in-vehicle dialog system products include those from Ford Sync (2007–2012). The GM CUE system took a major step in 2012 by incorporating natural speech beyond restricted grammars into its embedded dialog system, resulting in a noticeably improved user experience [28]. In the meantime, cloud-based intelligent assistance systems have popularized in-vehicle speech use, due to their relatively high language-understanding accuracy and rich applications not available in embedded in-vehicle systems. However, the impact of cloud-based speech technologies is mostly limited to the ASR accuracy and text-to-speech (TTS) naturalness for a dialog system development. Despite all the technology and product limitation, it is clear that a general direction for the in-vehicle interaction is conversational and intelligent dialog (CID) systems.

While many forward-looking building blocks in dialog systems have been investigated in research projects, the conversion rate into in-vehicle embedded products has been slow.

Challenges

In many ways, the technical challenges for in-vehicle intelligent dialog systems bear similarity to those intended for general-purpose dialog systems, especially in the context of hands-busy and

eyes-busy scenarios. Because of the special nature of in-vehicle use, however, additional challenges are apparent from in-vehicle data collection efforts and in-vehicle dialog system development activities mentioned in the section "Review of Past Major Activities." Some important challenges are summarized in Table 1.

One may look at these challenges as arising from the three interdependent factors: 1) the driver, 2) the environment, and 3) the automotive industry.

- *The driver.* Drivers typically have short attention spans during interaction with in-vehicle dialog systems, as driving is their primary task. From 2020 to 2030, drivers will begin to be exposed to autonomous driving technologies, freeing them from constant attention to vehicle control. However, according to projections by McKinsey [39], the adoption of self-driving vehicles will likely be less than 15% by 2030. HMI while driving will continue to be important challenge. As a result, CID systems have to handle challenges on many different levels [9]. Drivers' speech may be fragmented, disfluent, and repetitive. Drivers may need to hear shorter or simpler responses; they may expect the system to provide straightforward recommendations instead of a potentially overwhelming number of choices. When systems make recommendations, the systems should anticipate driver's contextual needs to the point that the driver has the feeling that she or he doesn't need to state the obvious, so driver behaviors and personal preferences need to be taken into consideration to avoid excessive conversational turns. Additionally, in one vehicle, there may be multiple passengers, each requiring a separate preference profile. A sophisticated dialog system should be able to coordinate requests from different users, know when to not interrupt human–human conversations, and come up with an optimized response for the group.
- *The environment.* The in-car environment is generally more dynamic and has higher stakes than the contexts in which other dialog systems are deployed. Inside the vehicle, information from sensors reflects vehicle status changes, some of which need immediate attention (an engine breaking down), while others can be handled at a later time (an oil change warning). Differences in design, interior materials, and mechanics create different acoustic environments and background noises inside vehicles. Outside the vehicle, the physical environment is also diverse and dynamic. A vehicle may be on a highway or a city road, accelerating or decelerating, on gravel or asphalt—creating significantly different background noises. Traffic conditions can vary significantly: the driver may be stuck in stop-and-go traffic, or may be traveling unimpeded at high speed. Harsh weather, such as wind, rain, hail, or thunder, typically requires additional attention from the driver and alters the in-vehicle acoustics. In such cases, the dialog system may need to use different dialog strategies with respect to taking the initiative to engage the driver. Available services (e.g., gas station, or parking) via TTS from users' mobile devices add another dimension of complexity and are competing for drivers' attention. Managing multiple assistance systems from different devices will become an important development consideration.
- *The automotive industry.* Two key issues are critical to automotive companies' decisions about whether or not to adopt new technologies. One is safety. Improper implementation of certain technologies could lead to fatal accidents. The other is the reliability. When a component is installed in a vehicle, it must work properly for a long time. Vehicle parts and features installed are typically required to function properly for at least ten years. For the safety and reliability reasons, this industry is highly regulated by the government. Relevant regulations and guidelines include ISO 26262 [29], ISO 15504 (Automotive SPICE), and IEC 61508 [30]. Due to such requirements, common practice in the consumer electronics or Internet world of fast product development cycle may not be directly applied in this industry. In the automotive industry, a rigorous process covering design specification, development, testing, and validation is followed to ensure the resulting product quality meets requirements. For example, to develop a CID system, one needs to specify the system coverage and performance by listing many phrasal and interactive variations under different noise conditions by speakers from different dialect regions, leading to a huge number of combined testing cases. A complete testing of these cases against various requirements becomes more and more challenging, especially with increasing coverage needs from the users.

The automotive industry has taken two different approaches to address the arrival of cloud-based, voice-enabled assistance systems from major IT companies.

