

# Spoken Language Interaction with Robots

## Research Issues and Recommendations



Matthew Marge, Carol Espy-Wilson, Nigel G. Ward

# **Spoken Language Interaction with Robots: Research Issues and Recommendations**

Report from the NSF Future Directions Workshop, October 2019

**Matthew Marge**, DEVCOM Army Research Laboratory  
**Carol Espy-Wilson**, University of Maryland, College Park  
**Nigel G. Ward**, University of Texas at El Paso

## **Workshop Participants and Report Contributors**

**Abeer Alwan**, University of California at Los Angeles  
**Yoav Artzi**, Cornell University  
**Mohit Bansal**, University of North Carolina at Chapel Hill  
**Gil Blankenship**, University of Maryland  
**Joyce Chai**, University of Michigan  
**Hal Daumé III**, University of Maryland, College Park  
**Debadeepa Dey**, Microsoft  
**Mary Harper**, DEVCOM Army Research Laboratory, emerita  
**Thomas Howard**, University of Rochester  
**Casey Kennington**, Boise State University  
**Tatiana (Tanya) Korelsky**, National Science Foundation  
**Ivana Kruijff-Korbayová**, DFKI  
**Dinesh Manocha**, University of Maryland, College Park  
**Cynthia Matuszek**, University of Maryland, Baltimore County  
**Ross Mead**, Semio  
**Raymond Mooney**, University of Texas at Austin  
**Roger K. Moore**, University of Sheffield  
**Mari Ostendorf**, University of Washington  
**Heather Pon-Barry**, Mount Holyoke College  
**Alexander I. Rudnicky**, Carnegie Mellon University  
**Matthias Scheutz**, Tufts University  
**Robert St. Amant**, DEVCOM Army Research Laboratory  
**Tong Sun**, Adobe  
**Stefanie Tellex**, Brown University  
**David Traum**, USC Institute for Creative Technologies  
**Zhou Yu**, University of California at Davis

November 12, 2020

## **Executive Summary**

With robotics rapidly advancing, more effective human-robot interaction is increasingly needed to realize the full potential of robots for society. While spoken language must be part of the solution, our ability to provide spoken language interaction capabilities is still very limited. The National Science Foundation accordingly convened a workshop, bringing together speech, language, and robotics researchers to discuss what needs to be done. The result is this report, in which we identify key scientific and engineering advances needed to enable effective spoken language interaction with robotics.

Our recommendations broadly relate to eight general themes. First, meeting human needs requires addressing new challenges in speech technology and user experience design. Second, this requires better models of the social and interactive aspects of language use. Third, for robustness, robots need higher-bandwidth communication with users and better handling of uncertainty, including simultaneous consideration of multiple hypotheses and goals. Fourth, more powerful adaptation methods are needed, to enable robots to communicate in new environments, for new tasks, and with diverse user populations, without extensive re-engineering or the collection of massive training data. Fifth, since robots are embodied, speech should function together with other communication modalities, such as gaze, gesture, posture, and motion. Sixth, since robots operate in complex environments, speech components need access to rich yet efficient representations of what the robot knows about objects, locations, noise sources, the user, and other humans. Seventh, since robots operate in real time, their speech and language processing components must also. Eighth, in addition to more research, we need more work on infrastructure and resources, including shareable software modules and internal interfaces, inexpensive hardware, baseline systems, and diverse corpora.

Research and development that prioritizes these issues will, we believe, provide a solid foundation for the creation of speech-capable robots that are easy and effective for humans to work with.

- 1. Introduction**
- 2. User Experience Design**
- 3. Robustness and Adaptability**
- 4. Infrastructure**
- 5. Audio and Speech Processing, Speech Recognition, and Behavior Signal Analysis**
- 6. Language Understanding**
- 7. Dialogue and Human Communication Dynamics**
- 8. Language Generation and Speech Synthesis**
- 9. Policy**

# 1 Introduction

As robotics advances, spoken language interaction is becoming increasingly necessary. Yet robot researchers often find it difficult to incorporate speech processing capabilities, and speech researchers seldom appreciate the special needs of robot applications. We hope to help members of both tribes, as well as social scientists, other researchers, and subject matter experts to better understand the difficulties, possibilities, and research issues in speech for robots; to catalyze new research projects in this area; and to thereby bring us closer to the vision of truly satisfying spoken language interaction with robots.

Thus this report aims to identify the key challenges that robotics brings for spoken interaction, and the key issues in designing robot systems able to make effective use of speech. We make 31 recommendations, relating broadly to issues of human needs, sociality and interaction, robustness, adaptation, multimodality, representations, timing, and infrastructure. We address these recommendations to funding agencies, leaders in industry, principal investigators, graduate students, developers, and system integrators.

Our work started with a two-day meeting, October 10 and 11, 2019, hosted by Carol Espy-Wilson at the University of Maryland with support from Tanya Korelsky's program at the National Science Foundation, followed by a year of discussion and reflection. Unlike more general roadmaps for research in robotics, dialogue systems, artificial intelligence, and related areas [3, 12, 26, 30, 38, 47, 98, 109, 112, 114, 116], this report focuses on issues that are especially critical or especially challenging for speech in robotics. This report is intended to complement the work of the similarly-themed Dagstuhl workshop on Spoken Language Interaction with Virtual Agents and Robots [36]: their report provides more information on recent research advances and trends, while here we focus on making specific recommendations.

## Why Spoken Language Interaction with Robots?

Across a wide range of applications, spoken language interaction with robots has great promise. As shown in Figure 1, among the most immediate applications are in education, healthcare, field assistance (both civilian and military), and the consumer market. Robot tutors can provide a sense of presence, which can stimulate interests and joy in learning and improve learning outcomes [7]. Moreover, the learning experience with a robot can be tailored to individual students to shape the learning process [91]. Speech-capable robots can amplify a healthcare worker's capabilities in remote care (e.g., a rehabilitation robot tracking and assisting a patient with physical therapy) [50], as well as create engaging social interactions with elderly patients that may be isolated due to medical reasons [21]. Robots fielded in search-and-rescue, humanitarian relief, or reconnaissance scenarios also require an expressive interface like spoken language, especially when situations call for rapid specification of robot goals (such as when the human teammate is engaged in other activities) [64]. Speech also allows an operator to oversee and command teams of robots doing tasks without the use of a handheld controller [72]. Finally, consumer robotics (including but not limited to entertainment, security, and household applications) represents the fastest-growing market segment, projected to be a \$23-billion market by 2025 [115]. Products such as Pepper, Jibo, and Anki Vector have shown the promise of speech interfaces to increase the ease-of-use of robots in homes by non-expert users.

Reasons why spoken language interaction with robots will greatly benefit human society include:

- Among the various ways to exchange information with robots, spoken language has the potential to often be the fastest and most efficient. Speed is critical for robots capable of interacting with people in real time. Especially in operations where time is of the essence, slow performance is equivalent to failure. Speed is required not only during the action, but also in the human-robot communication, both prior to and during execution.
- Spoken language interaction will enable new dimensions of human-robot cooperative action, such as the realtime coordination of physical actions by human and robot.



Figure 1: Spoken language interaction with robots across four application areas: education, healthcare, field assistance, and entertainment [65]. Image credits: RoboKind, SoftBank Robotics, and the United States Army. Reproduced with permission.

- Spoken language interaction is socially potent, and will enable robots to engage in more motivating, satisfying, and reassuring interactions, for example, when tutoring children, caring for the sick, and supporting people in dangerous environments.
- As robots become more capable, people will *expect* speech to be the primary way to interact with robots.
- Robots that you can talk with may be simply better liked, a critical consideration for consumer robotics.
- Robots can be better communicators than disembodied voices [34]; being co-present, a robot's gestures and actions can reinforce or clarify a message, help manage turn-taking more efficiently, convey nuances of stance or intent, and so on.
- Building speech-capable robots is an intellectual grand challenge that will drive advances across the speech and language sciences and beyond.

Not every robot needs speech, but speech serves functions that are essential in many scenarios. Meeting these needs is, however, beyond the current state of the art.

## Why Don't We Have It Yet?

At first glance, speech for robots seems like it should be a simple matter of plugging in some off-the-shelf modules and getting a talking robot [81]. But it's not that easy. This report will discuss the reasons at length, but here we give an initial overview of the relevant properties of robots and spoken communication.

**What is a robot, in essence?** While in many ways a robot is like any other AI system that may need to converse with a human, there are some fundamental differences [85]. Notably, in general:

1. A robot is situated; it exists at a specific point in space, and interacts with the environment, affecting it and being affected.
2. A robot provides affordances; its physical embodiment affects how people perceive its actions,

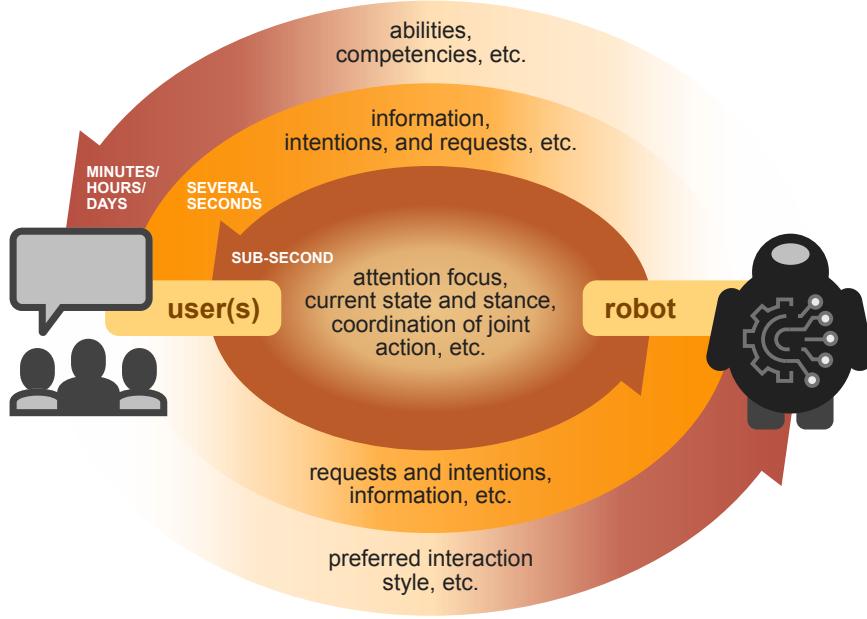


Figure 2: Speech-based communication between users and robots at three time scales. The audio signal at any moment may include information relating to all three time scales.

speech, and capabilities, and affects how they choose to interact with it.

3. A robot has very limited abilities, in both perception and action; it is never able to fully control or fully understand the situation.
4. A robot exists at a specific moment in time, but a time where everything may be in a state of change — the environment, the robot’s current plans and ongoing actions, what it’s hearing, what it’s saying, and so on.

Not every robot brings unique challenges for speech — a robot that just sits on a desk, chatting and smiling, can work much like any other conversational agent — but as robots become more capable, speech becomes more challenging.

