Computer Speech and Language 71 (2022) 101255

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Spoken language interaction with robots: Recommendations for future research

Matthew Marge [a,*], Carol Espy-Wilson [b], Nigel G. Ward [c], Abeer Alwan [d], Yoav Artzi [e], Mohit Bansal [f], Gil Blankenship [b], Joyce Chai [g], Hal Daumé III [b], Debadeepta Dey [h], Mary Harper [a], Thomas Howard [i], Casey Kennington [j], Ivana Kruijff-Korbayová [k], Dinesh Manocha [b], Cynthia Matuszek [l], Ross Mead [m], Raymond Mooney [n], Roger K. Moore [o], Mari Ostendorf [p], Heather Pon-Barry [q], Alexander I. Rudnicky [r], Matthias Scheutz [s], Robert St. Amant [a], Tong Sun [t], Stefanie Tellex [u], David Traum [v], Zhou Yu [w]

[a] DEVCOM Army Research Laboratory, United States of America
[b] University of Maryland, College Park, United States of America
[c] University of Texas at El Paso, United States of America
[d] University of California at Los Angeles, United States of America
[e] Cornell University, United States of America
[f] University of North Carolina at Chapel Hill, United States of America
[g] University of Michigan, United States of America
[h] Microsoft, United States of America
[i] University of Rochester, United States of America
[j] Boise State University, United States of America
[k] DFKI, Germany
[l] University of Maryland, Baltimore County, United States of America
[m] Semio, United States of America
[n] University of Texas at Austin, United States of America
[o] University of Sheffield, United Kingdom
[p] University of Washington, United States of America
[q] Mount Holyoke College, United States of America
[r] Carnegie Mellon University, United States of America
[s] Tufts University, United States of America
[t] Adobe, United States of America
[u] Brown University, United States of America
[v] USC Institute for Creative Technologies, United States of America
[w] University of California at Davis, United States of America

ARTICLE INFO

Keywords:
Research agenda
Issues
Challenges
Priorities
Users

ABSTRACT

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. In this article, based on the report of an interdisciplinary workshop convened by the National Science Foundation, we identify key scientific and engineering advances needed to enable effective spoken language interaction with robotics. We make 25 recommendations, involving eight general themes: putting human needs first, better modeling the social and interactive

* Corresponding author.
E-mail address: matthew.r.marge.civ@mail.mil (M. Marge).

https://doi.org/10.1016/j.csl.2021.101255
Received 2 December 2020; Received in revised form 17 March 2021; Accepted 3 June 2021
Available online 2 July 2021

aspects of language, improving robustness, creating new methods for rapid adaptation, better integrating speech and language with other communication modalities, giving speech and language components access to rich representations of the robot's current knowledge and state, making all components operate in real time, and improving research infrastructure and resources. Research and development that prioritizes these topics 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

As robotics advances, spoken language interaction is becoming increasingly necessary. Yet robot researchers often find it difficult to incorporate speech processing capabilities, speech researchers seldom appreciate the special needs of robot applications, and the overall path forward has not been clear. To address these problems, the National Science Foundation convened a workshop. The discussions started there continued over a year, culminating in a report (Marge et al., 2020a), listing both challenges that robotics brings for spoken interaction, and challenges in designing robot systems able to make effective use of speech.

This article presents key recommendations for research in this area, distilled from the full report. We refer readers wanting more detail to that report and to the recent report of a Dagstuhl workshop on the same topic (Devillers et al., 2020). Also relevant are various roadmaps for research in related areas: robotics, dialogue systems, artificial intelligence (AI), and so on Amershi et al. (2019), Beckerle et al. (2019), Chang et al. (2019), Christensen et al. (2009), Eskenazi and Zhao (2020), Gil and Selman (2019), Bohus and Horvitz (2010), Sheridan (2016), Ward and DeVault (2016), Wiltshire et al. (2013), Yang et al. (2018), Tangiuchi et al. (2019), McTear (2020). In this article we focus on making specific recommendations relevant to issues that are especially critical for spoken language interaction with robots.

We hope to help researchers in speech, in language, and in robotics, 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. We make 25 recommendations, addressed to funding agencies, leaders in industry, principal investigators, graduate students, developers, and system integrators. These recommendations relate broadly to issues of human needs, sociality and interaction, robustness, adaptation, multimodality, representations, timing, and infrastructure.