The combination of these three factors causes many technical challenges. We next highlight some critical ones that have long-term impact in the areas of speech recognition and understanding, multiple speaker conversation coordination, the effect of driver behaviors and states for safety, as well as the integration of general intelligent assistance systems.

Challenges in speech recognition and understanding

From a historical perspective, speech recognition in the car started with small vocabulary systems primarily for command and control, along with optimization of either microphone placement or multimicrophone array processing to suppress the diverse noise sources present for in-vehicle scenarios [31]. More recent efforts have focused on expanding speech recognition coverage to additional in-vehicle domains.

The in-vehicle acoustic environment is complex and dynamic. Factors in this acoustic environment include noise from air conditioning units, wiper blades, the engine, external traffic, the road surface, wind, open windows, and inclement weather. The level of background noise while windows were

open 1–2 inches traveling at 65 miles/hour can be very close to speech level in the frequency band 0–1 kHz. Noise is not only intrusive and reduces the drivers' concentration but also degrades the ASR performance and interrupts the dialog flow. While some noise conditions are quite similar across vehicle types, some in-vehicle noise conditions vary significantly across vehicles such as noise produced by engines, turn signals, or wiper blades. Within the same vehicle, noise levels can vary from very low when the engine is idle and windows are closed, to very high when traveling at high speeds with open windows. Similar to other dialog systems, environmental noise also impacts how speakers speak, affecting a wide range of speech characteristics (including task stress, emotion, and Lombard effects).

Another challenge for in-vehicle speech recognition and understanding comes from imperfect speech input.

Another challenge for in-vehicle speech recognition and understanding comes from imperfect speech input. When drivers are under stress, their speech can be less fluent and predictable. Their speech tends to contain more word fragments, restarts and repairs, hesitations, and alternative phrasings. These deviations from standard speech result in degraded speech recognition performance as a result of variation in acoustics and language.

For acoustic modeling, the classical cross-word modeling approach becomes less effective due to word fragmentation and hesitation. Similarly, speech endpoint detection becomes very difficult as the engine does not know whether a silence is a long pause or the end of a request, especially in the presence of noise. For language modeling, word fragments pose a challenge since there are often many proper names to cover and names can be interrupted leaving only a fragment. The more word fragments are included in the system, the more confusability is added, and it is harder to build a low-perplexity language model to constrain search space.

Initial attempts have been made to improve the detection of disfluent speech [32]. As DNN-HMM and Connectionist Temporal Classification (CTC) are overtaking GMM-HMM [18], [33], DNN-based technologies are predicted to better handle disfluent speech although their potential benefits need to be validated through real-world in-vehicle uses.

Challenges in coordinating conversations

Beyond the challenges of word recognition itself, interacting with increasingly capable voice technologies in the car will require more sophisticated coordination of human–machine conversations. Drivers freed from lower-level driving tasks will have more opportunity for social interactions both inside and (via mobile) outside the car. And the systems in the car that they do communicate with will have advanced knowledge and complexity. As a result, conversation management presents challenges and new opportunities even if one assumes that high-quality word recognition is available.

One basic challenge is determining the addressee of an utterance, since a driver (or passenger) may be speaking to the system, or to another person, or to a person outside the car (e.g., on a call), or even to an automatic assistant on, e.g., his or her mobile phone. A system needs to know when it is being addressed, and respond only then and not in other contexts. In addition, the interruption of conversation will generally need to put another "on hold"—something that people are used to doing with other people, but generally not with systems. Addressee detection will need to scale to be able to suspend as well as resume conversations. It is worth noting that the current practice of "hot" words (wake-up words that are used to engage with a system) serves to initiate an addressee, but not to effectively suspend, resume, or close these interactions.

Even given the correct addressee, another challenge is how to handle interruptions to fluency of incoming speech for the timely determination of system responses. As just noted, speech contains pausing and disfluencies [34], and the task of driving only enhances opportunities for distraction and coping with sudden changes in the car or surrounding environment. Such fluency breaks cause important ambiguities for turn-taking. For example, even a simple pause to a navigation system, as in "which road do I turn onto (long pause) after I cross the bridge" produces different results pre- (incorrect) versus post- (correct) pause that matter to timely interaction. Waiting induces system latency if the speaker was actually done. Work using acoustic-prosodic features of prepause speech [35], as well as incremental content [36], can be used to better determine whether a user is suspending versus finishing an utterance. Additional challenges exist for handling self-repairs [37] in the driver's speech in real time.

