Challenges in Building Highly Interactive Dialogue Systems

Nigel G. Ward, David DeVault

■ *Spoken dialogue researchers have recently demonstrated highly interactive systems in several domains. This paper considers how to build on these advances to make systems more robust, easier to develop, and more scientifically significant. We identify key challenges whose solution would lead to improvements in dialogue systems and beyond.*

Over the past decade, dialogue researchers have built several systems that go beyond rigid turn-taking and robotic interactions. In research-driven systems, we are beginning to see demonstrations of humanlike sensitivity to the user's behavior and state, swift and natural timing, and appropriately tailored behaviors. We are learning that the careful design of interactive skills in our systems can lead to improvements in naturalness, efficiency, feelings of rapport, and task-related outcomes. Research systems are providing a vision of what is possible.

However much work remains before such abilities are robust, widely useful, and generally available. This article identifies 10 key challenges, relating to modeling, systems architecture, and development methods. Of pressing importance for dialogue systems, these challenges are also relevant for intelligent and interactive systems more generally.

Given Siri's broad deployment and popular salience, one might imagine that it solved the problems of interacting in dialogue: we often meet people who are unaware how cleverly Siri and her sisters avoid dialogue. While they do use speech, their preferred interaction style is to map one user input to one system output, avoiding any of that messy interaction stuff.

While this helps them do well on simple tasks such as command recognition or question answering, many user needs are too complex to address in a single input-output exchange. These systems are missing half the promise of speech-based interaction.

We envision the creation of much more highly interactive systems. In broad strokes, these systems will be characterized by low latency and natural timing, a deft sensitivity to the multifunctional 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 real-time communicative signals. Their skill at understanding the multifunctional effects of utterances will mean, for example, that they decide not only what dialogue act to perform, but also with what prosody and nonverbal behavior, with what expected effects on the turn-taking state, and with what expected social effects such as implied attitudes, emotions, and potential for rapport building. Their flexibility about the interaction itself will mean users feel less constrained by obvious limitations in turn-taking protocols, supported dialogue flows, and expected speech patterns. The structure of their interactions will emerge more as a creative process than as a simple instantiation of a preplanned interaction template. As we develop the technology to support such interactive skills, we believe dialogue will become the interface of choice for a much broader range of applications.

The challenges in providing such interactivity are many. Our survey here is based on our experiences as researchers and developers, and on analysis of other recent advances in spoken dialogue systems, intelligent virtual agents, and human-robot interaction, including Gratch et al. (2007); DeVault, Sagae, and Traum (2009); Bohus and Horvitz (2011); Forbes-Riley and Litman (2011); Acosta and Ward (2011); Raux and Eskenazi (2012); Andrist, Mutlu, and Gleicher (2013); Meena, Skantze, and Gustafson (2014); Skantze, Hjalmarsson, and Oertel (2014); Ghigi et al. (2014); and Paetzel, Manuvinakurike, and DeVault (2015).

Challenge A: Surpassing Human Models

Implementing humanlike behavior has been a driving goal across AI, but increasingly the aim is instead to exceed human intelligence. For dialogue systems, super-human performance is a common vision, for example in science fiction movies. However, tellingly, such systems are portrayed as idiot savants: knowledgeable, logical, and well-spoken, but unable to interact smoothly with humans. We find it provocative to consider instead whether dialogue systems might one day be "interactionally superior" to the average human, or even most humans.

This is a very long-term goal, but the possibilities can be appreciated by listening to recordings of people in live conversation: we are surprisingly inefficient and awkward. While some disfluencies can be functional, most are regrettable. When listening to yourself, it's easy to see many things that you could have done or said better. For humans such insight is only possible when we can repeatedly replay a dialogue recording, noting every detail, and considering the options at leisure. Future dialogue systems, not subject to human cognitive limitations, might be able to do this in real time — to sense better, consider more factors, and plan further ahead — to be superhumanly efficient, easy to talk to, charming, and effective.

Among many other issues, this raises the question of how to surpass corpus-mining techniques. Although these are currently the mainstay for designing interactive behaviors, we will need methods to enable systems to do better than (at least some of) the human speakers in the corpus. Partial solutions to this problem exist already. One approach is to model consensus, to overcome the "noise" inevitable in the behavior of any individual (Huang, Morency, and Gratch 2010). A second approach is to somehow select the best interaction exemplars from a corpus, to learn from them. In training dialogue policies on large corpora, imitation learning, inverse reinforcement learning, and temporal supervised learning may help, although making them work at the time scale of interactive behaviors will clearly be difficult (Schaal 1999; Kim et al. 2014; Li, He, and Williams 2014).