**What is spoken communication, in essence?** It is not just audible text; nor is it just transmitting packets of information back and forth [93]. Rather, in general:

1. Spoken communication is a way for people to indicate their needs, desires, goals, and current state. State includes internal state, such as stress level, level of interest, and overall emotional state, and also stance, such as attitudes and intentions regarding the current activity, teammate actions, and specific objects in the environment.
2. Spoken communication can relate to the open world, as it calls out objects of interest, signals upcoming actions, enables coordination with teammates, supports timely action, and so on.
3. Spoken communication can accompany actions and gestures to emphasize or disambiguate intentions.
4. Spoken communication serves interpersonal functions — in motivating or guiding teammates, as well as in showing awareness of their contributions, their current state, their autonomy, their value, and so on.
5. Spoken communication styles can portray diverse information about the individual, or robot, including its abilities, level of competence, desired interaction style, and so on.
6. Finally, spoken communication operates at various timescales, as suggested by Figure 2. Of course, the robot’s audio output should accurately give the user information on the robot’s current knowledge state, needs, and intentions, and conversely the robot should understand

instructions from the user. Such calmly paced utterances and responses have been the primary focus of past research. Yet robots often also need to be able to interact swiftly with the user, enabling the direction of attention, exploitation of rapid dialogue routines, and tight coordination of joint action. Moreover, robots should be able to use spoken interaction when establishing long-term expectations. In one direction, the robot’s voice and turn-taking style should enable the user to infer the robot’s “personality,” including what the robot is capable of and how it can best be interacted with. In the other direction, the robot should be able to infer, from the user’s speaking style and interaction style, how this specific user likes to interact, and to adjust its behavior parameters accordingly.

Not every robot needs competence in all these functions of speech. If a robot’s job is just to pull weeds, it may need speech only for receiving simple commands and providing simple status reports. But to fully exploit the power of speech, roboticists will need to endow their creations with new representations and new functionality.

We offer the following contributions:

- A brief overview of the critical challenges associated with spoken language interaction with robots, spanning user experience design, robustness, adaptability, infrastructure, speech processing, behavior signal analysis, language understanding, dialogue, human communication dynamics, language generation, and speech synthesis.
- A set of concrete recommendations to researchers and program managers alike that we believe will accelerate progress in this area.

This report is organized as follows. The following sections describe research issues and recommendations associated with user experience design (Section 2), robustness and adaptability (Section 3), technical infrastructure (Section 4), audio and speech processing, speech recognition, and behavior signal analysis (Section 5), language understanding (Section 6), dialogue and human communication dynamics (Section 7), and language generation and speech synthesis (Section 8). We conclude the report with a list of recommendations tailored to funding agencies; these aim for improving the intellectual merit and broader impact of research in spoken language interaction with robots (Section 9).

## 2 User Experience Design

At the top level, we can say that there are the three driving forces that underlie most robotics research, and in particular, most projects relating to speech for robots: the visions, the technologies, and the needs. While many technical challenges remain, and we still need the inspiring visions, the field is now reaching the point where focus on the needs — the human needs — should become the main driver. This section discusses what this implies, organized around four broad recommendations.

**Recommendation U1:** *Focus on language not only as a way to achieve human-like behaviors, but also as a way to support limited but highly usable communications abilities.*

From the earliest days, an inspiration for robotics research has been the creation of human-like artifacts. However, experience with many user interfaces has shown that aiming to emulate a human too closely is often a recipe for failure [8]. Grand ambitions are good, but we also need to focus on engineering spoken interaction capabilities to maximize usability and utility. Simple, minimal interaction styles can even be natural, in their own way, for people. Empirically, even when designers aim to support natural “conversational” interaction, users often resort to formulaic language and focus on a handful of interaction routines that reliably work for them [83].

**Recommendation U2:** *Deliberately engineer user perceptions and expectations.*

People invariably form mental models of the artifacts they interact with. These mental models help them to predict what these artifacts are capable of and how best to interact with them. Without guidance, users can easily form misguided mental models, based for example on the interactions seen with robots in science fiction movies. But designers can help users form more accurate mental models of robots by choosing appropriate visual appearances, selecting appropriate voices, and implementing appropriate behavioral competencies. Using both first impressions and accumulated experience, users can thus come to feel comfortable in dealing with a robot.

A complicating factor here is that the state of the art in speech and robotics today is uneven: some components perform impressively, while others lag. In implemented systems, lack of coherence can be confusing. Obviously this implies the need for more work on the deficient modules, but varying levels of ability will always be a problem in real robots. Accordingly, the abilities exposed to users may need to be deliberately limited [82], to avoid giving an exaggerated perception of competence that can mislead users regarding how to behave and what to expect. More generally, designers need to avoid possible “habitability gaps” [90], where usability drops as flexibility increases. This can be an instance of the “uncanny valley” effect, in which a near human-looking artifact (such as a humanoid robot) can trigger feelings of eeriness and repulsion [84]. Of course there can be trade-offs between attracting users to engage in the first place and enabling truly effective interaction [68].

**Recommendation U3:** *Work to better characterize the list of communicative competencies most needed for robots in various scenarios.*

Today some research in speech for robotics follows well-worn paths, extending trajectories inspired by classic taxonomies of language and behavior. These topics and issues are not, however, always the most practically important for human-robot interaction. As suggested above in Figure 1 and elaborated below, there are many diverse communicative behaviors worth modeling. In particular, we see the need to model spoken interaction at rapid time scales, and to model it as centrally involving social intentions. These abilities do not represent merely nice-to-have features; rather they provide the very foundation of spoken interaction.

More generally, we see value in occasionally stepping back from the bustle at the speech technology research forefront, to observe how people actually communicate and what is most important for communicative success. This will enable us to thoughtfully determine what aspects of speech are truly the most important, across diverse scenarios, and thus to prioritize what to work on for the sake of maximizing the effectiveness of future speech-capable robots.

**Recommendation U4:** *Design for use in multi-party and team situations.*

Today human-robot interaction is generally designed to support single users. While we do not wish to pre-judge what new priorities might be uncovered by such diligent observation work, it is clear that one priority must be support for multi-party interaction. Many robots will function in environments with more than one person. These may include members of a team tasked to jointly work with the robot, complete bystanders, or anything in between. Moreover the roles of the humans may change over time.

Support for the ability to deal with multiple humans in the environment has implications for all components of robots, including audio processing, speech recognition, speaker diarization, language understanding, computer vision, situation planning, action planning, and speech generation and synthesis. In particular, a robot must be able to detect whether it is the intended recipient of some communication or not [28]. Multi-party interaction with robots has already been demonstrated in some situations, for example with the Furhat [2] and Directions [16] robots, and with Waseda’s Facilitation robot, used to help people who might otherwise get left behind in complex multi-party conversations [74], but enabling multi-part interaction more generally remains a challenge.

### 3 Robustness and Adaptability

In robotics research, like many other fields, successful demos are celebrated: the demonstration of a new technology that brings us closer to a noble vision of the future can be a source of great inspiration. However, success in demos is not very predictive of success in deployment, especially since most demos only illustrate an ideal case. However, the field is maturing, and, as we increasingly target solid experimental validation of capabilities and real deployments in the open world, considerations of robustness and adaptability are becoming ever more essential. This section recommends some strategic directions towards improving robustness and adaptability.

**Recommendation R1:** *Include partially redundant functionality.*

Human interaction is highly redundant, with the same message often being conveyed by words, prosody, gaze, posture, pose, facial expressions, hand gestures, actions, and so on [1, 43, 111]. While robots can perform well in demos with only one of these functions, this is only true when both the environment and the user are tightly constrained. Adding competence with these other modalities, beyond just the words alone, can contribute to robustness. Achieving this requires better scientific understanding of these aspects of behavior, more shareable software modules, more explorations of utility for various use cases, and more work on cross-modality integration, both for speech understanding and for speech and behavior synthesis.

**Recommendation R2:** *Make components robust to uncertainty.*

Demos can be staged so that the robot has complete knowledge of all relevant aspects of the situation, but in open worlds such knowledge is not possible. To give three examples: First, for intent recognition, a developer cannot assume that a robot can ever have a 100% correct understanding of the user’s goals and intents; rather it will invariably need to maintain a distribution of belief over multiple hypotheses. Second, a developer cannot treat the interface between the language understanding module and the response planning module as a single-predicate symbolic representation, given the inevitability of alternative possible real-world referents and meanings that might be ambiguously or deliberately bundled up in any user utterance. Third, a dialogue manager cannot be a simple finite state machine, as robots need to track multiple dimensions and facets of the current situation, typically none of which can be identified with full confidence.

It is easy to say that achieving robustness mandates that each component constantly tracks multiple hypotheses and maintains a probability estimate for each. However, doing so involves many challenges. One of these is the need for something of a change of mindset: we may need to accept the fact that simple, understandable, inspectable representations may not be generally adequate. Another is that, even when we know how to make one or two components probabilistic, it often remains difficult to integrate them, let alone to design a software architecture for a robot that operates entirely in this way.

One possible way out is end-to-end modeling, in which module boundaries are erased, everything is jointly optimized, and all mappings are learned directly from data. However, full end-to-end training will likely never be possible for robots, as we will never have enough training data for all the complex and diverse tasks that robots need to do.

Thus we need research on ways to move towards systems that can represent and model uncertainty throughout all components. As suggested by the recent explosive advances in spoken dialogue systems — in large part attributable to the adoption of this perspective [95], along with innovations in the design of suitably simple representations and the judicious application of machine learning from large data — the potential benefits are large.

**Recommendation R3:** *Focus not only on improving better core components, but also on cross-cutting issues.*

From a robot designer’s perspective, it would be convenient if natural language could be a simple add-on to an existing robot control architecture. Thus a naive approach to getting natural language onto robots is to add an automatic speech recognizer (ASR) whose output is fed into a finite state machine (FSM) that transitions to the appropriate next state and simultaneously selects the next output to send to the text-to-speech (TTS) synthesizer. While such systems can often meet the basic interaction needs in simple robotic applications — such as greeting a customer, prompting for a simple request, and following simple instructions — this simple ASR+FSM+TTS architecture is not sufficient for more natural interactions, especially not in “open worlds.” Unfortunately, research effort has tended to gravitate to these familiar components. While we can, and should, add components — such as a natural language understanding (NLU) module, a dialogue manager (DM), and a natural language generator (NLG) — there are many issues that fall through the cracks of such an architecture. These include ambiguity, grounding, social adeptness, prosody, and adaptation, some of which we discuss further below, and these, rather than further refinements to pure core technologies, are often of the first importance.

**Recommendation R4:** *Make every component able to support realtime responsiveness.*

Current dialogue-capable robots offer only slow-paced, turn-based interactions, with few exceptions [100]. This is now due less to actual processing time requirements, than to the architectures of our systems. In particular, it is far easier to build a system component if that component can delay the start of processing until the upstream model delivers a fully fleshed-out chunk of information. For example, it is easier to build a recognizer that waits until the user has produced a full turn and definitively ended it. Yet robots that operate on their own timescale can get out of sync with what the user is thinking, saying, and doing. Robots in general need to be responsive: to operate in real time.

Thus they will, in general, require incrementality in every component. That is, each component will need to process data as a continuous flow, incrementally and asynchronously updating its output representations or probability estimates as new information comes in. Incrementality in spoken dialogue has been an active area of research, with work on incremental turn management, speech recognition, semantics, dialogue management, language generation, and speech synthesis [11, 23, 55, 60, 62]. There are also general abstract models and toolkits for incrementality [10, 78, 97], but much remains to be done to make incrementality generally available to developers needing to add dialogue capabilities to robots [61].

Moreover, software for robots will generally need to model time explicitly, in every component. On the input side, robots have many sources of sensory input beyond the speech signal. These include cameras to take in scene information (e.g., for navigation or for identifying objects, people, or gestures), laser scanners to generate point clouds (e.g., for avoiding collisions), “introspection” on the internal states of the robot (e.g., state information about movement, location, or knowledge), as well as infrared sensors and GPS information. Different sensors operate at different sampling rates, and the downstream processes — for example speech recognition and object detection — have different processing speeds. These will produce different delays between events in the world and the time they are recognized. For example, if a user points to an object and then to a location while saying *put that there*, then the robot must appropriately fuse the information from the speech and visual inputs. The need for proper handling of time applies to all modules and aspects of processing, including perception, planning, action execution, speech recognition, gesture recognition, and speech and gesture production. The Microsoft Research Platform for Situated Intelligence [17] provides mechanisms for this, but issues of synchronization and temporal alignment still bring many challenges.