Future challenges in conversational management also include the system production of "conversational grounding," which becomes more necessary as utterance length and complexity increases. In natural conversation, partners employ speech back channels such as "uh-huh," as well as visual cues such as gaze and head nods, to convey to each other that they are "still listening." Mobile interfaces currently display visual information to ground users, but as autonomy increases, audio rather than visual grounding offers the benefit of an eyes-free option. Research on system-produced back channeling [38] offers promise for the future of natural interactive systems, but scaling to grounding in the safety and multiconversation environment requires a better understanding of how users interact with system-produced grounding mechanisms in real time and under cognitive load.

For all of these conversational management tasks, the in-vehicle environment offers unusually rich opportunities for speaker-dependent modeling. With fewer lower-level driving tasks to attend to, drivers are expected in particular to vary dramatically in use of voice for reasons unrelated

to driving, including phone calls. Systems can learn driver behaviors over time to achieve better performance and safety outcomes.

Challenges in incorporating driver behavior and driver states within a vehicle assistance system for enhancing system and safe driving

A further set of challenges for in-vehicle dialog systems concerns driver behavior and cognitive state in the context of additionally available input from in-vehicle sensors, Internet content, and Internet services, and their impact on dialog system design and development.

Driving is a highly dynamic process in which the level of environmental demands can change rapidly, while human cognitive resources are limited. Therefore, the most important goal in designing any in-vehicle driver–vehicle interaction system is to optimize information processing of drivers while operating such systems. Driver behavior is a crucial factor in traffic safety and interaction with in-vehicle intelligent systems. Intelligent dialog systems that minimize driver distraction require a thorough analysis and a good understanding of driver behavior. In addition, dialog systems should operate together with vehicle Advanced Driver Assistant Systems such as Driver Inattention Monitoring Systems and Driver Alert Control Systems [36]. In future autonomous vehicles, drivers may engage in many tasks other than driving. Accurate identification of driver state and behavior plays an important role in deciding hand over of vehicle control.

Driver behavior is a crucial factor in traffic safety and interaction with in-vehicle intelligent systems.

Two basic challenges are how to accurately observe driver behavior using different kinds of sensors and then, based on the observations, how to objectively identify driver behavior and cognitive states, which are highly subjective. Previous studies have used nonintrusive techniques combining eye movements, gaze variables, head orientation, heart rate, CAN-bus signal, vehicle position, and road geometry to capture driver-behavior signals. These approaches could be enhanced by integrating personal characteristics such as gender, age, and medical conditions. Earlier attempts monitored the behavior of driver and vehicle separately, whereas recent attempts focused on monitoring driver, vehicle, and driving environment simultaneously to effectively associate driver's behavior and states corresponding to contextual information. To infer the driver behavior from sensor data, recent common approaches perform some kind of probabilistic techniques to capture both static and dynamic behavioral characteristics. The challenge is how to define to normal versus abnormal driving status. Nevertheless, drivers possess an ability to manage their workload capacities and adjust their behavior to the environment under hazardous situations—but not in every situation. In addition to safety, monitoring driver behavior is important for a dialog system to dynamically adapt itself to the driver state (e.g., emotion) to avoid negative experience of users from system errors, as well as nurture the positive experience.

Challenges to integrate general intelligent assistance technologies

As the demand for general cloud-based IPAs continues to increase, drivers will bring them into vehicles while driving. The general cloud-based IPAs have been designed primarily for uses other than in-vehicle; for example, responses are mostly displayed on the screen. Significant adaptation will be necessary to accommodate in-vehicle uses. Another challenge for the in-vehicle use of general IPAs would be to address different types of noises effectively to achieve high understanding accuracy with drivers and passengers. Yet another one will be about how well the general IPAs will be integrated into embedded spoken dialog systems with overlapping functionalities and different personalities. For such cases, one needs to decide which IPA will take the task and provide a solution. If multiple IPAs are used, the integration of multiple solutions for a task is rather challenging and has to be resolved to offer consistent user experience. This would be especially important if different IPAs keep different user profiles with different learned preferences.

Taking into the consideration of all these factors and their possible combinations in a dialog system realization, we summarize some major technical challenges in Table 1.

Future trends and CID system outlook

Looking toward the future, voice-enabled in-vehicle assistance technologies will be influenced by two major trends on the horizon: increased automation in driving, with independent sensing and artificial intelligence capabilities; and increased vehicle connectivity to online IPAs, with driving-related services enhanced by traffic infrastructure and sensor advancement.