Challenge B: Enabling Exploration

Today most interactive systems require tightly controlled user behavior. The constraints are often implicit, relying on a genre that sets up expectations and various hints that lead the user to perform only a very limited set of behaviors (Cohen, Giangola, and Balogh 2004): to follow the intended track. Under such constraints a system can behave strictly according to a role, which simplifies design and reduces 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.

This is not a problem for systems intended only for use in a controlled experiment, to demonstrate the utility of some new interaction ability. Like amusement-park rides, it's possible to deliver amazing expe-

riences if the users have no real freedom. However in general we want systems to be robust to users stepping off the expected track.

Consider for example back-channeling behavior. This is a prototypical interactive behavior, probably the best studied one, and one already implemented in several systems that back channel at appropriate times in response to user speech (Ward and Tsukahara 1999; Fujie, Fukushima, and Kobayashi 2005; Gratch et al. 2007; Schröder et al. 2012; Kawahara et al. 2015). Back-channeling demonstrations today work if the user has been guided to perform a specific type of dialogue, such as retelling a story, solving a puzzle, or engaging in small talk. Within one such activity type, good back channeling is possible, and this can give a powerful impression of engagement and improve users' perceptions and cooperativeness. However, any single back-channeling policy does not generalize to other activity types. If a user diverges from the designed-for interaction style, the illusion of engagement comes crashing down.

While it may never be possible to build a system that behaves appropriately for all possible user behaviors, across all domains, we do want to do better at providing more track options. At least we want to be able to tell whether the current state of the dialogue is properly within a system's zone of competence, and if not, to guide the user back on track.

Lacking this, the benefit of other system abilities will be limited. Because users are good at learning to avoid crashing the system, by sticking with a limited set of behaviors that seem to work, it is all too easy for the intended richness of an interaction to collapse into a two-dimensional caricature. Users can cope with single-track systems, but we want to enable them to be comfortable exploring.

Challenge C: Integrating Learned and Designed Behaviors

Today even the most interactive systems have a fixed skeleton specifying the overall dialogue flow. This reflects two pernicious common assumptions: that a dialogue system designer should know how to control the dialogue flow, and that by transitioning through a finite state machine, users will be able to reach their goals. The field has long struggled to overcome these assumptions (Pieraccini et al. 2009). Of course it is possible for a designer to leave a few decision points underspecified, for subsequent filling in with data-derived decision rules (Watanabe et al. 2014). On the other extreme, there are algorithms that learn policies for all branch points (Young et al. 2013; Li et al. 2016; Gasic et al. 2015).

Such elegant, learn-everything approaches ignore the reality that most interactive systems need to include both learned and designed behaviors. One reason everything cannot be learned is that customers for dialogue systems may not trust fully learned dialogue policies; they want to be able to verify that certain behaviors are guaranteed and others not allowed, and that overall the system meets specifications and achieves design goals (Pieraccini et al. 2009). Another reason is that dialogue systems are increasingly situated. All but chat-style dialogue systems interact with some back end, and substantial integration work is sometimes needed to connect learned policies to API actions. Finally, in embodied systems, the dialogue manager usually can't own the top-level perception-action loop. Rather, the system as a whole needs to be responsive at run time to externally generated goals (Raux and Nakano 2010; Bohus, Kamar, and Horvitz 2012). In a word, dialogue abilities generally need to be provided as integratable and controllable modules, not as stand-alone demonstrations.

Thus it is important to find better ways to integrate learned and designed behaviors. We are inspired by what has been done for animating motion, where it has become possible to synthesize behavior that smoothly combines reactions to the environment and agent goal directedness (Lee et al. 2014). We feel this could also be done for dialogue, combining two styles of modeling, one metaphorically the kinematics, modeling motion as it follows observed patterns, and important for local coherence, and the other metaphorically the dynamics, modeling motion as directed by external forces, such as system goals.

Challenge D: Synthesizing Multifunctional Behaviors

Much as realizing natural motion requires the simultaneous control of multiple actuators at multiple joints, effective dialogue fundamentally involves performing multiple actions in parallel. This crucial fact is completely obscured by our typical representations: when transcribed into words, speech looks purely sequential. From this perspective, narrow tracks seem natural, since they match the (mis)perception that dialogues pursue one goal at a time and involve one action at a time.