Similarly on the output side: spoken output must be timed and synchronized in concert with actions in other modalities, as discussed later. With advances in the synthesis of non-verbal actions, the need here is becoming more pressing. For example, a gesture at the wrong time can be far worse than no gesture at all, and small variations in the timing of responses to questions have large effects in the interpretation of what people really mean by those responses [18, 59, 94]. While we understand some aspects of these issues, so far our knowledge is mostly isolated to specific aspects of specific dialogue acts in specific contexts. In general, there is a need for more general models of how to time and align multimodal actions.

**Recommendation R5:** *Make systems and components adaptable to users.*

Every successful robot application today involves careful engineering to make it work for a specific user population. This is especially true for speech interfaces. This process is expensive and slow, so we need to face up to the challenges of making robots able to readily adapt, either to groups or to specific users and teams. This adaptation might be partly automatic, partly based on small sets of training data, and partly handled by exposing parameters that developers can easily adjust. Adaptation is also necessary as a way to overcome whatever biases might exist in training data, since no training set will ever precisely represent the desired robot behaviors.

Further, even within a target population, each user is an individual, and individuals will differ in age, gender, dialect, domain expertise, task knowledge, familiarity with the robot, and so on. One particular open challenge is that of adapting to the user’s interaction style preferences. Today our understanding of interaction style differences is limited. We do know, for example, that in multimodal interaction some people tend to make the pointing gesture in synchrony with the deictic, as in *put it there*, while others tend to point after the word *there* [86]. We know that some prefer swift turn taking with frequent overlaps, while others prefer to wait until the other is silent before speaking [103]. We know that some people like to explain things by a brief low-pitch monologue, while others tend to explain by interleaving short pieces of an explanation with frequent checks that the listener is following [111]. We know that dialogue partners often accommodate to resemble the interlocutor’s behavior in terms of surface-level features such as pitch height or speaking rate [48, 66]. In addition there is a rich folk vocabulary for describing interaction styles — including terms like stiff, withdrawn, shy, domineering, nerdy, oblivious, goofy, chatterbox, quick-witted, lively, and supportive — reflecting the importance of these styles for success in interactions. In the past, interaction style differences have not been a burning issue, since most people are able, entirely subconsciously, to model and adapt to the interaction styles of their teammates. Existing explorations in interaction styles and tendencies [45, 52, 54, 57, 77, 92] need to be extended to model more aspects of behavior. Beyond basic research, we need to develop ways for robots to effectively embody plausible and consistent interaction styles, and that can adapt to work with different interaction styles.

In general, if robots are to become effective partners, we need better models of the relevant dimensions of human variation, and of how to adjust behavior to work well with diverse human partners.

**Recommendation R6:** *Develop new ways to make components more reusable across tasks and domains.*

Robots need to be able to adapt to new tasks and domains. Linked to system-level adaptability and reusability, there is also the question of component-level adaptability. Developers of software components have a general strategic choice of aiming to optimize performance for a specific task by a specific robot, or of aiming to create reusable components that can be plugged into any architecture and used for any task. This is an essential tension, but one that can be partially alleviated. One direction is to investigate how to best define inter-component interfaces, either APIs or intermediate representations, to enable better information fusion and thus better decisions. A second direction is

to develop improved ways for rapid adaptation to new contexts of use [102], to enable the creation of components that are simultaneously high-performing and highly reusable. This may involve pre-training on massive data sets, with mechanisms for easily and robustly ablating or adapting the models to perform well on specific small domains, including, for some experimental purposes, exceedingly narrow domains.

## 4 Infrastructure

Research in spoken dialogue for robots has high barriers to entry. An ever-present problem is the intellectually demanding nature of conducting such interdisciplinary research: speech researchers must have access to a robot (or virtual robot), and robotics researchers must have access to some form of speech processing. To conduct research in this area requires mastering knowledge about robot platforms and spoken dialogue frameworks, including individual components of both. Significant effort is required to create systems that work, even minimally, for example because individual components, such as automatic speech recognition, even when well-tested in other domains, often don't transfer well to robots. This section describes infrastructure needs to conduct research on spoken language interaction with robots.

While the ultimate robot dialogue architecture will be complex and satisfy all the requirements discussed above, in the meantime, the community needs systems that make it easier for newcomers to get started, in the form of accessible robotic platforms that come coupled with accessible spoken dialogue systems. There have been many notable efforts that partially address this need, of which we can mention only a few.

**Robotics Technologies** The Robot Operating System (ROS) is a well-established platform for robotic systems. It enables low-cost robots built on ROS such as Duckietown (Ground and Air), MIT's Race Car platform, UW's MuSHR platform, and Artzi's drone platform. Other efforts are not necessarily tied to ROS but endeavor to make conducting robot research accessible: Microsoft Research's AirSim, Intel Research's CARLA, Facebook Research's AI Habitat, AI2's Thor, Nvidia's Isaac Platform, Semio Arora, the Furhat Robot, Anki Cozmo and Vector, and Misty Robotics's Misty II.

Modern architectures for human-robot interaction lack the ability to capture context with humans. If they do exist, they are usually in standalone models separate of the ROS software stack. As middleware, ROS does permit the capability to operate at latencies supporting realtime human interaction. In practice, however, researchers have generally made limiting and sometimes inappropriate assumptions about human interaction, in particular regarding highly time-sensitive tasks of the kind common in social robotics. The tools presently available using ROS for human-robot interaction can serve as a possible starting point for speech researchers interested in getting started in robotics. In particular, users can get started with ROS before buying any hardware since there are many virtual robot simulation packages.

**Speech Technologies** On the speech technology side, there are many available components, both commercial and open-source. For speech recognition, these include Kaldi, Sphinx, and Deep Speech 2, and proprietary cloud-based systems including Google ASR, Azure, and Alexa, although of course the lags of cloud-based speech recognizers are prohibitive for many robotics applications. The enduring problem here is that the openly available speech technologies generally have steep learning curves to getting started, while the proprietary commercial technologies do not support control of various components (e.g., controlling for vocabulary or retraining the language model to a robot task domain). For natural language understanding, there is Microsoft's LUIS, Rasa, and PolyAI. However, for reasons already noted, none of these components are truly "robot-ready." There are also a number of notable integrated frameworks and toolkits, several of which have been successfully used in robots, including InproTK, OpenDial, IrisTK, DIARC, RETICO, Plato, and the ICT Virtual Human Toolkit. These technologies provide evidence that a fully integrated platform is possible.

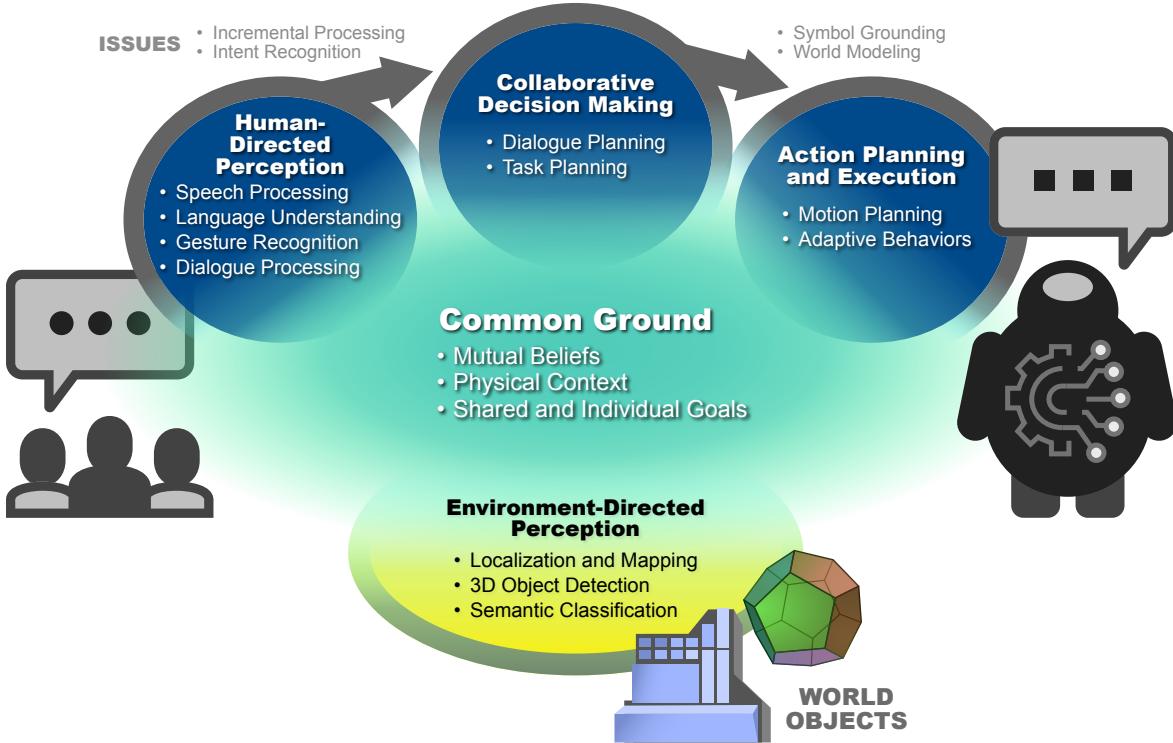


Figure 3: Core capabilities for robust spoken language interaction with robots include monitoring the human or humans a robot is interacting with and its physical surroundings, all of which contribute to common ground.

**Recommendation I1:** *Create and distribute one or more minimal speech and dialogue-capable robot systems.*

Although there are many tools and components, the easy-to-use ones are few and far between. Ideally there should be a basic dialogue-capable robot that people could simply buy and use out of the box. Of course, what such a robot should include is not obvious, given the many ranges of desired uses. They range from fun, for hobbyists, to serious, for conducting interesting extensions and experiments. They could support individual use or serve as a shared platform to help bring together researchers from robotics, speech, social interaction, computer vision, and so on. Possibilities span simply concatenated modules to systems engineered to support data capture, replay, visualization, informative experimentation, and performance analysis. They could vary from highly inclusive to exceedingly minimal, in the extreme case consisting of just a recognizer, synthesizer, and robot, with everything else to be designed (or kludged) for the intended use.

While no single solution will serve all needs, the availability of a basic dialogue-capable robot would greatly increase the number of researchers able to contribute to this area. To improve the likelihood of wide adoption, the platform should have an extensible simulated version compatible with ROS. In fact, a ROS-friendly simulated platform would be free for users to try out without the immediate costs of hardware. A critical step towards identifying worthy platforms requires considering the capabilities that robots will need to support spoken language interaction with robots, described below.

**Core Capabilities** Robots will need several features to converse with humans in a natural way. Figure 3 presents four primary capabilities that contribute towards common ground between humans and robots: (1) Human-Directed Perception, (2) Collaborative Decision Making, (3) Action Planning and Execution, and (4) Environment-Directed Perception. This diagram serves as a high-level

overview of how robot intelligence architectures break down the problem of converting all of a robot’s inputs into meaning representations of the physical world and actions that we want a robot to take, goals that we want to specify, constraints that we want them to satisfy, or facts about the world that we want to convey. The next four paragraphs discuss issues related to these four capabilities.

For robust spoken language communication with robots, the robot must first perceive the human or humans that are seeking to interact with it. Software supporting speech processing extracts sequences of words, along with audio-spectral data such as prosody, from raw audio signals produced by human speech. Components for language understanding convert word sequences into semantic representations that convey the relationship between words in the speech, determine a task intention, and attempt to ground the intention in the robot’s immediate physical context. Dialogue processing determines if the robot has produced an initial understanding of what was said. For example, if the speech is too noisy, it could initiate a clarification (e.g., request the user to repeat). For these processes to work in real time and support fluid human communication dynamics, they should be running incrementally as data is fed into the robot’s input streams. Ultimately, the robot’s sensing of humans will need to be distilled into one or more intents that it can recognize.