1.1. Why spoken language interaction with robots?

Across a wide range of applications, spoken language interaction with robots has great promise. The possibilities for education, healthcare, field assistance (including search and rescue, humanitarian relief, and reconnaissance) and the consumer market (including entertainment, security, and household) are vast.

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.
- Spoken language interaction is socially potent (Bainbridge et al., 2011), 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 Deng et al. (2019); 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.

1.2. 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 (Moore, 2015). But it's not that easy. This article 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 some ways a robot is like any other AI system that needs to converse with humans, there are also fundamental differences. 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, 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 (Reddy, 1979). 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. 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. At the other extreme, robots should be able to use spoken interaction when establishing long-term expectations (Kruijff-Korbayová et al., 2015). 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.

In this article our contribution to this endeavor is a set of research recommendations that we believe will accelerate progress in this area, each briefly motivated and contextualized. These recommendations are organized into seven sections: user experience design (Section 2), audio processing, speech recognition, and language understanding (Section 3), speech synthesis and language generation (Section 4), dialogue (Section 5), other sensory processes (Section 6), robustness and adaptability (Section 7), and infrastructure (Section 8).

2. User experience design

At the top level, we can say that there are 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 some implications, organized around four broad recommendations.

Recommendation 1. *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 (Balentine, 2007).

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 (Moore, 2017a).

As a corollary, we advocate generally preferring evaluations to be 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 datasets. 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 (Moller et al., 2009) that the generality can always be questioned. Thus we suggest continued research towards new evaluation methodologies that are both efficient and highly informative.

Recommendation 2. *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. Designers can, however, help users form a more accurate mental model of a robot, 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 (Moore, 2017b), 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", where usability or acceptability drops as flexibility increases (Phillips, 2006; Mori, 1970). Of course there can be trade-offs between attracting users to engage in the first place and enabling truly effective interaction (Luger and Sellen, 2016).

Recommendation 3. *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. Among other communicative behaviors, 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 ways to maximize the effectiveness of future speech-capable robots.

Recommendation 4. *Design for use in multi-party and team situations.*

Today human–robot interaction is generally designed to support single users, but many robots will function in environments with more than one person. These may include bystanders, members of a team tasked to jointly work with the robot, or anything in between. Moreover the roles of the humans may change over time.

A need to interact 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 or not it is the addressee of some communication. While multi-party interaction with robots has already been demonstrated in some situations (Al Moubayed et al., 2012; Bohus et al., 2014; Matsuyama et al., 2015), enabling multi-party interaction more generally remains a challenge.

3. Audio processing, speech recognition, and language understanding

With this section we start making component-by-component recommendations, starting with issues relating to the robot's ability to understand what the user is saying. Roboticists looking to exploit speech recognition today face numerous challenges. This is often a source of surprise, since as consumers we are familiar with high-performing speech recognizers, exceeding human performance in some cases. Nevertheless, making use of the technology for robots is still very hard. This section summarizes some issues and recommends some directions for overcoming them.

Recommendation 5. *Develop general techniques and toolkits for front-end audio processing.*

For many speech processing systems, audio input is a complex mixture of user speech, the system's own speech output, and noise from the environment, convolved with reverberation. Aspects of this problem have been well-researched, and smart speakers, for example, do quite well even in complex environments, thanks to intensively-tuned algorithms. However, robots in addition often face other signal sources, including speech from multiple speakers and meaningful sounds from the environment, raising the challenges of speaker diarization and audio scene analysis. Additional challenges often include noise from the robot's own motors, and ongoing change in the relative positions of sound sources as the robot moves through the environment. We recommend more research on these topics, with the aim of creating reusable general toolkits for front-end processing.

Recommendation 6. *Develop speech recognizers designed for robotics applications.*

Roboticists evaluating the options for speech recognition currently face some unpleasant trade-offs. First, while cloud-based systems are often highly accurate, the latency in accessing them is often too high to support effective natural interaction with robots, and conversely, local speech recognizers, while avoiding this problem, typically are inferior in vocabulary size and accuracy. Second, pre-trained recognizers, while convenient and powerful, are invariably tuned using datasets and objective functions that are quite unlike those needed for robots, but training one's own models can be a major task in itself.

Speech recognition for robots also often faces spontaneous and fragmentary utterances, as users may speak in rapidly-changing situations, and under conditions of high cognitive load. Other requirements include robustness in the face of noise and multiple speakers, retrainability to perform well in narrow domains, fast incremental processing, and awareness of time in order to work well with components for environment tracking, prosody processing, multimodal input, realtime output, and so on. While we expect that existing engines can be extended and adapted to work well for robotics, it is also worth considering creating a branch of an existing open source recognizer to specifically target the needs of robotics.