CID systems in the context of autonomous driving

Because the average American drives alone nine times as often as he or she drives or rides in a car with someone else, the primary emphasis regarding in-car speech systems has been and today still is on such systems' interactions with drivers and mitigating driver distraction. As technological advances push automotive design toward increasing autonomy in coming years, perhaps eventually culminating in fully self-driving cars, drivers will become increasingly like passengers. The nature as well as the risks of distraction vary along this spectrum of automotive autonomy. Even as driving becomes increasingly automated, the modality of speech retains essential advantages over other modalities for several reasons: listening does not detract from the visual attention needed for driving, language has richest expressive capability, speech technology is very flexible

with respect to form factor, and the speech channel is little utilized in single-occupant vehicles.

At all points on the vehicle autonomy spectrum, drivers and passengers can avail themselves of Internet connectivity to rich content and services only recently accessible in-vehicle. Time in the car can be used to entertain oneself or complete useful tasks, to the extent that these activities do not interfere with driving responsibilities. Audio channel is preferable to visual for these activities because reading in cars causes motion sickness for many people and because vehicle vibration caused by road conditions interferes with reading but listening is much less affected. Therefore, CID systems will remain a desirable interface channel for content and service consumption.

When drivers have greater responsibility for controlling their vehicles, CID systems hold the potential to reduce distraction and increase safety by, for example, communicating information about vehicle health and road conditions in natural language instead of with cryptic warning lights or not at all. When autonomous cars take over greater responsibility for controlling vehicles, the drivers' trust or sense of a safe ride need to be built up over time. This trust-building process for the adoption of autonomous vehicles can be facilitated by the CID systems through communicating information about vehicle controlling capability together with vehicle health and road conditions.

At all points on the vehicle autonomy spectrum, drivers and passengers can avail themselves of Internet connectivity to rich content and services only recently accessible in-vehicle.

Further along the autonomy spectrum, where smarter cars and intelligent traffic systems provide ever greater driver assistance, the temptation for drivers to pay less attention to driving and more to unrelated content and services will naturally grow as the primary driving tasks diminish. The potential for boredom increases drivers' risk of becoming inattentive and occupying themselves with nondriving activities. Smarter cars equipped with smarter speech systems can reduce this risk of distraction by keeping drivers engaged to an appropriate extent, and preparing them to take over greater vehicle control as necessary. For instance, such cars could explain aloud their automated responses to significant changes in driving circumstances. When approaching a complex intersection, encountering difficult vehicle or pedestrian traffic, or upon detecting an unexpected road closure, cars can announce pertinent information preparing the driver to handle the situations.

When fully autonomous driving becomes available, transitions from autonomous-driving mode to human-driver mode and vice versa will need to be very intuitive and natural without any additional training. When the vehicle requires a driver to take over the control, the most natural way for drivers will be for the alert to come via speech request. Likewise, allowing drivers to request the vehicle to take over control using speech commands will be natural and convenient for drivers.

The increasing automation of driving may allow cars to contain larger screens. Virtual reality or mixed reality could make use of this screen space to present content, possibly as a three-dimensional (3-D) virtual world. CID systems may play a special role in navigating such 3-D worlds while driving; for example, one may, with a speech request, switch the display to a place not shown on a screen and preview its surroundings.

Recognizing that CID technology can both increase driving safety and improve user experience allows one to see new useful opportunities for in-car speech systems. These, however, require more conversational intelligence than is currently available to ensure that drivers' and passengers' voice interactions with systems are natural and not cognitively demanding or distracting.

CID systems in the context of the Internet of Things

Today, a single car can contain more than 100 sensors, supporting various vehicle functionalities and detecting occupant states. As vehicle technology advances, increasing numbers and types of sensors will appear in cars. Some of these will improve vehicle automation and safety, some will sense the external environment, for example, measure air quality, and others will sense occupants' physical states and actions, such as posture or alertness. Such sensors can provide much of the information needed for improving the conversational intelligence of CID systems.

With the increase of connectivity in the Internet of Things (IoT), devices and sensors will increasingly become more diverse with much richer information-exchange functionalities, expressing additional device status, features, operating instructions, or maintenance needs, etc. With new form factors of devices and sensors, the success of touch screens on smart phones for information exchange may not be easily repeated in cars where physical spaces are limited or design flexibility is required. As the screen size decreases, speech will become a more preferred information exchange modality.

The IoTs will also bring in much more content and many more services for drivers to access. One may receive a scenic spot description or hotel vacancy advertisement along a highway not through traditional billboards but rather from the Internet in real time. The general IPAs may intend to support people for these needs. However, given the typically fast-changing environment involved in driving, the in-vehicle use cases add much more dynamic and context-sensitive requirements for such assistance. In-vehicle CID systems have the potential in using in-vehicle sensors, such as gas-tank level, to find better solutions for the drivers. One will expect a tight integration of the general IPAs and embedded CID systems to offer the drivers synergized benefits from both systems.