Intents observed by the robot contribute to how the human(s) and robot are to make decisions. These will often be joint, in the sense that achieving common ground will require coordination from both entities. Dialogue planning policies dictate how a back-and-forth dialogue would play out given the robot’s observation of an intent from a human, a notion of the current context (physical, task-oriented, audio, and otherwise), and any shared history. While some actions will be verbal (such as replies or requests for clarification), some require broader task planning so the robot can perform physical actions in its surroundings. Dialogue planning also accepts information from the downstream robot actions to report on successes, failures, and errors to humans when appropriate.

Human intents must be encoded as goals (with corresponding constraints) so that they can be mapped to robot actions. One such process, symbol grounding, attempts to associate intents to the robot’s environment. The robot’s actions are determined by the adaptive behaviors it can do (i.e., behaviors that change based on the context, such as performing an exploration task or following another moving agent). Motion planning dictates the precise movements the robot makes based on what it senses from world modeling. Based on the current context or in reply to a human, response generation allows the robot to verbally or non-verbally provide output to complete the bi-directional communication.

Without an environment model, the robot would not be able to meaningfully refer to task-relevant objects in the physical world. For a robot to contextualize speech from a human, it must first understand where it is in the world, which is achieved by localization and mapping (e.g., SLAM). Computer vision technologies from object detection and recognition contribute to a robot’s ability to perform 3D object detection. Finally, a broader environment model is derived from semantic mapping that accepts observations from a set of processes representing semantic classification.

It is not the intent of Figure 3 to serve as a definitive and conclusive answer for how to design a robot intelligence architecture capable of performing bi-directional communication with humans through speech, but rather it serves as a representative framework that situates many important problems in spoken language interaction with robots. Many subtleties, across the range of lower-level processes in audio, speech, and robotics processing, are omitted for clarity. From this diagram we can also see how high error rates in speech and audio processing or deficiencies in how dialogue is tracked over time could significantly impair how a robot reasons about the world and selects actions to perform various tasks.

**Common Ground** All four capabilities described above contribute to building a common ground representation. All signals interpreted by the robot contribute to its building of a set of inferred mutual beliefs with one or more humans. Meanwhile, the robot must also build a notion of its physical context, which consumes non-linguistic auditory, visual, and tactile inputs into observations

to infer distributions about metric and semantic models of the environment.

**Recommendation I2:** *Update the representation of a robot’s physical surroundings continuously.*

Successful interactions between humans and robots will require not only robust speech processing, but also the ability to quickly and accurately contextualize spoken language dynamically with interpretations from the robot’s immediate physical context. These interactions will be open world in that they will take place in unconstrained environments and rely on continuous properties like space and time [14]. This demands an ability to quickly assess multiple sources of uncertainty that will inform a robot’s ability to perform actions. Tasks that have proven successful in open world contexts include learning from demonstration [5] and interactive task learning for robots [25], though spoken language has not traditionally been the primary source of communication.

Open world interactions form the robot’s physically situated context [63]: (1) perception, which includes sensor readings from visual and occupancy-based sensors, (2) knowledge about pre-defined or recently acquired knowledge about the physical world or relations within it, (3) spatial reasoning about relationships between the robot, nearby objects, and other agents, and (4) temporal reasoning about how things in the robot’s world change over time. Although speech recognition will often be an inadequate source of information [71], these other sources will provide crucial insight on possible misunderstandings and therefore require the development of novel representations that support continuous reasoning in the physical world. Compounding this problem is the diversity of sensors: they range from microphones (both headset-based and array-based) to a robot’s onboard cameras, depth sensors, laser scanners, haptic and proprioceptive sensors, and other peripherals. Ensemble machine learning methods and representations will likely be required for robots to process speech and other sensors while reasoning in the physical world.

## 5 Audio and Speech Processing, Speech Recognition, and Behavior Signal Analysis

Roboticians looking to exploit speech recognition today face numerous challenges. While everyone is familiar with high-performing speech recognition, approaching or exceeding human performance in some cases, making use of the technology for robots is very hard. As noted above in Section 4, speech researchers must have access to a robot (or virtual robot), and be willing to develop the knowledge required to work with robot platforms. This section recommends the development of resources, speech-related technologies and the use of paralinguistic information to aid in the development of a speech recognition system specifically for robots.

**Recommendation A1:** *Develop the following shared resources: general toolkits for front-end audio processing, a database of robot-directed speech, and a challenge task on speech recognition for robot-directed speech.*

The first set of challenges speech for robotics faces relates to the acoustic environments where robots operate, which are often very complex, as suggested by Figure 4. Even for a robot in a quiet, non-reverberant room with only a single user present, the sounds of the robot’s own motors and speech will be present. Multiple speakers bring a further challenge. Thus, prerequisite to speech recognition, there is a need to separate out the speech signal(s) from the confounding contributions to the audio signal. Aspects of this problem have been well-researched, with various techniques for source separation, both in software such as single-channel speech segregation [58, 69] and using hardware, such as microphone arrays (e.g., [29]). Smart speakers, for example, do quite well even in complex environments, thanks to intensively-tuned algorithms. However, reusable general toolkits

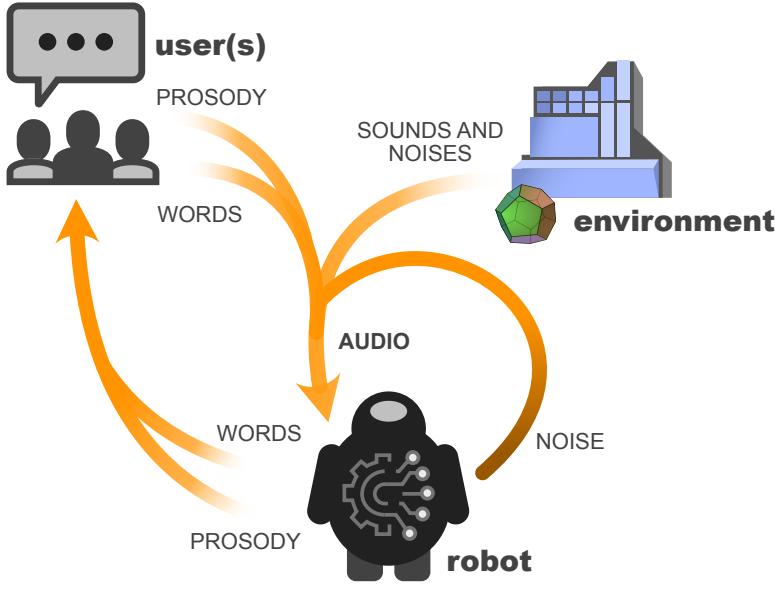


Figure 4: A robot’s audio input will be a mixture of lexical and prosodic information from the user, plus meaningful sounds from the environment and the robot’s own speech output, plus noise from the environment and its own motors. Disentangling these is a huge challenge.

appear to not yet exist. These should, ideally, be pre-trained on massive datasets, but easily and robustly adaptable to specific contexts of use. In addition, current techniques will need to adjust to wheeled or bipedal robots that are moving around in their environment.

In the case of multi-human, multi-robot communication, a robust speaker diarization system will be needed to partition the speech signal into homogenous segments corresponding to each speaker. While such partitioning is not necessarily needed before ASR is performed, it will be needed in order for the robot to understand what was said, and who said it. The degree of difficulty of this task will depend on whether a supervised approach that allows for the system to be trained on speech from each speaker is possible or not. There are several speaker diarization tools available including the open source tools available with pyAudioAnalysis [46] and Kaldi, and also others that are meant to work in conjunction with specific ASR systems such as LIUM, which integrates well with Sphinx, Microsoft’s VideoIndexer which includes speaker diarization and ASR among other tools, and Google’s joint speech recognition and speaker diarization system [37].

One issue with cloud-based, pre-trained models like Google’s speech-to-text system and Microsoft’s Azure is latency and the need for an internet connection. Because the recognizers are cloud-based and not local to the robot, the latency in obtaining the recognition results is often too slow for practical and natural interaction with robots. Local speech recognizers can process faster, but the tradeoff is that locally trained models don’t have as large of a vocabulary as the cloud-based models. Moreover, cloud-based recognizers are invariably trained using data sets and objective functions that are quite unlike those needed for robots. Naturally, the performance obtained on speech to robots is much poorer. Issues include most centrally the sorts of things that people tend to say to robots, and how they say them [73].

To balance these tradeoffs, we suggest the development of, for example, a thousand-hour dataset of human-robot speech to enable training models with existing tools, and the evaluation of new techniques. Realistically, no single dataset could handle all the types of speech and situations needed for robotics, so any such corpus would need to be diverse, across multiple dimensions: speech directed to both humanoid and other mobile robots; in office, warehouse- or airport-sized spaces, and vari-

ous outdoor environments; for a variety of tasks; for various user demographics; and for a variety of microphones including headset, on-board, and microphone-array.

**Recommendation A2:** *Better exploit context and expectations in speech recognition.*

Speech directed to robots will always be a challenge to recognize, but in partial compensation, the context can be expected to be highly informative. For example, if a robot has just started to move, the probability of hearing words like *stop*, *wait*, or *no* will increase, and the probability of hearing words like *pick*, *lift*, and *explain* will decrease. In other words, the robot can use its interpretation of the environment, task plan, and available actions to bias its language and speech understanding. Among the many ways to do this, perhaps the most practical is dynamic language modeling, ideally involving a model able to map from the entire context to a probability distribution over all the words in the vocabulary, continuously updated. Of course, speech recognizers need APIs that support this.

**Recommendation A3:** *Consider creating a speech recognition system focused on the issues encountered in robotics (e.g., speech segregation and speech in noise where the noise may come from the robot and/or the surrounding environment, possibly while the robot is moving).*

While many speech recognizers exist, none are robot-ready. In practice, high error rates remain a key limitation for creating social robots. While we expect that existing engines can be extended and adapted to work well for robotics, it is also worth considering a branch of an existing open source recognizer to create one specialized for robotics. Reiterating some points made earlier, this should be robust to high amounts of noise and multiple speakers, robust to spontaneous and fragmentary utterances, retrainable to perform well in narrow domains, incremental and fast, time-aware, and designed to work well with other components for environment tracking, prosody processing, multi-modal input, and realtime output.

**Recommendation A4:** *Better exploit prosody, emotion, and mental state.*

Speech includes both words and prosody. Speech recognizers handle only the former, meaning that much of the information in the speech signal is discarded. Prosody comprises features of the speech input that are not governed by the phoneme sequences of the words said. These include features of pitch, energy, rate, and voicing. In many applications the lack of prosodic information is not an issue: if the user wants to set an alarm or to get today's weather, it's enough to detect the words, without worrying about whether the user is confused, preoccupied, distressed, unsure, or about how the utterance relates to the user's goals or the temporal context. Yet for robots, all these aspects, and many more, can be critical. In some cases, the prosody can matter more than the words: an *oops* can flag an embarrassing little mistake that can be ignored or a major surprise that requires everything to be replanned, and only the prosody may indicate the difference. Configurations of prosodic features convey information of three main kinds: the paralinguistic, conveying user traits and states, the phonological, relating to the lexical and syntactic components of the message, and the pragmatic, relating to turn taking, topic structure, stance, and intention. Today it is easy to compute many prosodic features, and, given enough training data, to build classifiers for any specific decision. However, we would like tools that can extract prosodic information in real time and provide a continuous read-out of the results, in terms that are directly useful as input for robot task planning.