Recommendation 7. *Build a database of robot-directed speech, and organize a challenge task on speech recognition for robot-directed speech.*

Building a suitable speech recognizer will require suitable data. Today we lack good models of the sorts of things that people tend to say to robots, and how they say them (Marge et al., 2020b). Thus we suggest the development of a large dataset of human–robot speech, perhaps on the order of a thousand hours. This would support both the training of better models with existing tools and the evaluation of new techniques. As no single dataset could handle all the types of speech and situations needed for robotics, this 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 various outdoor environments; for a variety of tasks; for various user demographics; and for a variety of microphones including headset, on-board, and microphone-array.

Such a dataset would support shared tasks on speech recognition for robot-directed speech. We envisage that this will foster the development of fast and accurate recognizers pre-trained on massive data, but easily and robustly adaptable to specific contexts of use.

Beyond the improvement of speech recognizers, data is, more generally, the lifeblood of research in spoken language interaction with robots. 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. While some of these restrictions protect privacy or exist for other good reasons, we still need to work to find ways to better share data. While we see no simple solutions, one initial step is for researchers, whenever possible, to design data collections to be fully shareable. In some countries, this may be as simple as having participants dedicate their "work" to the public domain, or using the Creative Commons CC0 license.

Recommendation 8. *Better represent context and expectations to support speech recognition.*

Speech directed to robots thus brings many challenges, 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. In support of such dynamic language modeling, research should target the ability to fully exploit the robot's knowledge for this. We envision methods to map from the entire state and context to a probability distribution over all the words in the vocabulary, continuously updated, and speech recognizers with suitable APIs for ingesting this information.

Recommendation 9. *Develop language understanding models for robots that resolve referential and other ambiguities in spoken dialogue.*

Language understanding in general is a much-studied problem, but for robots we require in addition grounded (or situated) language understanding, where success depends on correctly interpreting language that refers to, or has meaning only in the context of, the physical environment (Tellex et al., 2020). A major complication is the fact that a robot's environment model is never veridical, and is never the same as a human's model of the same environment. Aspects of language understanding may need to be designed or tuned to effectively support specific downstream tasks, such as identifying or disambiguating referents (objects, regions, and paths), following instructions, obeying constraints on motion, planning and scheduling actions to meet goals, and communicating back to the user. Since the task of grounded language acquisition goes hand-in-hand with understanding the environment, and given the complex physical worlds in which robots may operate, language understanding abilities may need to be learned and tuned *in situ*, rather than predefined for all situations and tasks.

Recommendation 10. *Better exploit prosodic information.*

Speech includes both words and prosody. Speech recognizers handle only the former, meaning that much of the information in the speech signal is discarded. 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 how the utterance relates to the user's goals or the temporal context, or whether the user is confused, preoccupied, distressed, or unsure. 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 not only extract prosodic information in real time and provide a continuous read-out of the results, but also output information that is directly useful for robot task planning and other downstream components.

4. Speech synthesis and language generation

We turn our attention to the most salient user-facing functionality: speech synthesis.

Recommendation 11. *Develop the ability to tune speech synthesizers to convey a desired tone, personality, and identity.*

The voice of an artificial agent 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. While one can argue that robots should be purely functional, and that designers should not bother to produce robots that project a specific personality, 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. While effective for some applications, a robot interacting with a small child should talk very differently, and a robot assisting in a disaster recovery effort should sound different again.

Further, although contemporary speech synthesis is capable of generating utterances virtually indistinguishable from those produced by a human being, this is inappropriate if it encourages people to overestimate a robot's linguistic and cognitive capabilities. Rather "robotic" voices — by which we mean not low-quality voices, but high-quality voices that sound like a robot should sound — can be more appropriate (Wilson and Moore, 2017).

Recommendation 12. *Extend the pragmatic repertoire of speech synthesizers.*

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. But even this is not adequate for most robotics applications.

Robots operate in real time and real space. Speech synthesis in this context needs access to the full expressive power of spoken language. In particular, this includes prosody, that is, those features of the speech input that are not governed by the phoneme sequence of the words said, including features of pitch, energy, rate, and voicing. 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 to 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. 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. In others, it suffices to convey just simple propositions and intents. But in general, there is a need to convey many kinds of information, along with aspects of the robot's attitude and internal state. Effective language generation remains a challenging problem (Gatt and Krahmer, 2018), especially for robots, for reasons already noted.