But consider the simple two-word utterance "okay, well." Depending on the prosody, this may serve to: accept that the interlocutor's proposal is a valid one, flag that it was unexpected, indicate that there are countervailing factors that he's probably not aware of, convey that the speaker needs a second to marshal his thoughts, and project that he momentarily will propose an alternative. In fact, all these things can be conveyed at once. In general, most human utterances are richly multifunctional (Bunt 2011), conveying things like the speaker's attitude and state, turn-taking intentions, how the utterance fits into the larger discourse, interpersonal feelings and stances, and so on. Among these are the social dimensions of dialogue, as humans in dialogue unavoidably engage all the perceptions and behav-

iors associated with social goals (Nass and Brave 2007). These happen in parallel with the other processes, and are typically not very accessible to introspection.

Multifunctional behaviors are a challenge for dialogue systems in many ways. On the input side, they require recognizing more than just the words (Clark 2014). In general, there is a lot more going on in human interaction than we are modeling today. These aspects are not always hard to detect — for example there have been good demonstrations of how to recognize user uncertainty and various emotion-related user states (Forbes-Riley and Litman 2011; Schuller et al. 2013) — but productively using such information remains a challenge. One problem is that today most behavior-related decisions are made in isolation. For example, a system might decide whether to produce a back channel, and if so which word, and then for that word which prosodic form to use. While such individual decisions simplify design, modularity can be taken too far. Instead, making decisions jointly, optimized together, could help produce better outputs that serve multiple functions.

Multifunctionality is also a problem for speech synthesizers. Current speech synthesis techniques support concatenation but not much superposition. Yet the latter is important because overlaying multiple prosodic patterns is essential in conveying multiple things simultaneously (Xu 2011, Ward 2014). Prerecorded speech or simple synthesis techniques are fine for systems whose expressive needs are limited, but more flexibility is required for systems whose expressive goals include combinations that are not statically predetermined. This is true not only for speech, but also for multimodal behaviors involving speech, animation, and action (Nijholt et al. 2008; Chao and Thomaz 2012; Huang and Mutlu 2014; Bailly et al. 2015; Ward and Abu 2016).

As humans, we're good at noticing sequential behaviors, and as systems builders, we're also comfortable introspecting on them and designing with them. To implement systems capable of using multifunctional and parallel behaviors, however, requires builders to move out of their comfort zones, to enable users to be in theirs.

Challenge E: Low-Resource Learning

Developing a highly interactive system, even one exhibiting only one or two forms of responsiveness, today requires a major engineering effort. Machine learning can of course decrease the need for analysis by hand, but brings its own costs and limitations.

Obviously huge corpora can have great value. In the image geolocation task (Weyand, Kostrikov, and Philbin 2016), where a machine must identify the location on Earth where a photo was taken, superhuman performance is partially attributed to the machine's training set including many more scenes than a human could encounter in a lifetime of travels. Similarly, highly interactive systems might learn interactive tactics by observing billions more dialogues than any individual human speaker could ever participate in. The dialogue agent Eve (Paetzel, Manuvinakurike, and DeVault 2015), which plays a picture-matching game where it identifies pictures as users describe them word-by-word, is able to achieve near human-level performance by training its dialogue policy on hundreds of human-human games — far more than its individual users ever play.

Some commercially deployed dialogue systems interact with millions of users per month, and it is possible to vary aspects of a system to find out what works better (Pieraccini et al. 2009). However in such contexts the experimentation must stay within the narrow bounds of acceptability. People are not generally eager to talk with experimental systems, and thus in practice machine learning of interactive behaviors is a low-resource problem. Much current work addresses this problem by building user simulations, including learning aspects of such simulations from data (Young et al. 2013), but this is still in its infancy. In particular, no current user simulations model the time course of human understanding or production, although dialogue is fundamentally a process in time (Clark 2002). One of the things that human conversants do, subconsciously, is monitor their listener's degree of understanding and cognitive capacity, moment by moment. Much of what we do in dialogue — including many of the "false starts," pauses, and repetitions — is purposeful, done to actively manage the listener's attention, to increase their receptiveness so that at a specific moment some crucial information can be delivered and have impact (Yu, Bohus, and Horvitz 2015).

Another learning-related issue is that of multiple levels. Most algorithms today work at just one level, for example addressing only issues in turn-taking, or only issues in next-move selection. Dialogue is however intrinsically multilevel and multiscale. Convolutional and deep neural networks have proven very useful for analogous problems in vision and other tasks, where they are able to model both low-level features and higher-level, wide-span features, and these may be useful for dialogue also (Kalchbrenner and Blunsom 2013).