Expectation for better in-vehicle CID systems

In smarter cars with high autonomy and connectivity, speech systems need more conversational intelligence, as we have seen. Fortunately, the popular demand for better IPAs and mobile voice technology is driving improvements in conversational agents by increasing conversational intelligence in many of the ways previously mentioned in this article to be necessary for in-vehicle CIDs. Integrating developments in safety-optimized autonomy and conversational intelligence offers much promise for the CIDs that can meet automotive requirements for naturalness, ease of use, low cognitive demand on users, and minimal distraction from interacting with the CID itself. Of course, automobiles constitute a unique environment for speech in a multimodal setting, and additional research is needed specific to the automotive milieu, as discussed previously.

A key aspect with in-vehicle dialog systems is that the interaction is often carried out while a driver is operating a vehicle as a secondary task.

Moving forward, we foresee that CID systems would increasingly offer explanations about the vehicle itself, including functional operations, vehicle status, maintenance requirements, or even recommended driving styles for extended uses or environmental impact. CID systems will further act as a mediator to synchronize the content and services from different IPAs and integrate them with in-vehicle information. They will collaboratively support drivers on various activities with expert advisories, and communicate properly based on their cognitive and emotional state (Figure 3).

A successful deployment of sophisticated in-vehicle CID systems in the future would require a breakthrough in the system development process from specification, development, testing, and validation in the automotive industry to ensure high-quality but low-cost software. The complexity in introducing many additional sensors into the vehicle combined with much more content and many more services from the cyberworld should not be underestimated. It is quite possible that additional new layers will be introduced in the infotainment architecture to simplify the development and testing process. New standards may need to be introduced to facilitate industry-wide collaboration and incentivize the adoption of new technologies with affordable cost. Additional features may be also introduced to support the privacy and security in both the physical or cyberworlds by using CID systems to recognize and track drivers' identity and set up proper access or operation restrictions while driving.

Conclusions

An in-vehicle dialog system is a complex system that involves broad interdisciplinary knowledge and technologies from automatic speech recognition, spoken language understanding,