It can be helpful to use the term "speech generation", 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 — is

the norm, but this is problematic (Bulyko and Ostendorf, 2001), as the language and speech decisions are often interdependent. 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 a challenge. Instead, researchers will need to explore 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.

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 (Wang et al., 2018), 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 (Ward, 2019) that robots most need. There is thus a need for more work specifically targeting the speaking needs of robots, as they perform actions in time and space.

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 for a collaborating entity is not to minimize one's own effort, but rather to minimize the joint effort to come to a common ground (Clark, 1996). Thus a robot should often make the extra effort to ensure 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, or by proactively first describing essentials of its own internal representation, to make subsequent grounding more efficient (Chai et al., 2014; Fang et al., 2015). 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 13. *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 not be paying attention, and robots should do the same. 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. To coordinate spoken language with a robot's physical gestures and motion, synthesizers must need to be able to output sync points and to support fine-grained timing control.

Recommendation 14. *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.

5. Dialogue

Dialogue is more than the sum of processing speech input and creating speech output. We want spoken language interaction with robots to include coherent, flowing, natural, and efficient user experiences.

Recommendation 15. *Focus on highly interactive dialogue.*

Fluent 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 Ward and DeVault (2016), 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 (Cohen et al., 2004) 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 (DeVault et al., 2014; Gratch et al., 2007; Yu et al., 2015; Kawahara, 2019). Such abilities are, moreover, often highly valued: users interacting with systems (or people) with better interaction skills may trust them more, like them more, and use them more (Fusaroli and Tylén, 2016). Yet 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. 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. We see the modeling of dialogue as an intellectual grand challenge, which requires both scientific and engineering advances.

Recommendation 16. *Make every component able to support realtime responsiveness.*

Current dialogue-capable robots offer only slow-paced, turn-based interactions, with few exceptions (Skantze, 2021). This is now due less to 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 complete 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, each component will probably 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, speech synthesis, and general abstract models and toolkits for incrementality, but much remains to be done (Kennington et al., 2020).

As an important special case, systems should strive to update the representation of a robot's physical surroundings continuously. This will support not only robust speech processing, but also the ability to quickly ground the inputs in the context.

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, including cameras, laser scanners, infrared sensors, GPS modules, and so on. Different sensors operate at different sampling rates, and the downstream processes — speech recognition, object detection, planning, execution monitoring, and so on — have different processing speeds. These cause different delays between events in the world and the time they are recognized, and so, for example, if a user points to an object and then to a location while saying *put that there*, it is nontrivial to properly fuse the information from the speech and visual inputs. Similarly on the output side: spoken output must be timed and synchronized in concert with actions in other modalities. 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 their meaning (Boltz, 2005; Itoh et al., 2009; Roberts and Francis, 2013). While we understand some aspects of these issues, we need more general models of how to time and align multimodal actions. The need for proper handling of time applies to all modules and aspects of processing. The Platform for Situated Intelligence (Bohus et al., 2017) illustrates how it is possible to provide mechanisms for this, but issues of synchronization and temporal alignment still bring many challenges.

6. Other sensory processes

To support effective language use, robots need to broadly understand what is going on. This is needed both in order to fully understand what the user means and to effectively convey information to the user. We have two specific recommendations.

Recommendation 17. *Improve audio scene and event analysis methods to better understand the environment.*

Robots need to understand the sounds in their environment, as these can 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 use the information in sounds from vehicles and bystanders to understand the environment, the events, and the social context, and to enable it to understand what its human teammates are hearing and reacting to.

Recommendation 18. *Develop methods to infer and represent more information about human interactants.*

Table 1
Common robot components and functions.

Human-Directed Perception: recognizing individual people, gesture recognition, sentiment analysis …
User Intent Inference: speech recognition, language understanding, inferring the user's beliefs, goals, plans, and intentions …
Environment-Directed Perception: scene interpretation (both metric and semantic), identifying objects, entity recognition, grounding …
Action Planning, Execution, and Monitoring: task modeling, action-sequence selection, spatial and temporal reasoning, path planning, collision avoidance …
Dialogue: next-action selection, achieving common ground, clarification requests, joint decision making, using downstream information to report on successes, failures, and errors to humans when appropriate, facial expression generation, gaze control …

For many robots today, a user is just a disembodied source of speech input, or at best, an approximate image region with an estimated location and velocity. This may be fine for basic competencies such as not bumping into people, but we would like robots to be able to infer much more, such as what the user is doing, what they are paying attention to, what they are likely to do next, aspects of their energy level, current cognitive load level, emotional and physical state, and so on.