Learning for dialogue, as generally, is never a free lunch. On the one hand, supervised learning requires labeling or preprocessing the data, to split it into turns or otherwise identify the decision points to which machine learning will be applied. On the other hand, unsupervised methods, although able to discover many things (Goldwasser and Daume III 2014), rely on strong and limiting assumptions about the nature of the dialogue. Thus, even when employing machine learning, most dialogue systems today incorporate a host of hand-crafted intermediate quantities, such as language understanding results,

representation of dialogue states, possible actions, and so on. Other fields in AI have seen steady progress by removing hand-crafted intermediate representations in favor of learned representations. Perhaps the same can be done for dialogue.

Unlike most tasks where machine learning has been effective, dialogue is not about simple input-output mappings, but about interacting. Systems need to interact with the real world (or APIs) to actually get stuff done. Systems also need to interact, of course, with the user. While massive amounts of dialogue data are available, simply training on human-human dialogue is seldom directly useful. System behavior changes affect user behavior, which affects the system's next actions, so any improvement to the system can have unpredictable consequences. In short, learning interactive behaviors is fundamentally different from learning input-output mappings, and is likely to require new approaches.

In summary, this section has argued that dialogue involves many challenges for learning, and that addressing them will require new methods that: make fewer assumptions, require less data, can model the details of human cognitive processing, and can operate across multiple levels.

Challenge F: Compositional Policy Specification

Today dialogue systems are hard to port to new domains. This is especially true for interactive behaviors: today typically tied to specific decision points, current "nodules of interactivity" are limited in applicability to the exact context in which they appear. We would instead like to build interactive systems from more general conversational skills, as reusable components. For example, imagining that we have developed a general policy for choosing the next interview question, a general policy for showing empathy, and a general policy for supportive turn taking, we could imagine that these could be composed to produce a system capable of effective, natural, and warm first-encounter dialogues. That is, dialogue systems might be built from communications skills that are decoupled from the overall dialogue policy. Ultimately we would like to be able to compose policies learned from different corpora, to increase reuse and reduce development costs. (This challenge emphasizes the ability to reuse behaviors and conversational skills across multiple domains, while Challenge D emphasizes an agent's ability to achieve multiple goals with a single behavior.)

Challenge G: Modeling User Variation

Every dialogue system today is carefully designed to work well with some target population of users. Adapting one to work well for a different population requires a significant engineering effort. Finding better ways to adapt is a major challenge. Recent explorations of interaction styles (Grothendieck, Gorin, and Borges 2011; Ranganath, Jurafsky, and McFarland 2013) suggest what is possible. Effective adaptation is not just a question of better algorithms; there is also a design challenge. We would like to be able to design a family of systems, with the same basic functionality but with different personalities or behavior styles, that can be used for users of different types or preferences. This means that we need ways to enable system behavior to be parameterized and adjusted at a high level.

Challenge H: Continuous Processing

Today's dialogue systems have a lot of inertia in interaction. After one makes a decision (which usually happens infrequently, such as once per user turn-end), it sticks with it, usually until it has delivered a full utterance and heard the user's response. Despite innovations in incremental processing, in practice these have been used so far just to add a few more decision points, for example when the user barges in or when a user's key words are recognized.

To enable more effective interaction we need more continuous decision making. Consider figure 1, representing an exchange between players of Fireboy and Watergirl, a maze game where the players run and jump and coordinate to overcome obstacles, such as the fatal green mud. Here, after the expert player cues the novice to jump, he lands in the green mud, realizes what happened, and apologizes, and then the expert reviews the relevant game rule.

When written out like this, the interaction looks like a nice sequence of turns, each responding to the previous one. But the reality is more interesting, as seen in figure 2. The timing suggests that each person's speech at each moment reflects his rapidly changing understanding of the situation. Here E realizes what's happened, N realizes it a half second later, then quickly diagnoses the problem "oh, green." Then E and N both speak, E to make sure that N understands what went wrong, and N to clarify that he already understands it. Their speech actions are a real-time reflection not only of the state of the game play, but also inferences about the other person's understanding of the situation, and about their communicative intentions.

This is perhaps an extreme example, but the point is a general one: dialogue is a continuous process (Clark 2002), and it is common for speakers to continuously monitor the state of the dialogue, at every moment making a fresh decision as to what they will do next. These decisions must be based on the most current information but also consider the recent context (Geiger et al. 2013).

Thus success in dialogue can require continuous

Expert: you first
(novice moves and lands in the green mud)
Expert: (laughs)
Novice: oh. oh, green.
Expert: yeah.
Novice: got it. yeah, okay, green.
Expert: there's green, yeah, neither of us can touch the green
Novice: my bad. okay.
Expert: nah, you're good, you're good

Figure 1. Transcript of a Highly Interactive Dialogue Fragment.