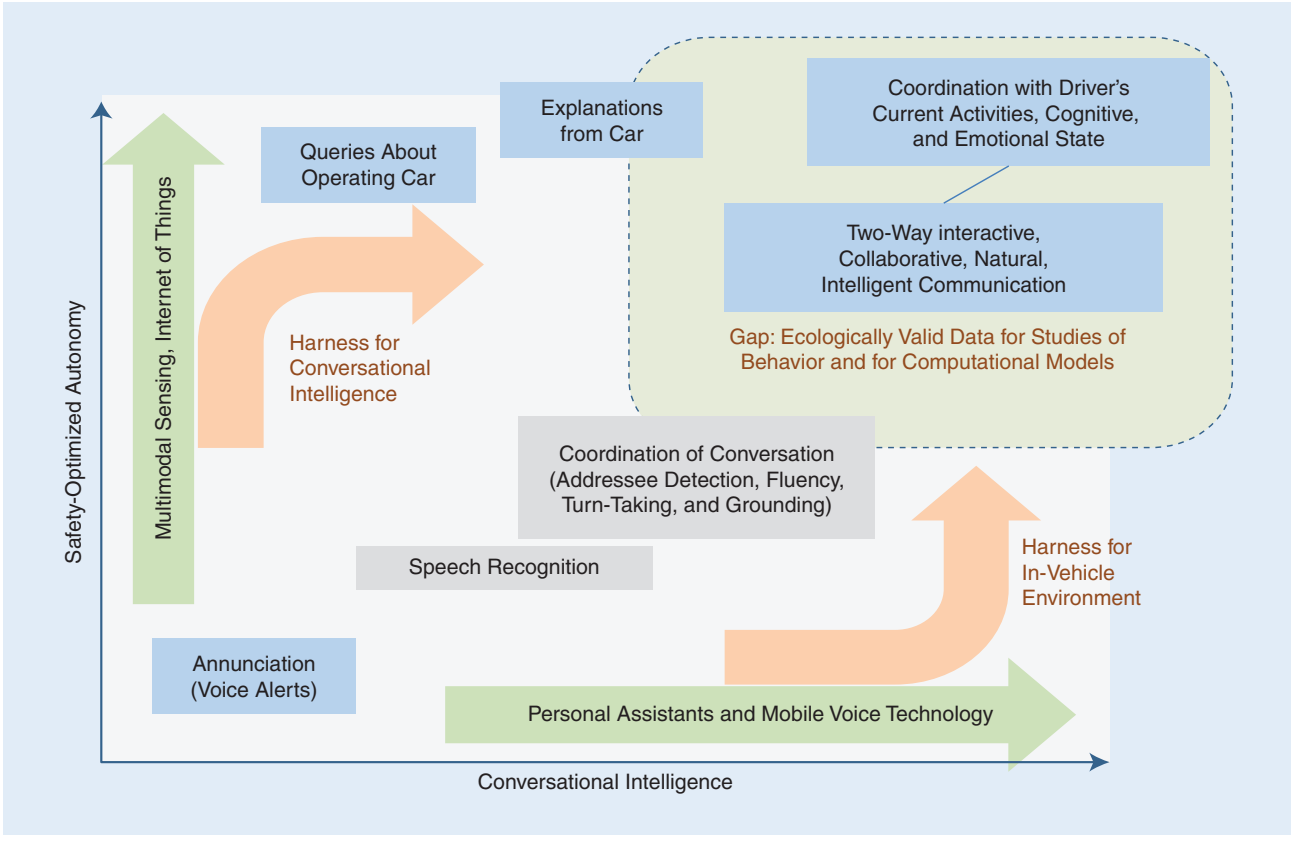


FIGURE 3. Future directions of CID systems.

DM, natural language generation, and TTS synthesis. In addition to challenges in development of other spoken dialog systems, a key aspect with in-vehicle dialog systems is that the interaction is often carried out while a driver is operating a vehicle as a secondary task. To ensure driving safety and intelligent interaction, it is necessary to provide an effective and user-friendly interaction between the driver and vehicle. The development of such a system requires the support of both the automotive and high-tech industries.

To ensure driving safety and intelligent interaction, it is necessary to provide an effective and user-friendly interaction between the driver and vehicle.

As sensors have become more reliable and accurate, the incorporation of multimodal interaction beyond audio itself allows the dialog system to detect additional nonverbal communications such as the intention and emotion of drivers. This information is advantageous for the dialog system to manage its interaction toward accomplishing its task based on the state or status of the driver, vehicle, and environment. Design of the intelligent in-vehicle dialog system also exploits such information for modeling behavior and usage patterns of a driver to adapt itself toward more effective interaction with an individual driver.

In the context of autonomous driving and the IoT, we expect to see more integration of speech-enabled technology with general IPAs fully connected with in-vehicle systems. We believe that in-vehicle dialog system technology will remain on demand with much more enriched features in the future for both the current human-centric driving paradigm and autonomous driving paradigm.

Acknowledgment

We would like to thank Dr. Liberty Lidz for her support on the preparation of the manuscript.

Authors

Fuliang Weng (Fuliang.weng@us.bosch.com) received his B.S. degree from Fudan University, Shanghai, China, in 1984. He was in a joint M.S. and Ph.D. program of Fudan University from 1984 to 1989, and completed all requirements but his Ph.D thesis. In addition, Fuliang received an M.S. degree from New Mexico State University, Las Cruces, in 1993. He is the general manager of Bosch Language Services. Previously, he served as chief expert and director of user technologies at Bosch Corporate Research. He has more than 70 research publications and 60 issued or pending patents and has received multiple national and industrial awards.

Pongtep Angkititrakul (pongtep.angkititrakul@us.bosch.com) received his B.Eng. degree from Chulalongkorn University, Thailand, and his M.S. and Ph.D. degrees in electrical engineering from the University of Colorado at Boulder. Currently, he is a lead engineer at Robert Bosch LLC. Previously, he was a visiting researcher at Toyota Central R&D, Japan, where he contributed to algorithm development for advanced driver assistance system and driver monitoring technology, as well as the Toyota award-winning project for drowsiness detector product. He organized and served as the general cochair of the Biennial Workshop for Digital Signal Processing for In-Vehicle Systems and Safety in 2015.

Elizabeth E. Shriberg (elizabeth.shriberg@sri.com) is a principal scientist at SRI International. She is also affiliated with Johns Hopkins University, the University of Maryland, and the International Computer Science Institute, and she was recently at Microsoft Research. Her government and commercial work models spoken communication with focus areas in emotion, cognition, health, prosody, fluency, dialog, and speaker modeling, in which she holds multiple patents. She has authored more than 250 research publications and serves on the International Speech Communication Association (ISCA) Advisory Council and on journal, conference, and review boards for the speech science and speech technology communities. She is a fellow of SRI and ISCA.