In addition to recognizing what the user has said, an ultimate goal will be for robots to be able to use speech and language as behavioral signals to assess the user's current state: mental, physical, and otherwise. There is considerable ongoing research in emotion recognition, sentiment analysis and mental health condition based on unimodal and multimodal systems using speech, language, facial gestures, and various biological signals (e.g., [33]). Having robots take advantage of behavioral

signal processing can be crucial in determining how to best be of assistance. For example, when helping a child read, if the robot is able to ascertain that the child is getting frustrated, then it can change its approach, and continue to do so until it determines what will be most effective in helping the child learn. As another example, a robot that is participating in elderly care might alert family members or a doctor if it is able to recognize negative symptoms in emotional state that could suggest the onset of depression.

**Recommendation A5:** *Use audio scene and event analysis to better understand the environment.*

Finally, not only is high-performing speech recognition needed and the determination of mental state desirable, robots also need to understand the sounds in their environment, which provide contextual information that can be crucial to their performance. As an example, consider a robot assisting first responders in a disaster relief setting. The robot should be able to sense sounds from vehicles and bystanders that could pose a danger to relief efforts and notify human teammates.

## 6 Language Understanding

Language understanding focuses on interpreting communications from human to robot into a machineusable representation. While intelligent personal assistants primarily focus on deriving intent from a user’s speech to determine a system request, robots must do more. Specifically, they must perform grounded (or situated) language understanding, where success depends on correctly interpreting language that refers to, or has meaning only in the context of, some sensed physical environment. In this section, we describe current approaches to grounded language understanding and offer recommendations that leverage this context to improve spoken language interaction with robots.

Grounded language understanding is often necessary for robots to follow human-provided instructions. The task can be framed as the problem of interpreting a linguistic structure, possibly in the context of a perceived or *a priori* known environment model, to produce a representation that is meaningful to a robot. In robots, interpretations drawn from language understanding can then be shared with a dialogue manager to determine actions, goals, or constraints to mission planners, motion planners, or planning and scheduling modules. Meanwhile, a language generation module can communicate a statement or question back to the human interacting with the robot. The task of grounded language acquisition goes hand-in-hand with understanding, where the agent learns a correspondence between some interpreted natural language text and the environment with the goal of fulfilling one or more tasks. In practice, given the complex physical world in which robots operate, such models need to be learned *in situ*, rather than predefined for all language understanding tasks.

Generally speaking, grounded language understanding seeks to infer a representation that associates natural language to a robot’s environment model as perceived through sensing (e.g., vision, sound, haptics, etc. [107]). This environment model may contain spatial and semantic information that can be used to resolve the meaning of utterances where resulting interpretations relate in a meaningful way to physical objects in the environment that humans and robots share. The ability to understand expressions that relate to or inform a robot’s model of the environment is critical for collaborative robots. Such robots may be required to follow sequences of navigation [4, 27, 70] or manipulation [25, 89] instructions using spatial terms [88] or report back on observations in areas beyond a human teammate’s line of sight. Several representations have proven effective, such as Generalized Grounding Graphs [104, 105], Distributed Correspondence Graphs [56], Dynamic Grounding Graphs [87], linear temporal logic propositions [19, 20], combinatory categorial grammars [6, 75, 108], and neural networks [13, 76, 79].

There are a number of open problems in grounded language understanding. As the recent lit-

erature has shown, the field has made significant progress in the ability to efficiently infer accurate representations using machine learning [106]. Nevertheless, gaps still remain in our ability to (1) track user intention in spoken dialogue, (2) resolve complex spatial and temporal relationships between referents (i.e., objects and regions) in the physical environment, and (3) represent the world to most efficiently interpret the meaning of an instruction. Most evaluations have been conducted in simulated environments with clean and often well-formed written language inputs. An inherent limitation to these methods is that they are not necessarily able to derive user intention from the shorter, less formally structured content in spoken language. However, the back-and-forth nature of these communications can be exploited to make language understanding more robust.

**Recommendation L1:** *Develop language understanding models for robots that resolve referential and other ambiguities in spoken dialogue.*

Not only will humans and robots be speaking to each other, they will be co-present in a shared environment. A touchstone task to assess progress can be to resolve referents in spoken dialogue to accomplish tasks. This is because it satisfies the following three core requirements: an interaction that uses speech, grounding language to the environment, and bi-directionality. This challenge is non-trivial because humans and robots have mismatched levels of knowledge and abilities in recognizing speech, understanding language, perceiving the environment, and reasoning about goals, actions, and plans. To mediate perceptual differences, grounded language understanding will need to go beyond single utterances and rely on collaborative discourse [67]. If more efforts focus on resolving referential ambiguities in dialogue, we can identify coping strategies that mitigate this fundamental difference. Another key challenge of grounded language understanding in spoken dialogue involves the timing demands: accurately inferring the meaning of an instruction in a partially known or highly noisy representation of the environment in real time. Efficiency is particularly important because without realtime language understanding algorithms (as described in Section 3), speech and language-based interaction becomes impractical.

**Recommendation L2:** *Develop methods to infer and represent more information about humans from robot sensors.*

For many robots today, a user is treated as just a disembodied source of speech input, or at best, an approximate image region with an estimated location and velocity. Sometimes such a simplistic representation is appropriate (such as navigating in a space away from humans) and other times this representation is sorely inadequate (like social interaction with a hospital patient). In the latter case robots need to know more, both to properly understand what the user is saying and to appropriately speak to them. A robot should be able to infer what the user is doing, what they are paying attention to, what their posture suggests they are likely to do next, aspects of their energy level, current cognitive load level, and so on. Doing this reliably requires advances in both speech and robotic perception.

## 7 Dialogue and Human Communication Dynamics

Dialogue can be broadly defined as interactive communication (speech, gesture, gaze, etc.) between two or more interlocutors [40]. We wish to enable robots to converse with people, providing a coherent, flowing, natural, and efficient user experience, but this requires far more than just incorporating the core language components. Realistic androids like ERICA [49] have shown the promise of immersive realtime dialogue with robots, but there are still many active areas of research. This section describes the need for realtime interactivity in robot-centric dialogue processing and identifies some key challenges.

**Recommendation D1:** *Focus on highly interactive dialogue.*

Dialogue allows a robot to interpret a user’s speech and prompt for clarification if something is unclear. At the same time, the robot can provide verbal status updates of its progress on a task. For robots to build common ground with users, they will need to coordinate joint activities [9], with speech playing a central role. Speech provides a direct signal to a robot that a person is instructing it, and the acoustic interpretation can serve as a source of information for managing dialogue and making decisions.

A critical capability for robots will be the ability to ask for help or clarify when something is unclear. Such decision-making will require determining the level of uncertainty about a robot’s sources of evidence when performing a task. Such sources are not limited to only the human-robot interaction itself, but also the physical context, including objects in the immediate physical surroundings and their properties, and actions and their immediate feasibility.

**Improving Dialogue with Recovery Strategies** With dialogue, a robot can mitigate failures to perfectly understand language by asking for clarification using recovery strategies to get the conversation back on track. This can be achieved in multiple levels of human-robot communication, whether they are related to the speech that was uttered [96] or to the broader physical context [15]. Clarifications can be broadly categorized as responding to non-understandings and misunderstandings [53]. Non-understandings mean the robot would have no clear interpretation of what was said by a human, while with misunderstandings, the robot would arrive at an incorrect interpretation, though it could provide evidence of its reasoning if needed. While a great deal of recovery strategies exist in the dialogue literature, few are designed for human-robot communication.

**Recommendation D2:** *Explore the broad space of recovery strategies in spoken language interaction with robots, including when and how.*

While there exists a well-established list of recovery strategies for speech recognition (e.g., clarifying a word or phrase, or asking the user to rephrase their instruction [99]), only a handful exist for clarifying physical context in a dialogue, such as resolving ambiguous or impossible-to-execute instructions [71, 108]. For robots, the space of recovery strategies expands to include not only speech but also actions involving gesture and movement. Early signaling of the need for recovery becomes possible, for example, with a raised eyebrow or a sudden slowing of motion.

**Human Communication Dynamics in Realtime Interaction** Realtime social interaction is a uniquely human ability. For many human abilities, AI is on track to approximate or exceed human performance, but that is not the case with human communication dynamics. To reap the full benefit of other advances, robots need to be able to work effectively with humans, which requires research in this area. Realtime interaction is essential in joint-task situations where time is of the essence, but also has more general value. It is, indeed, something that people often seek out. Texting and emails have their place, but if we want to get to know someone, negotiate plans, make lasting decisions, get useful advice, resolve a workplace issue, or have fun together, we usually seek a real-time spoken interaction. For robots to be widely useful and widely accepted, they similarly need to master real-time interaction.

However, this is currently beyond the state of the art. To quote from [112], given the broad acceptance of systems like Siri and Alexa, one might imagine the problems of interaction are solved. But this is an illusion: in fact, these systems rely on numerous clever ways of avoiding true interaction. Their preferred style is to simply map one user input to one system output, and they employ all sorts of stagecraft to guide users into following a rigid interaction style. Thus today most interactive systems require tightly controlled user behavior. The constraints are often implicit, relying on ways to set up expectation and hints that lead the user to perform only a very limited set of behaviors [32] to follow the intended track. Such constraints greatly simplify design and reduce the likelihood

of failures due to unplanned-for inputs. However, designing around narrow tracks of interaction has led system builders to adopt impoverished models of interactive behavior, useful only for very circumscribed scenarios.

In the research arena, researchers have shown how we can do better, producing prototype systems with amazing responsiveness [35, 51]. For example, some researchers demonstrated a system that could pick up on subtle indications of a student’s cognitive state and respond in ways that increased learning gains [41] and others demonstrated how a robot in dialogue could dynamically shape the user’s attention [117]. These illustrate that elements of true realtime interaction are possible. Moreover these are often highly valued: experiments have shown that users interacting with systems (or people) with better interaction skills trust them more, like them more, and use them more [42].

In general, we envision the creation of highly interactive systems. Imagine that you’re moving heavy furniture, performing surgery, or cooking with the aid of a robot. You would want it to be alert, aware, and good at coordinating actions, and this would require competent realtime interaction. Borrowing again the words of [112], we think that in broad strokes, these systems will be characterized by low latency and natural timing, a deft sensitivity to the multi-functional nature of communication, and flexibility about how any given interaction unfolds. Their skill with interaction timing will be manifest in the way they are attuned to and continuously respond to their users with an array of realtime communicative signals.

**Recommendation D3:** *Work to elucidate the fundamental questions in realtime social interaction, both scientific and engineering.*

As our AI systems grow more capable, the need for realtime social interaction competence will only increase. Existing techniques are limited. Some involve custom datasets, careful policy design, and intense engineering and tuning, and these do not scale. Others model only single dimensions of interaction [101]. At the same time, current deep learning models, though they have worked so well in many areas of AI, are not directly applicable to realtime, situated interaction.

## 8 Language Generation and Speech Synthesis

So far our focus has largely been on communication from human to robot, but robots also need to speak to humans. Speech synthesizers were, historically, designed to create an audio signal to intelligibly encode any given sentence. The target was read speech, in a neutral tone. More recently, synthesizers have become able to produce speech that is not only intelligible but also highly natural, and even expressive in some ways. While adequate for some purposes, robots often need more.

**Recommendation S1:** *Extend the pragmatic repertoire of speech synthesizers.*

Robots operate in real time and real space. Speech synthesis in this context needs access to the full expressive power of spoken language. For example, consider the use of language to direct attention (*hey look!*), convey uncertainty (*the red one?*), establish priorities (*help!*), or coordinate action (*ready ... go!*). With appropriate timing, voicing and prosody, such phrases can be powerfully effective; without this, users may be confused or slow to respond. Or, for example, imagine a robot prefacing its next movement with *okay, over behind that truck*. Beyond the words, a cooperative utterance may also convey the robot’s view of the likely difficulty of moving behind the truck and its desire for follow-on information about what it should do once it gets behind the truck. Robot speech thus needs to be able not only to convey propositions and speech acts, but to be able to simultaneously convey nuances of information state, dialogue state, and stance. Enabling robots to do such things requires advances of several kinds.