There are many sensors that can serve this purpose; beyond microphones and cameras, these include depth sensors, laser scanners, haptic and proprioceptive sensors, and other peripherals. Semantically, these inferences can exploit many types of information beyond speech and language: also facial gestures, hand gestures, posture shifts, skin color changes, and so on. There is a considerable body of research already on emotion recognition, sentiment analysis, mental health assessments and so on, but this needs to be extended to support robot needs. This is most obvious in scenarios where caring for the user is part of the job — for example, when helping a child read, a robot should be able to tell whether the child is getting frustrated, and if so change its strategy, or a robot involved in elderly care that notices depressive symptoms might alert family members — but the need is more general: understanding more about the user can in many ways help it better accomplish the task at hand.

7. Robustness and adaptability

In this section we change our focus, from the desirable properties of specific components to general properties that most or all components will need. We focus on properties necessary for robustness and adaptability. While the desirability of such properties is obvious, so far they have received less attention than they deserve. One likely reason is that in robotics research, as in many other fields, successful demos are celebrated: seeing a new technology in action can be a source of great inspiration. However, success in demos is not very predictive of success in deployment, and most demos only illustrate an ideal case. But robotics is increasingly targeting solid experimental validation of capabilities and real deployments in the open world, and robustness and adaptability are becoming more essential.

The recommendations in this section are relevant broadly across the components and functionalities of robots, with a partial, suggestive list given in Table 1.

Recommendation 19. *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 Admoni and Scassellati (2017), Gaschler et al. (2012), Ward (2019). While robots can perform well in demos with only one of these functions, this requires that both the environment and the user be tightly constrained. Adding competence with other modalities, beyond just the words alone, can contribute to robustness. Achieving this requires better scientific understanding of these aspects of behavior, more explorations of utility for various use cases, more work on cross-modality integration, and more shareable software modules.

Recommendation 20. *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 illustrate with three examples: First, for intent recognition, a developer cannot assume that a robot will 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 each component should constantly track multiple hypotheses and maintain a probability estimates for them, but to do so involves many challenges. One of these is the need for something of a change of mindset: developers probably need to accept that simple, understandable, inspectable representations may not be generally adequate. Another is that, even when we know how to make individual components probabilistic, integration with other components remains difficult.

Recommendation 21. *Explore the broad space of recovery strategies in spoken language interaction with robots.*

A critical capability for robots will be the ability to ask for help or clarify when something is unclear or confusing. Even if a robot's uncertainty modeling is accurate and it can infer which uncertainties truly need resolving, deciding whether the human can help and if so how to ask are challenging problems. An important special case is uncertainty that arises from the communication itself, including non-understandings, misunderstandings, unresolvable ambiguities, and other miscommunications (Marge and Rudnicky, 2019). While the dialogue literature includes many general recovery strategies, robots bring more complexity: with more possible ways to misunderstand, but also an expanded space of recovery strategies, including not just speech but also actions involving gesture and movement. For example, early signaling of the need for recovery may need just a raised eyebrow or a sudden slowing of motion.

Recommendation 22. *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 also 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 done by customizing using small sets of training data, and partly handled by exposing parameters that developers can 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* (Oviatt et al., 1997). We know that some prefer swift turn taking with frequent overlaps, while others prefer to wait until the other is silent before speaking (Tannen, 1989). We also 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 Ward (2019). 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. In addition to basic research, we need to develop ways for robots to not only embody plausible and consistent interaction styles, but to select among these to adapt to the (implicit) interaction style preferences of specific users.

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 23. *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, to enable the creation of components that are simultaneously high-performing and highly reusable. This may involve pre-training on massive datasets, with mechanisms for easily and robustly ablating or adapting the models to perform well on robots with different hardware or abilities, or on specific small domains, including, for some experimental purposes, exceedingly narrow domains.

Recommendation 24. *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. It would be even more convenient if this could be done by simply pipelining together a speech recognizer, a natural-language understander, a finite-state dialogue manager, and a speech synthesizer. While such systems can be built, and may suffice for simple interactions — such as greeting a customer, prompting for a simple request, and confirming execution — they are not sufficient for more natural interactions, especially not in open worlds. While research effort has tended to gravitate to improving these familiar components, there are many issues that fall through the cracks of such an architecture. 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 always be difficult for robots, as we will never have enough training data for all the complex and diverse tasks that robots need to do.