Larry Heck (larry.heck@ieee.org) received his B.S.E.E. degree from Texas Tech in 1986 and his M.S.E.E. and Ph.D. degrees from the Georgia Institute of Technolgy in 1989 and 1991, respectively. He worked at SRI (1992–1998) in the Speech Technology and Research Laboratory. He was with Nuance (1998–2005) as vice president of research and development, where he led natural language processing, automatic speech recognition, voice authentication, and text-to-speech synthesis. At Yahoo! (2005–2009), he was the vice president of search and advertising sciences. While at Microsoft (2009–2014), he cofounded the Cortana personal assistant as chief scientist of speech. He is currently a research director for dialog at Google. He is a Fellow of the IEEE and holds more than 50 U.S. patents.

Stanley Peters (peters@csli.stanford.edu) is emeritus professor of Linguistics and Symbolic Systems at Stanford University, where he chaired the Department of Linguistics (1996–2002) and directed the Center for the Study of Language and Information (1987–1990 and 2008–2013). He previously taught at the University of Texas at Austin (1966–1983). His research interests are mathematical properties of grammar, logical semantics of natural languages, artificially intelligent tutoring systems, human–robot interaction, computational modeling of conversation, and spoken dialog systems. He has authored three books and more than 120 papers and holds five patents. He is a fellow of the Linguistic Society of America.

John H.L. Hansen (john.hansen@utdallas.edu) received his B.S.E.E. degree from Rutgers University, New Jersey, in 1982 and his M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology in 1983 and 1988, respectively. He serves as associate dean for research at the University of Texas–Dallas, where he founded the Center for Robust Speech Systems. He has been an associate editor for numerous IEEE publications

and organized the International Speech Communication Association (ISCA) INTERSPEECH 2002 and coorganized and was the tehnical program chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing 2010 and the IEEE Spoken Language Technology Workshop 2014. He is a Fellow of the IEEE and ISCA.

References

[1] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The SPEECHDAT-CAR multilingual speech databases for in-car applications: Some first validation results," in *Proc. European Conf. Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 2279–2282.

[2] J. H. L. Hansen, P. Angkititrakul, J. Plucienkowski, S. Gallant, U. Yapanel, B. Pellom, W. Ward, and R. Cole, "CU-move: Analysis and corpus development for interactive in-vehicle speech systems," in *Proc. ISCA Interspeech Conf.*, Aalborg, Denmark, Sep. 2001, pp. 2023–2026.

[3] N. Kawaguchi, S. Matsubara, I. Kishida, Y. Irie, H. Murao, Y. Yamaguchi, K. Takeda, and F. Itakura, "Construction and analysis of a multi-layered in-car spoken dialogue corpus," in *Proc. ISCA Interspeech Conf.*, Aalborg, Denmark, Sep. 2001, pp. 2027–2030.

[4] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. Di Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker, "The AT&T–DARPA communicator mixed-initiative spoken dialogue system," in *Proc. Int. Conf. Spoken Language Processing*, 2000, pp. 122–125.

[5] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. ISCA Interspeech Conf.*, Korea, 2004, pp. 2489–24924.

[6] 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 Trans. Intell. Transport. Syst.,* vol. 12, no. 4, pp. 1609–1623, Dec. 2011

[7] P. Geutner, F. Steffens, and D. Manstetten, "Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments," in *Proc. 3rd Int. Conf. Language Resources and Evaluation*, Canary Islands, 2002, pp. 1588–1593.

[8] O. Lemon, K. Georgila, J. Henderson, M. Gabsdil, I. Meza-Ruiz, and S. Young, "D4.1: Integration of learning and adaptivity with the ISU approach," TALK Project No. IST-507802, Information Society Technologies, Germany, 2005.

[9] 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, and H. Schmidt, "CHAT: A conversational helper for automotive tasks," in *Proc. ISCA Interspeech Conf.*, 2006, pp. 1061–1064.

[10] B. Lathrop, H. Cheng, F. Weng, R. Mishra, J. Chen, H. Bratt, L. Cavedon, C. Bergmann1, T. H. Bender, H. P. Barry, B. Bei, M. Raya, and E. Shriberg, "A Wizard of Oz framework for collecting spoken human- computer dialogs: An experiment procedure for the design and testing of natural language in-vehicle technology systems," in *Proc. 12th World Congress Intelligent Transport Systems*, San Francisco, CA, 2005.

[11] B. Thomson and S. Young. "Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems," *Comp. Speech Lang.* vol. 24, no. 4, pp. 562–588, Oct. 2010.