To produce such richly informative outputs, speech synthesizers need rich input: far more than just sequences of words. For the above examples, effective speech synthesis would also need access to information from the user model, environment and plan representations, and mission context representation. Current software architectures for robots generally do not expose such information: it may be buried down in some component-specific internal data structures or parameter values. To support adequately expressive synthesis, we will need new ways to explicitly represent and expose more of a robot’s instantaneous internal state.

Beyond issues of speech synthesis, effective communication also requires appropriate choice of words. For some applications, a robot may need to produce only one of a finite set of sentences, or use only a finite set of templates. However, effective language generation remains a challenging problem [44], especially for robots, for reasons already noted. The term “speech generation” is sometimes used to indicate that speech synthesis and language generation are essentially one, tightly integrated problem. Today a pipelined approach, with two separate modules — concept-to-text and text-to-speech (TTS) — is the norm, but this is problematic [22], as the language and speech decisions are often interdependent. In some cases the interdependencies may be mild, for example, for TTS systems used to produce good quality isolated sentences based on read speech example. Robots, however, require more expressive power. Today, end-to-end training may in principle solve this problem, but in practice, the limited data associated with many human-robot interaction scenarios will make this problematic. Instead, researchers will need to pursue loosely coupled language and speech generation, where the generated “language” comprises both text and control signals for speech synthesizers. Of course, this means that speech synthesis systems must be designed to allow such control. These control signals will need not only to support wide expressive abilities, but also the needs of a robot that performs actions in time and space.

The nature of these control signals is a question in itself. A particular challenge is that of appropriate prosodic control signals. Clearly the use of punctuation marks is not enough. For example, an exclamation point can indicate emphatic agreement (*exactly!*), enthusiasm (*let’s go!*) or urgency (*help!*). Also, while an exclamation point can accurately indicate emphasis in a short phrase (*over here!*), it would not be useful in a sentence where meaning can differ depending on the emphasis location, as in *this* versus *today* in: *We need to use this one today!* It may be possible to learn from data an appropriate set of prosodic control signals [110], but it is not clear how “style tokens” or other methods for representing tone in simple applications — like audiobook synthesis and emulating acted emotions — can be extended to support speech adequate for the here-and-now communicative functions [111] that robots most need. More work specifically targeting the speaking needs of robots, perhaps involving adaptation methods and multi-domain training, is needed.

Audience design is another major issue: robots need to produce speech that is not only clear and correct, but understandable. For example, if the robot recognizes an object as an *orange* but the human cannot see the object due to view occlusion, a simple referring expression, such as *next to the orange at the corner* will not be understandable. Robust models need to consider the human’s perspective and knowledge state. In general, the goal is not to minimize the speaker’s own effort, but rather to minimize the joint effort to come to a common ground [31]. Thus a robot should often make the extra effort to make sure the human understands. This may involve, for example, generating a description in small pieces, giving the human the chance to give interleaved feedback to verify that the knowledge states are aligning [39], or by proactively first describing essentials of its own internal representation, to make subsequent grounding more efficient [24]. Future language generation and speech synthesis modules will need more systematic techniques for applying theory-of-mind reasoning to model humans’ mental models and perspectives, and methods for collaborative grounding. Further, in an environment containing multiple human agents, a robot needs to design its utterances to make them clear to specific individuals or groups of individuals, and to craft them to make clear at each time who its utterances are addressing.

**Recommendation S2:** *Develop speech generators that support multimodal interaction.*

Robots are embodied and multimodal. To be effective, actions in the linguistic channel must be coordinated with other channels, such as physical gestures and eye gaze. This involves not only selection of appropriate word sequences but also utterance prosody. This is especially important for robots that need to be able to refer to specific objects in the environment, and need to show ongoing awareness of the environment as things change.

Fortunately, many robots have physical attributes that enable them to communicate more efficiently. For example, they often have capabilities for gestures and postures or gaze to show direction of attention. Generated language should incorporate deictic expressions and be coordinated with the timing of a robot's physical gestures. Robots with capabilities for facial expressions need to coordinate those with the timing of prosodic emphasis or intonational cues associated with a question. In addition, language should be coordinated with path planning and motion planning, as when a robot needs to convey that it is about to move *over here*. Robots that have these abilities will be able to communicate more efficiently, often using just a few words deftly augmented with multimodal and prosodic signals. Indeed, robot-to-human information transfer may evolve from being a sequence of individual communicative actions to something more continuous: an ongoing display of state and intention.

**Recommendation S3:** *Create synthesizers that support realtime control of the voice.*

Robots operate in real time, so synthesizers must also. There are several aspects to this. As robots must respond to dynamic changes in the environment, the generation and speaking processes need to be interruptible and the plans must be modifiable. For example, human speakers reflexively pause if a loud noise occurs in the environment, or if an addressee seems to be not paying attention, and robots should do the same. To coordinate spoken language with a robot's physical gestures and motion, synthesizers must need to be able to output synch points and to support fine-grained timing control. Moreover, a robot's speech may need to be timed to support, guide, or complement the user's actions and utterances. Incremental synthesis is also commonly needed. Timing is thus a major issue, with progress needed on many fronts.

**Recommendation S4:** *Develop the ability to tune speech generators to convey a desired tone, personality, and identity.*

One can take the perspective that robots should be purely functional, and that designers should not bother to produce robots that project a specific personality. However, designing a robot to have no detectable personality is itself a design choice. It is not uncommon today to hear robots with voices chosen only on the basis of intelligibility, and this guides the user to expect a formal, tedious interaction partner. Many highly capable agent systems, such as Siri and Alexa, have clear, dominant voices to convey to the user that they should adopt a formal turn-taking style and keep their utterances short and to the point. Yet this strategy is not always appropriate; for example, a robot interacting with a small child should talk very differently, and a robot assisting in a disaster recovery effort should sound different again.

In general, the voice of an artificial agent, together with its visual appearance, tells the user what to expect of it. Thus we need voices that are parameterizable — to be a little more childlike, more rigid, more helpless, more businesslike, and so on — to meet the needs of an application, and guidelines for making such choices. Further, although contemporary speech synthesis is capable of generating utterances virtually indistinguishable from those produced by a human being, this will often be inappropriate or even unethical: the use of human-like voices for artificial devices encourages people to overestimate their linguistic and cognitive capabilities. In many cases, the voice should make the agent clearly identifiable as an artificial individual: a “robotic” voice can be natural for a

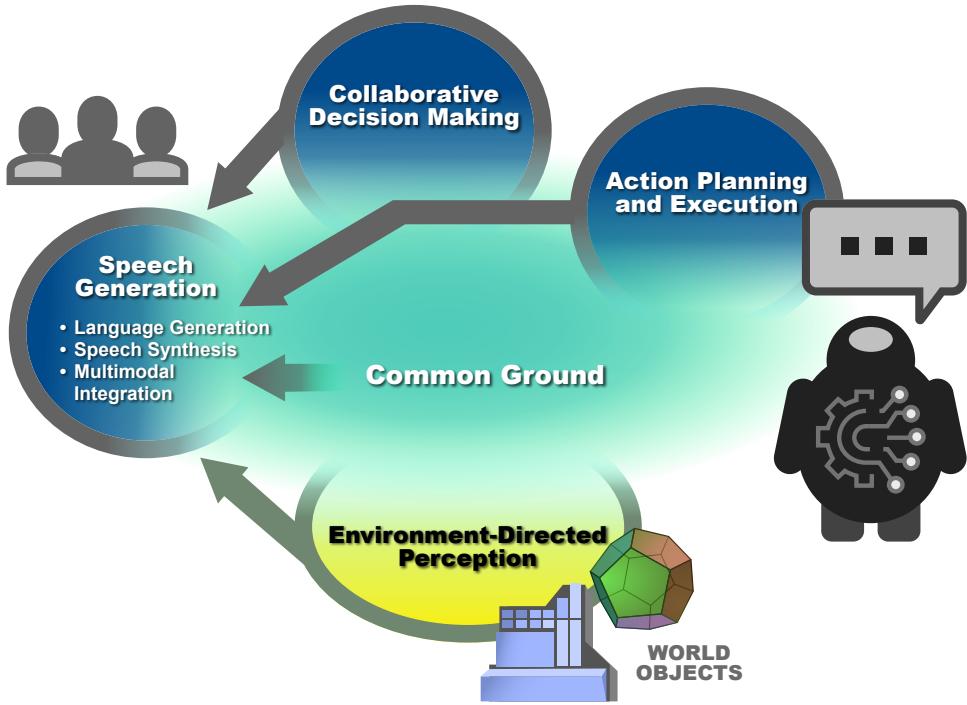


Figure 5: Speech generation in context, derived from Figure 3. There the focus was on human-to-robot communication; here, despite the different direction of information flow, many of the same knowledge sources and representations are relevant.

robot; yet such a voice need not be a low-quality voice, it being perfectly possible to generate high-quality robotic-sounding speech [113]. Beyond just the qualities of the voice itself, variation in style more generally, including word choice and interaction style, can also guide user expectations, but so far has been very understudied.

Figure 5 illustrates some of the points made above. Speech Generation needs to be able to convey not only propositions and intents (top arrow), but also other aspects of the robot's internal state, stance, and knowledge. Speech Generation also needs to be appropriately timed and coordinated with robot actions and multimodal signaling (middle arrow). To do so effectively, it needs access to the robot's models of what the user knows, and of the current state of the environment (bottom arrows).

## 9 Policy

Our recommendations to this point have been directed to researchers and developers: scientific and technical challenges call for scientific and technical advances. But to enable these advances, a healthy research ecosystem is needed, and this also requires supportive decisions by funders and policy makers.

**Recommendation P1:** *Create funding opportunities specifically for spoken language interaction with robots.*

None of the issues noted in this report are entirely new: all have been discussed before. Yet they remain as unsolved issues today, in part because they have tended to fall through the cracks. Few of

the challenges we have identified are core robotics topics, and few are pure speech science topics. In competitions for funding, work on these issues may seem only marginally appealing to pure robotics programs, and also marginal to pure speech science programs. The challenges and needs in spoken language interaction with robots are specific enough and non-mainstream enough that funders need to use special care to nurture this area. While there are bottom-up efforts to foster research in this area, such as the recent RoboDial workshop, top-down support is also critical.

Spoken language interaction with robots is a highly interdisciplinary area, requiring researchers and developers who are competent in multiple disciplines — speech, language, vision, robotics, machine learning, etc. — and able to collaborate broadly. Great benefit could result from the creation of a multi-disciplinary curriculum or consortium to train the next generation of researchers and developers. These might include summer schools or summer camps (perhaps like the human language technology summer programs organized by Johns Hopkins) dedicated to the intersection of speech, language, vision, and robotics. We also see the need for more fellowship and internship programs, some perhaps joint public-private efforts, to encourage Ph.D. students to do multi-disciplinary research that connects spoken language and robotics.

At the same time, care must be taken to ensure that work in this area does not become inward-looking, instead remaining well-connected to other work in speech, language, and robotics.

**Recommendation P2:** *Prefer evaluation based on use cases.*

There is an essential tension between intrinsic and extrinsic evaluation. Today in most areas of speech processing the former is pervasive: researchers commonly tackle an existing dataset and develop a new algorithm that improves on previous results according to a standard metric. Yet ultimately we need to evaluate research extrinsically, judged by its contribution towards providing useful communicative capabilities to users. Doing so brings greater likelihood of leading to novel results and perspectives, and of driving real progress. However, extrinsic evaluation is much more time-consuming and expensive. This is true especially for interaction, as the evaluation of interactive behaviors cannot really be done by reference to static data sets. Even for the best-understood aspects of extrinsic evaluation, relating to user satisfaction, meaningful measurement is difficult, and the results depend on so many factors [80] that the generality can always be questioned. Still, we hope that funders and proposal reviewers will prefer extrinsic evaluation, and, at the same time, researchers will work to address the need for new evaluation methodologies that are both efficient and highly informative.