Thus we see the need for improved ways to share typically component-initial models, such as that of what the user knows, and that of the current state of the environment. Doing this will support advances on cross-cutting issues, such as grounding, ambiguity, social adeptness, prosody, and adaptation. Evaluation of component-level performance is now routine, but evaluating progress on cross-cutting issues is still a challenge. We may need to develop testbeds and evaluation metrics for touchstone tasks, such as grounding, where success requires the successful integration of many sources of information.

8. Infrastructure

Our recommendations so far have mostly targeted research questions. However there is another important way:

Recommendation 25. *Create and distribute one or more minimal speech and dialogue-capable robot systems.*

Research in spoken dialogue for robots has high barriers to entry. 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, not least because individual components, such as automatic speech recognition, even when well-tested in other domains, often don't transfer well to robots. Thus 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.

We have already discussed how currently available speech technologies might be improved to meet roboticists' needs. Conversely, currently available robot hardware and software offerings could be extended to make them easy for speech researchers to experiment on.

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 could support individual use, for hobbyists and small projects, or serve as a shared platform to bring together researchers from robotics, speech, social interaction, computer vision, and so on. These may range from, on the one hand, just a recognizer, synthesizer, and robot, with everything else to be custom-built (or kludged) for the intended use, to, on the other hand, integrated highly functional systems, engineered to support data capture, replay, visualization, performance analysis, and informative experimentation.

Such shared infrastructure will of course need to incorporate a powerful and flexible software architecture. In this context we must note the Robot Operating System[1] (ROS), a well-established platform for robotic systems, including virtual robots. As middleware, ROS does permit the capability to operate at latencies supporting realtime human interaction. In practice, however, researchers building on ROS have tended to make limiting assumptions about human interaction, in particular regarding highly time-sensitive tasks of the kind common in social robotics. Demonstrations which are able to go beyond this, to broadly capture the context of interact with humans, have been mostly implemented as standalone models, separate from the ROS software stack. Nevertheless, 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. Other platforms are actively under development, so the trend is very positive.

While no single solution will serve all needs, the creation of shareable infrastructure will greatly increase the number of researchers able to contribute to this area.

9. Conclusion

In general, we envisage 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 the words of Ward and DeVault (2016), 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.

This vision will only be realized, we believe, *by deliberate and concerted action targeting the issues we presented.*

None of the issues are entirely new: all have been discussed before. Yet they remain as unsolved issues today, in large part because they have fallen through the cracks. Thus, to enable spoken language interaction with robots, we will need advances not only in speech science and in robotics, but also 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 article, by providing specific recommendations, 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.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Tanya Korelsky and the National Science Foundation, United States of America for initiating our discussions and for support, via Grant #IIS-1941541. We also thank the government observers at the workshop for their contributions to the discussions: Jonathan Fiscus, Susan G. Hill, Nia Peters, Erion Plaku, Christopher Reardon, and Clare Voss. We also thank Erin Zaroukian for further discussion.

[1] https://www.ros.org

References

Admoni, H., Scassellati, B., 2017. Social eye gaze in human-robot interaction: A review. J. Hum.-Robot Interact. 6 (1), 25–63.

Al Moubayed, S., Beskow, J., Skantze, G., Granström, B., 2012. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In: Cognitive Behavioural Systems. Springer, pp. 114–130.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., 2019. Guidelines for human-AI interaction. In: Proceedings of CHI, pp. 1–13.

Bainbridge, W.A., Hart, J.W., Kim, E.S., Scassellati, B., 2011. The benefits of interactions with physically present robots over video-displayed agents. Int. J. Soc. Robot. 3 (1), 41–52.

Balentine, B., 2007. 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.

Beckerle, P., Castellini, C., Lenggenhager, B., 2019. Robotic interfaces for cognitive psychology and embodiment research: A research roadmap. Wiley Interdiscip. Rev. Cogn. Sci. 10 (2).

Bohus, D., Andrist, S., Jalobeanu, M., 2017. Rapid development of multimodal interactive systems: A demonstration of platform for situated intelligence. In: Proceedings of ICMI, pp. 493–494.