[12] D. Hurwitz, E. Miller, M. Jannat, L. Boyle, S. Brown, A. Abdel-Rahim, and H. Wang, "Improving teenage driver perceptions regarding the impact of distracted driving in the Pacific Northwest," *J. Transport. Safety Security*, vol. 8, no. 2, pp. 148–163, 2016.

[13] C. Muller, G. Weinberg, and A. Vetro, "Multimodal input in the car, today and tomorrow," *IEEE Multimedia*, vol. 18, no. 1, pp. 98–103, 2011.

[14] Y. C. Cheng, K. Li, Z. Feng, F. Weng, and C.-H. Lee, "Online whole-word and stroke-based modeling for hand-written letter recognition in in-car environments," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 847–852.

[15] G. Tur, A. Stolcke, L. Voss, D. Hakkani-Tur, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, M. Purver, K. Riedhammer, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting assistant system," *IEEE Trans. Audio Speech Lang. Processing*, vol. 18, no. 6, pp. 1601–1611, Aug. 2010.

[16] Deep learning (2016). [Online]. Available: https://en.m.wikipedia.org/wiki/Deep_learning

[17] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Process. Mag.,* vol. 12, no. 3, pp. 25–42, May 1995.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.,* vol. 29, pp. 82–97, Nov. 2012.

[19] L. Heck, Y. Konig, M. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Commun.*, vol. 31, no. 2, 2000, pp 181–192.

[20] Y. Konig, L. Heck, M. Weintraub, and M. Sonmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," presented at the RLA2C, ESCA Workshop, Avignon, France, 1998.

[21] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing,* Vancouver, Canada, 2013, pp. 7398–7402.

[22] X. Feng, B. Richardson, S. Amman, and J. Glass, "On using heterogeneous data for vehicle-based speech recognition: A DNN-based approach," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Brisbane, Australia, 2015a, pp. 4385–4389.

[23] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. (2014). Deep speech: Scaling up end-to-end speech recognition [Online]. Available: https://arxiv.org/abs/1412.5567

[24] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: Analytical evaluation," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 241–252, May 1999.

[25] J. H. L. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward "CU-Move: Robust speech processing for in-vehicle speech systems," in *Proc. Int. Conf. Spoken Language Processing*, 2000, vol. 1, pp. 524–527.

[26] X. Zhang and J. H. L. Hansen, "CSA-BF: A constrained switched adaptive beamformer for speech enhancement and recognition in real car environments," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 733–745, Nov. 2003.

[27] Z. Zhou, R. Botros, H. Lin, Y. Hao, Z. Feng, and F. Weng, "Hybrid speech recognition using pairwise classification and deep neural network based frameworks with long-distance constraints," Bosch Internal Report No. CR/RTC-NA-847, Robert Bosch LLC, Palo Alto, CA, 2016.

[28] UKIP Media & Events Ltd [Online]. Available: http://www.automotive-interiors-expo.com/english/awards_14_winners.php

[29] ISO [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:26262:-6:ed-1:v1:en

[30] IEC [Online]. Available: http://www.iec.ch/functionalsafety/

[31] F. Soong and E. Woudenberg, "Hands-free human-machine dialogue-corpora, technology and evaluation," in *Proc. Int. Conf. Spoken Language Processing*, 2000, vol. 4, pp. 41–44.

[32] Q. Zhang, F. Weng, and Z. Feng, "A progressive feature selection algorithm for ultra large feature spaces," in *Proc. 21st Int. Conf. Computational Linguistics,* 2006, pp. 561–568.

[33] X. Maas, D. Jurafsky, Z. Xie, and A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," presented at the North American Association for Computational Linguistics, Denver, CO, 2015.

[34] H. H. Clark, "Managing problems in speaking," *Speech Commun.,* vol. 15, no. 3–4, pp. 243–250, Dec. 1994.

[35] H. Arsikere, E. Shriberg, and U. Ozertem, "Computationally-efficient endpointing features for natural spoken interaction with personal-assistant systems," in *Proc. Int. Conf. Acoustics Speech Signal Processing*, 2014, pp. 3241–3245.

[36] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog," in *Proc. Int. Conf. Spoken Language Processing*, 2002.

[37] E. E. Shriberg, "To 'errrr' is human: Ecology and acoustics of speech disfluencies," *J. Int. Phonetic Assoc.,* vol. 31, no. 1, pp. 153–169, 2001.

[38] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *J. Pragmatics,* vol. 32, no. 8, pp. 1177–1207, 2000.

[39] P. Gao, H.-W. Kass, D. Mohr, and D. Wee, "Automotive revolution: Perspective towards 2030—Advanced industries," McKinsey & Company, New York, Jan. 2016.

SP