**Recommendation P3:** *Support many kinds of research and development activities.*

Not everything we need to advance this area can be formulated as a classic research project. Some things are pure infrastructure development, including the development of shareable data, software, and hardware. Others, like dealing with reverberation, are now engineering problems more than scientific problems. We encourage funding agencies to look at the big picture, and support, when appropriate, any activity that is critical to progress. At the same time, we recognize that federal funding is not the answer to every question. The field needs to figure out how industry, including start-ups, can carve off important problems, solve them, enable those solutions to be widely used, and make a profit from doing so.

One important type of activity to value is truly focused research. Research solicitations naturally reflect funders' multitudinous ambitions and desires, and, as a result, they increasingly call for every research proposal to address all of these at once — from the creation of data resources and software infrastructure, to demo-ing on real robots and addressing true scientific questions, to doing outreach and technology transfer — as suggested by Figure 6. Yet focused work is also important: advances in only one or two areas can still be highly significant. For example, a research project to answer core

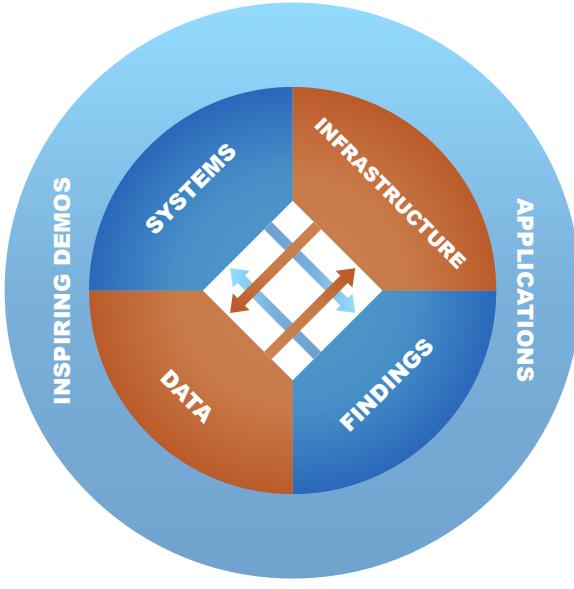


Figure 6: Some of the many facets of speech for robotics. Funders should recognize this diversity of needs and the diversity of methods they require, and be able to support projects that focus on only one aspect or approach.

questions in this area may not need to involve real robots. Similarly, research on software architectures that make robots' internal representations more useful for speech-based communication may not need to promise advances in any core speech or robotics technology.

**Recommendation P4:** *Work to overcome the barriers to data sharing.*

Data is the lifeblood of research in spoken language interaction with robots. In particular, recordings of real humans interacting with real or simulated robots can be used for analysis, discovery, and model training. Unfortunately, today almost all such data is trapped within individual institutions, barricaded by restrictions that prevent sharing. Many of these restrictions exist for good reasons, but we still need to work to find ways to share data that are respectful of privacy concerns. While we see no simple solutions, we do see two possible ways to address this problem. First, researchers should, whenever possible, design data collections to be fully shareable, which may in some cases be as simple as having participants dedicate their "work" to the public domain, for example using the Creative Commons CC0 license. Second, researchers should try to educate their local Institutional Review Boards (IRBs) on both the costs and benefits of tight restrictions on data sharing. To this end, the creation of a model IRB proposal that shows how to address the issues and concerns could be very helpful, especially if endorsed by the NSF or another prestigious organization.

**Recommendation P5:** *Explore novel public-private partnerships for open source software.*

Open source software is enormously valuable for both research and development, but in this area we see a missed opportunity. Traditionally, open source software comes from universities, but industry has a great deal of valuable software that might be released. While some companies aim mostly to protect intellectual property, others are potentially willing to open-source some or all of their products. Examples include Willow Garage — and now the Open Source Robotics Foundation — which provided Robot Operating System (ROS) free and open-source, as well as the PR2 as a reference robot hardware system, both of which catalyzed research in autonomous robotics. However, releasing clean, well-documented open-source software takes resources and is seldom a direct revenue

generator, so external encouragement and support can be necessary. Thus we see the opportunity for novel private-public partnership to help create and maintain open reference implementations of software and hardware for enabling social robotics. We might call this a “reverse SBIR,” in the sense that a for-profit company and a research institution collaborate, not to commercialize a high-risk/high-reward research project (as with a traditional SBIR), but, rather, to document and publish open-source software implementations and hardware designs of company products. For companies the benefits may include extended product lifetimes and visibility, and for the research community new tools and testbeds for research.

Enabling spoken language interaction with robots will require advances in speech science, in robotics, and at the intersection. There is a lot to do: not just one single problem to solve, but a multifaceted challenge, needing attack from many fronts, over years and decades. Our hope is that this report, by providing a clearer view of the key issues, will help researchers and funders optimally choose what to tackle, and ultimately, after much hard work, bring us to the day when we can interact with robots effectively and smoothly, just by talking with them.

## Acknowledgments

The workshop that led to this report was sponsored by the National Science Foundation, Grant #IIS1941541. We thank Tanya Korelsky for her vision, support, and guidance. We also thank the government observers who attended the workshop and contributed to the discussions: Jonathan Fiscus, Susan G. Hill, Nia Peters, Erion Plaku, Christopher Reardon, and Clare Voss. A hearty thanks also to Erin Zaroukian for her careful reading and thoughtful comments on the writing, Chad M. Smith for illustrating the title page, Evan Jensen for improving the technical figures, and Amber Bennett-Groves for proofreading.

## References

- [1] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.
- [2] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems*, pages 114–130. Springer, 2012.
- [3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Journey, Besmira Nushi, Penny Collison, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-AI interaction. In *Proceedings of CHI*, pages 1–13, 2019.
- [4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of CVPR*, pages 3674–3683, 2018.
- [5] Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.
- [6] Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62, 2013.
- [7] Wilma A. Bainbridge, Justin W. Hart, Elizabeth S. Kim, and Brian Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1):41–52, 2011.
- [8] Bruce Balentine. *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. ICMI Press, 2007.
- [9] Adrian Bangerter and Herbert H. Clark. Navigating joint projects with dialogue. *Cognitive Science*, 27(2):195–225, 2003.
- [10] Timo Baumann and David Schlangen. The InproTK 2012 Release. In *Proceedings of the NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 29–32, 2012.
- [11] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. Recognising conversational speech: What an incremental ASR should do for a dialogue system and how to get there. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots*, pages 421–432. Springer, 2017.
- [12] Philipp Beckerle, Claudio Castellini, and Bigna Lenggenhager. Robotic interfaces for cognitive psychology and embodiment research: A research roadmap. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(2), 2019.
- [13] Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *Proceedings of the Conference on Robot Learning*, pages 1415–1438, 2019.
- [14] Dan Bohus and Eric Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of SIGdial*, pages 225–234, 2009.
- [15] Dan Bohus and Eric Horvitz. On the challenges and opportunities of physically situated dialog. In *Proceedings of the AAAI Fall Symposium on Dialog with Robots*, 2010.

- [16] Dan Bohus, Chit W. Saw, and Eric Horvitz. Directions robot: In-the-wild experiences and lessons learned. In *Proceedings of AAMAS*, pages 637–644, 2014.
- [17] Dan Bohus, Sean Andrist, and Mihai Jalobeanu. Rapid Development of Multimodal Interactive Systems: A Demonstration of Platform for Situated Intelligence. In *Proceedings of ICMI*, pages 493–494, 2017.
- [18] Marilyn G. Boltz. Temporal dimensions of conversational interaction: The role of response latencies and pauses in social impression formation. *Journal of Language and Social Psychology*, 24(2):103–138, 2005.
- [19] Adrian Boteanu, Thomas M. Howard, Jacob Arkin, and Hadas Kress-Gazit. A model for verifiable grounding and execution of complex natural language instructions. In *Proceedings of IROS*, pages 2649–2654, 2016.
- [20] Adrian Boteanu, Jacob Arkin, Siddharth Patki, Thomas M. Howard, and Hadas Kress-Gazit. Robot-initiated specification repair through grounded language interaction. In *Proceedings of the AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.
- [21] Joost Broekens, Marcel Heerink, and Henk Rosendal. Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94–103, 2009.
- [22] Ivan Bulyko and Mari Ostendorf. Joint prosody prediction and unit selection for concatenative speech synthesis. In *Proceedings of ICASSP*, volume 2, pages 781–784, 2001.
- [23] Okko Buß and David Schlangen. DIUM – An incremental dialogue manager that can produce self-corrections. In *Proceedings of SemDial*, 2011.
- [24] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of HRI*, pages 33–40, 2014.
- [25] Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to Action: Towards Interactive Task Learning with Physical Agents. In *Proceedings of IJCAI*, pages 2–9, 2018.
- [26] Shih-Fu Chang, Alex Hauptmann, Louis-Philippe Morency, Sameer Antani, Dick Bulterman, Carlos Busso, Joyce Chai, Julia Hirschberg, Ramesh Jain, Ketan Mayer-Patel, Reuven Meth, Raymond Mooney, Klara Nahrstedt, Shri Narayanan, Prem Natarajan, Sharon Oviatt, Balakrishnan Prabhakaran, Arnold Smeulders, Hari Sundaram, Zhengyou Zhang, and Michelle Zhou. Report of 2017 NSF Workshop on Multimedia Challenges, Opportunities and Research Roadmaps. arXiv preprint arXiv:1908.02308, 2019.
- [27] David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of AAAI*, pages 859–865, 2011.
- [28] Wei Chen and Jack Mostow. A tale of two tasks: Detecting children’s off-task speech in a reading tutor. In *Proceedings of Interspeech*, pages 1621–1624, 2011.
- [29] Amit Chhetri, Philip Hilmes, Trausti Kristjansson, Wai Chu, Mohamed Mansour, Xiaoxue Li, and Xianxian Zhang. Multichannel audio front-end for far-field automatic speech recognition. In *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, pages 1527–1531, 2018.

- [30] Henrik I. Christensen, Tom Batzinger, Kostas Bekris, Karl Bohringer, Joe Bordogna, Gary Bradski, Oliver Brock, Jeff Burnstein, Thomas Fuhlbrigge, Roger Eastman, Aaron Edsinger, Erica Fuchs, Ken Goldberg, Tom Henderson, William Joyner, Lydia Kavaraki, Clint Kelly, Alonzo Kelly, Vijay Kumar, Dinesh Manocha, Andrew McCallum, Pieter Mosterman, Elena Messina, Todd Murphey, Richard Alan Peters, Stuart Shepherd, Sanjiv Singh, Larry Sweet, Jeff Trinkle, Jason Tsai, James Wells, Peter Wurman, Tom Yorio, and Mingjun Zhang. *A Roadmap for US Robotics: From Internet to Robotics*. Computing Community Consortium (CCC), 2009.
- [31] Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.
- [32] Michael H. Cohen, James P. Giangola, and Jennifer Balogh. *Voice User Interface Design*. Addison-Wesley, 2004.
- [33] Nicholas Cummins, Stefan C. Scherer, Jarek Krajewski, Sebastian Schneider, Julien Epps, and Thomas F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.
- [34] Eric Deng, Bilge Mutlu, and Maja J. Matarić. Embodiment in socially interactive robots. *Foundations and Trends® in Robotics*, 7(4):251–356, 2019.
- [35] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of AAMAS*, pages 1061–1068, 2014.
- [36] Laurence Devillers, Tatsuya Kawahara, Roger K. Moore, and Matthias Scheutz. Spoken language interaction with virtual agents and robots (SLIVAR): Towards effective and ethical interaction (Dagstuhl Seminar 20021). In *Dagstuhl Reports*, volume 10. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
- [37] Laurent El Shafey, Hagen Soltau, and Izahak Shafran. Joint speech recognition and speaker diarization via sequence transduction. In *Proceedings of Interspeech*, pages 396–400, 2019.
- [38] Maxine Eskenazi and Tiancheng Zhao. Report from the NSF Future Directions Workshop, Toward user-oriented agents: Research directions and challenges. arXiv preprint arXiv:2006.06026, 2020.
- [39] Rui Fang, Malcolm Doering, and Joyce Y. Chai. Embodied collaborative referring expression generation in situated human-robot dialogue. In *Proceedings of HRI*, pages 271–278, 2015.
- [40] Terrence Fong, Charles Thorpe, and Charles Baur. Collaboration, dialogue, and human-robot interaction. In *Robotics Research*, pages 255–266. Springer, 2003.
- [41] Kate Forbes-Riley and Diane Litman. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9-10):1115–1136, 2011.
- [42] Riccardo Fusaroli and Kristian Tylén. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40(1):145–171, 2016.
- [43] Andre Gaschler, Sören Jentzsch, Manuel Giuliani, Kerstin Huth, Jan de Ruiter, and Alois Knoll. Social behavior recognition using body posture and head pose for human-robot interaction. In *Proceedings of IROS*, pages 2128–2133, 2012.