Bohus, D., Horvitz, E., 2010. On the challenges and opportunities of physically situated dialog. In: Proceedings of the AAAI Fall Symposium on Dialog with Robots.

Bohus, D., Saw, C.W., Horvitz, E., 2014. Directions Robot: In-the-wild experiences and lessons learned. In: Proceedings of AAMAS, pp. 637–644.

Boltz, M.G., 2005. Temporal dimensions of conversational interaction: The role of response latencies and pauses in social impression formation. J. Lang. Soc. Psychol. 24 (2), 103–138.

Bulyko, I., Ostendorf, M., 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In: Proceedings of ICASSP, Vol. 2, pp. 781–784.

Chai, J.Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., Hanson, K., 2014. Collaborative effort towards common ground in situated human-robot dialogue. In: Proceedings of HRI, pp. 33–40.

Chang, S.-F., Hauptmann, A., Morency, L.-P., Antani, S., Bulterman, D., Busso, C., Chai, J., Hirschberg, J., Jain, R., Mayer-Patel, K., Meth, R., Mooney, R., Nahrstedt, K., Narayanan, S., Natarajan, P., Oviatt, S., Prabhakaran, B., Smeulders, A., Sundaram, H., Zhang, Z., Zhou, M., 2019. Report of 2017 NSF workshop on multimedia challenges, opportunities and research roadmaps. arXiv preprint arXiv:1908.02308.

Christensen, H.I., Batzinger, T., Bekris, K., Bohringer, K., Bordogna, J., Bradski, G., Brock, O., Burnstein, J., Fuhlbrigge, T., Eastman, R., Edsinger, A., Fuchs, E., Goldberg, K., Henderson, T., Joyner, W., Kavaraki, L., Kelly, C., Kelly, A., Kumar, V., Manocha, D., McCallum, A., Mosterman, P., Messina, E., Murphey, T., Peters, R.A., Shepherd, S., Singh, S., Sweet, L., Trinkle, J., Tsai, J., Wells, J., Wurman, P., Yorio, T., Zhang, M., 2009. A Roadmap for US Robotics: From Internet to Robotics. Computing Community Consortium (CCC).

Clark, H.H., 1996. Using Language. Cambridge University Press.

Cohen, M.H., Giangola, J.P., Balogh, J., 2004. Voice User Interface Design. Addison-Wesley.

Deng, E., Mutlu, B., Matarić, M.J., 2019. Embodiment in socially interactive robots. Found. Trends® Robot. 7 (4), 251–356.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., Xu, Y., Rizzo, A., Morency, L.-P., 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In: Proceedings of AAMAS, pp. 1061–1068.

Devillers, L., Kawahara, T., Moore, R.K., Scheutz, M., 2020. Spoken language interaction with virtual agents and robots (SLIVAR): Towards effective and ethical interaction (dagstuhl seminar 20021). In: Dagstuhl Reports. Tech. Rep., 10, Schloss Dagstuhl: Leibniz-Zentrum für Informatik, pp. 1–51.

Eskenazi, M., Zhao, T., 2020. Report from the NSF future directions workshop, toward user-oriented agents: Research directions and challenges. arXiv preprint arXiv:2006.06026.

Fang, R., Doering, M., Chai, J.Y., 2015. Embodied collaborative referring expression generation in situated human-robot dialogue. In: Proceedings of HRI, pp. 271–278.

Fusaroli, R., Tylén, K., 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. Cogn. Sci. 40 (1), 145–171.

Gaschler, A., Jentzsch, S., Giuliani, M., Huth, K., de Ruiter, J., Knoll, A., 2012. Social behavior recognition using body posture and head pose for human-robot interaction. In: Proceedings of IROS, pp. 2128–2133.

Gatt, A., Krahmer, E., 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. J. Artificial Intelligence Res. 61, 65–170.

Gil, Y., Selman, B., 2019. 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).

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R.J., Morency, L.-P., 2007. Can virtual humans be more engaging than real ones? In: Proceedings of HCI, pp. 286–297.

Itoh, T., Kitaoka, N., Nishimura, R., 2009. Subject experiments on influence of response timing in spoken dialogues. In: Proceedings of Interspeech, pp. 1835–1838.

Kawahara, T., 2019. Spoken dialogue system for a human-like conversational robot ERICA. In: 9th International Workshop on Spoken Dialogue System Technology. Springer, pp. 65–75.

Kennington, C., Moro, D., Marchand, L., Carns, J., McNeill, D., 2020. rrSDS: Towards a robot-ready spoken dialogue system. In: Proceedings of SIGdial, pp. 132–135.