- [44] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
- [45] Jeroen Geertzen. Exploring age-related conversational interaction. In *Proceedings of SemDial*, pages 42–47, 2015.
- [46] Theodoros Giannakopoulos. pyAudioAnalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015.
- [47] Yolanda Gil and Bart Selman. *A 20-Year Community Roadmap for Artificial Intelligence Research in the US*. Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI), 2019.
- [48] Howard Giles, Anthony Mulac, James J. Bradac, and Patricia Johnson. Speech accommodation theory: The first decade and beyond. In *Annals of the International Communication Association*, volume 10, pages 13–48. Taylor & Francis, 1987.
- [49] Dylan F. Glas, Takashi Minato, Carlos T. Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. Erica: The ERATO intelligent conversational android. In *Proceedings of RO-MAN*, pages 22–29, 2016.
- [50] Rachel Gockley and Maja J Matarić. Encouraging physical therapy compliance with a hands-off mobile robot. In *Proceedings of HRI*, pages 150–155, 2006.
- [51] Jonathan Gratch, Ning Wang, Anna Okhmatovskaia, Francois Lamothe, Mathieu Morales, Rick J. van der Werf, and Louis-Philippe Morency. Can virtual humans be more engaging than real ones? In *Proceedings of HCI*, pages 286–297, 2007.
- [52] John Grothendieck, Allen L. Gorin, and Nash Borges. Social correlates of turn-taking style. *Computer Speech and Language*, 25(4):789–801, 2011.
- [53] Graeme Hirst, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15(3-4): 213–229, 1994.
- [54] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. An end-to-end conversational style matching agent. In *Proceedings of IVA*, pages 111–118, 2019.
- [55] Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the International Conference on Computational Semantics*, pages 206–216, 2015.
- [56] Thomas M. Howard, Stefanie Tellex, and Nicholas Roy. A natural language planner interface for mobile manipulators. In *Proceedings of ICRA*, pages 6652–6659, 2014.
- [57] Kristelle Hudry, Catherine Aldred, Sarah Wigham, Jonathan Green, Kathy Leadbitter, Kathryn Temple, Katherine Barlow, and Helen McConachie. Predictors of parent-child interaction style in dyads with autism. *Research in Developmental Disabilities*, 34(10):3400–3410, 2013.
- [58] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey. Single-channel multi-speaker separation using deep clustering. In *Proceedings of Interspeech*, pages 545–549, 2016.
- [59] Toshihiko Itoh, Norihide Kitaoka, and Ryota Nishimura. Subject experiments on influence of response timing in spoken dialogues. In *Proceedings of Interspeech*, pages 1835–1838, 2009.

- [60] Casey Kennington and David Schlangen. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech and Language*, 41:43–67, 2017.
- [61] Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. rrSDS: Towards a Robot-ready Spoken Dialogue System. In *Proceedings of SIGdial*, pages 132–135, 2020.
- [62] Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. A multimodal in-car dialogue system that tracks the driver’s attention. In *Proceedings of ICMI*, pages 26–33, 2014.
- [63] Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138, 2007.
- [64] Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen Hindriks, Mark Neerincx, Petter Ögren, Tomáš Svoboda, and Rainer Worst. TRADR project: Long-term human-robot teaming for robot assisted disaster response. *KI-Künstliche Intelligenz*, 29(2):193–201, 2015.
- [65] Maria Lazarte. Robots to the rescue! <https://www.iso.org/news/Ref2169.htm>, 2017. Accessed 2020-10-29.
- [66] Rivka Levitan, Agustin Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of NAACL-HLT*, pages 11–19, 2012.
- [67] Changsong Liu and Joyce Y. Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of AAAI*, pages 2288–2294, 2015.
- [68] Ewa Luger and Abigail Sellen. “Like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In *Proceedings of CHI*, pages 5286–5297, 2016.
- [69] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.
- [70] Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of AAAI*, pages 1475–1482, 2006.
- [71] Matthew Marge and Alexander I. Rudnicky. Miscommunication detection and recovery in situated human–robot dialogue. *ACM Transactions on Interactive Intelligent Systems*, 9(1), 2019.
- [72] Matthew Marge, Stephen Nogar, Cory J. Hayes, Stephanie M. Lukin, Jesse Bloecker, Eric Holder, and Clare Voss. A Research Platform for Multi-Robot Dialogue with Humans. In *Proceedings of NAACL (Demonstrations)*, pages 132–137, 2019.
- [73] Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz, and Antonio Roque. Let’s do that first! A comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In *Proceedings of SemDial*, 2020.
- [74] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech and Language*, 33(1):1–24, 2015.

- [75] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Proceedings of ISER*, pages 403–415, 2013.
- [76] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of AAAI*, pages 2772–2778, 2016.
- [77] Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. Mirroring to build trust in digital assistants. In *Proceedings of Interspeech*, pages 4000–4004, 2019.
- [78] Thilo Michael. RETICO: An incremental framework for spoken dialogue systems. In *Proceedings of SIGdial*, pages 49–52, 2020.
- [79] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. *Proceedings of EMNLP*, pages 1004–1015, 2017.
- [80] Sebastian Moller, Klaus-Peter Engelbrecht, Christine Kuhnel, Ina Wechsung, and Benjamin Weiss. A taxonomy of quality of service and quality of experience of multimodal human-machine interaction. In *Proceedings of the International Workshop on Quality of Multimedia Experience*, pages 7–12, 2009.
- [81] Roger K. Moore. From talking and listening robots to intelligent communicative machines. In Judith Markowitz, editor, *Robots That Talk and Listen*, pages 317–335. de Gruyter, 2015.
- [82] Roger K. Moore. Appropriate voices for artefacts: Some key insights. In *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)*, pages 7–11, 2017.
- [83] Roger K. Moore. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots*, pages 281–291. Springer, 2017.
- [84] Masahiro Mori. Bukimi no tani (The uncanny valley). *Energy*, 7:33–35, 1970.
- [85] National Science Foundation. Foundational research in robotics (Solicitation). [https://www.nsf.gov/funding/pgm\\_summ.jsp?pims\\_id=505784](https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505784), 2020. Accessed 2020-09-23.
- [86] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of CHI*, pages 415–422, 1997.
- [87] Jae Sung Park, Biao Jia, Mohit Bansal, and Dinesh Manocha. Efficient generation of motion plans from attribute-based natural language instructions using dynamic constraint mapping. In *Proceedings of ICRA*, pages 6964–6971, 2019.
- [88] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M. Howard. Grounding abstract spatial concepts for language interaction with robots. In *Proceedings of IJCAI*, pages 4929–4933, 2017.
- [89] Rohan Paul, Jacob Arkin, Derya Aksaray, Nicholas Roy, and Thomas M. Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms. *International Journal of Robotics Research*, 37(10):1269–1299, 2018.
- [90] Mike Phillips. Applications of spoken language technology and systems. In *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology (SLT)*, page 7. IEEE, 2006.

- [91] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. Personalized Robot Tutoring using the Assistive Tutor POMDP (AT-POMDP). In *Proceedings of AAAI*, volume 33, pages 8050–8057, 2019.
- [92] Rajesh Ranganath, Dan Jurafsky, and Daniel A. McFarland. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, 27(1):89–115, 2013.
- [93] Michael Reddy. The conduit metaphor — A case of frame conflict in our language about language. *Metaphor and Thought*, 2:285–324, 1979.
- [94] Felicia Roberts and Alexander L. Francis. Identifying a temporal threshold of tolerance for silent gaps after requests. *Journal of the Acoustical Society of America*, 133:471–477, 2013.
- [95] Ruhi Sarikaya. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine*, 34(1):67–81, 2017.
- [96] David Schlangen. Causes and strategies for requesting clarification in dialogue. In *Proceedings of SIGdial*, pages 136–143, 2004.
- [97] David Schlangen and Gabriel Skantze. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse*, 2(1):83–111, 2011.
- [98] Thomas B. Sheridan. Human-robot interaction: Status and challenges. *Human Factors*, 58(4):525–532, 2016.
- [99] Gabriel Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341, 2005.
- [100] Gabriel Skantze, Martin Johansson, and Jonas Beskow. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of ICMI*, pages 67–74, 2015.
- [101] Justin S. Smith, Crystal Chao, and Andrea L. Thomaz. Real-time changes to social dynamics in human-robot turn-taking. In *Proceedings of IROS*, pages 3024–3029, 2015.
- [102] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of NAACL*, pages 2610–2621, 2019.
- [103] Deborah Tannen. *That's Not What I Meant! How Conversational Style Makes or Breaks Relationships*. Ballantine, 1989.
- [104] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4):64–76, 2011.
- [105] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of AAAI*, pages 1507–1514, 2011.
- [106] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.
- [107] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney. Learning multi-modal grounded linguistic semantics by playing “I Spy”. In *Proceedings of IJCAI*, pages 3477–3483, 2016.

- [108] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, 67:327–374, 2020.
- [109] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D’Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.
- [110] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of ICML*, pages 5180–5189, 2018.
- [111] Nigel G. Ward. *Prosodic Patterns in English Conversation*. Cambridge University Press, 2019.
- [112] Nigel G. Ward and David DeVault. Challenges in building highly-interactive dialog systems. *AI Magazine*, 37(4):7–18, 2016.
- [113] Sarah Wilson and Roger K. Moore. Robot, alien and cartoon voices: Implications for speech-enabled systems. In *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR)*, pages 40–44, 2017.
- [114] Travis J. Wiltshire, Daniel Barber, and Stephen M. Fiore. Towards modeling social-cognitive mechanisms in robots to facilitate human-robot teaming. In *Proceedings of HFES*, volume 57, pages 1278–1282, 2013.
- [115] Meldon Wolfgang, Vladimir Lukic, Alison Sander, Joe Martin, and Daniel Küpper. Gaining robotics advantage. The Boston Consulting Group (BCG), 2017.
- [116] Guang-Zhong Yang, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, Vijay Kumar, Marcia McNutt, Robert Merrifield, Bradley J. Nelson, Brian Scassellati, Mariarosaria Taddeo, Russell Taylor, Manuela Veloso, Zhong Lin Wang, and Robert Wood. The grand challenges of science robotics. *Science Robotics*, 3(14), 2018.
- [117] Zhou Yu, Dan Bohus, and Eric Horvitz. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *Proceedings of SIGdial*, pages 402–406, 2015.