Kruijff-Korbayová, I., Colas, F., Gianni, M., Pirri, F., de Greeff, J., Hindriks, K., Neerincx, M., Ögren, P., Svoboda, T., Worst, R., 2015. TRADR project: Long-term human-robot teaming for robot assisted disaster response. KI-Künstliche Intell. 29 (2), 193–201.

Luger, E., Sellen, A., "Like Having a Really Bad PA": The Gulf between user expectation and experience of conversational agents. In: Proceedings of CHI, pp. 5286–5297.

Marge, M., Espy-Wilson, C., Ward, N., 2020a. Spoken language interaction with robots: Research issues and recommendations, report from the NSF future directions workshop. arXiv preprint arXiv:2011.05533.

Marge, M., Gervits, F., Briggs, G., Scheutz, M., Roque, A., 2020b. Let's do that first! A comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In: Proceedings of SemDial.

Marge, M., Rudnicky, A.I., 2019. Miscommunication detection and recovery in situated human–robot dialogue. ACM Trans. Interact. Intell. Syst. 9 (1).

Matsuyama, Y., Akiba, I., Fujie, S., Kobayashi, T., 2015. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. Comput. Speech Lang. 33 (1), 1–24.

McTear, M., 2020. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Morgan & Claypool Publishers.

Moller, S., Engelbrecht, K.-P., Kuhnel, C., Wechsung, I., Weiss, B., 2009. 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, pp. 7–12.

Moore, R.K., 2015. From talking and listening robots to intelligent communicative machines. In: Markowitz, J. (Ed.), Robots that Talk and Listen. de Gruyter, pp. 317–335.

Moore, R.K., 2017a. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In: Jokinen, K., Wilcock, G. (Eds.), Dialogues with Social Robots. Springer, pp. 281–291.

Moore, R.K., 2017b. 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, pp. 7–11.

Mori, M., 1970. Bukimi no tani (The uncanny valley). Energy 7, 33–35.

Oviatt, S., DeAngeli, A., Kuhn, K., 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In: Proceedings of CHI, pp. 415–422.

Phillips, M., 2006. Applications of spoken language technology and systems. In: Proceedings of the IEEE/ACL Workshop on Spoken Language Technology. SLT, IEEE, p. 7.

Reddy, M., 1979. The conduit metaphor — A case of frame conflict in our language about language. Metaphor Thought 2, 285–324.

Roberts, F., Francis, A.L., 2013. Identifying a temporal threshold of tolerance for silent gaps after requests. J. Acoust. Soc. Am. 133, 471–477.

Sheridan, T.B., 2016. Human-robot interaction: Status and challenges. Hum. Factors 58 (4), 525–532.

Skantze, G., 2021. Turn-taking in conversational systems and human-robot interaction: A review. Comput. Speech Lang. 67, 101–178.

Tangiuchi, T., Mochihashi, D., Nagai, T., Uchida, S., Inoue, N., Kobayashi, I., Nakamura, T., Hagiwara, Y., Iwahashi, N., Inamura, T., 2019. Survey on frontiers of language and robotics. Adv. Robot. 33 (15–16), 700–730.

Tannen, D., 1989. That's Not What I Meant! how Conversational Style Makes or Breaks Relationships. Ballantine.

Tellex, S., Gopalan, N., Kress-Gazit, H., Matuszek, C., 2020. Robots that use language. Ann. Rev. Control Robot. Auton. Syst. 3, 25–55.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., Saurous, R.A., 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: Proceedings of ICML, pp. 5180–5189.

Ward, N.G., 2019. Prosodic Pattterns in English Conversation. Cambridge University Press.

Ward, N.G., DeVault, D., 2016. Challenges in building highly-interactive dialog systems. AI Mag. 37 (4), 7–18.

Wilson, S., Moore, R.K., 2017. 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, pp. 40–44.

Wiltshire, T.J., Barber, D., Fiore, S.M., Towards modeling social-cognitive mechanisms in robots to facilitate human-robot teaming. In: Proceedings of HFES, Vol. 57, pp. 1278–1282.

Yang, G.-Z., Bellingham, J., Dupont, P.E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B.J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z.L., Wood, R., 2018. The grand challenges of science robotics. Science Robotics 3 (14).

Yu, Z., Bohus, D., Horvitz, E., 2015. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In: Proceedings of SIGdial, pp. 402–406.