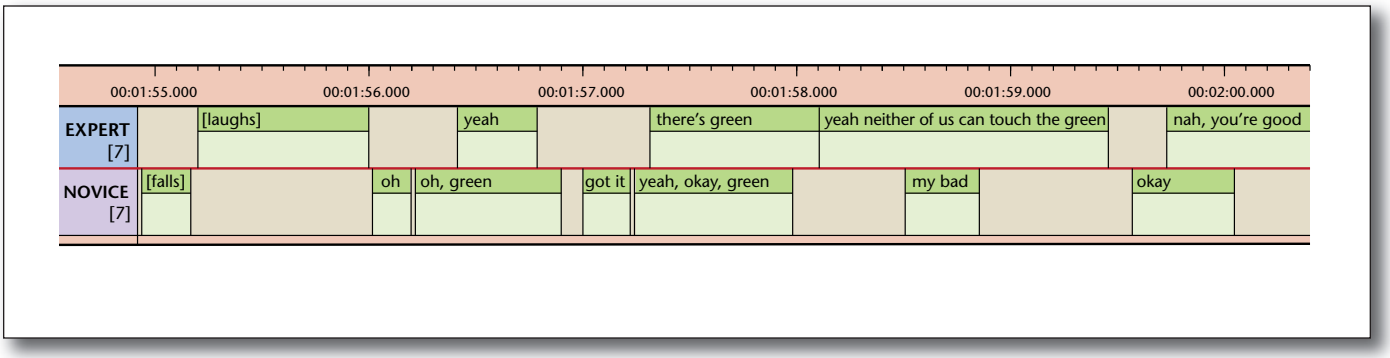

Figure 2. Timeline View of the Example in Figure 1.

tracking of the current state, regardless of who "has the floor." This involves not only fine attention to the interlocutor's gaze, gesture, and backchannels, but also self-monitoring: speakers monitor the words they are saying, what they sounded like after they said them, and what things are pending in their output buffers. For this they use their own recognition/understanding circuits to simultaneously emulate the listener's uptake and to compare their own actual performance with their intended message (Pickering and Garrod 2013). In this way human dialogue systems tightly integrate perception and action, and automated systems should also.

While implementing continuous state tracking won't be easy, the potential value is significant. Systems that react more quickly to new information may be seen as more efficient and natural conversational partners (Paetzel, Manuvinakurike, and DeVault 2015). Additionally, systems will be relieved of the pressure to make perfect decisions: if they can track appropriateness and make midcourse corrections, then the risk associated with any individual decision is less, and initial choices of how to start turns can be more approximate.

Essential to this is handling uncertainty. While this has recently become a focus in dialogue research, most designers still prefer the fantasy that they can be certain of the user's state and intention. Commercial dialogue systems employ various techniques that help designers maintain this fantasy, including slowing down the interaction pace, intimidating users into producing unnaturally clear inputs, and using tedious extra subdialogues to confirm the intended meaning. In practice, all of these techniques can hurt the user experience.

Today handling uncertainty is acknowledged as a central problem for dialogue systems, and there are good techniques for one aspect of the problem, accumulating and updating evidence for members of a set of preenumerated semantic hypotheses (Foster, Keiz-

er, and Lemon 2014; Williams et al. 2014). However the general question is unsolved. In particular, we need techniques able to track the probabilities of multiple possible interpretations of the user's instantaneous state, including cognitive load, receptiveness, and immediate turn-taking intentions. Designing for uncertainty and incrementality affects not only the recognizer and the synthesizer, but every component.

Challenge I: Making Evaluation more Informative

Today, evaluating highly interactive systems usually involves user studies with a final questionnaire. This is costly and not very informative. For example, consider the virtual interviewer in the SimSensei Kiosk system (DeVault et al. 2014). This system is deliberately slow to take the floor after user speech ends, in support of the design goal of encouraging users to talk as much as possible. If this system's turn-taking were made lower latency and more natural, it could work against system design goals. Or so we believe; but making choices like this is, today, more art than science, and our evaluations are not fine-grained enough to help much. In particular, it is difficult to relate user perceptions of system style — such as attentive, polite, considerate, supportive — to the details of the actual behaviors and the design choices underlying them — such as whether a certain state has a time-out of 1.2 or 1.8 seconds. We think this could be addressed in part by elaborating causal models of the relations between system properties and user perceptions (Möller and Ward 2008; Möller, Engelbrecht, and Schleicher 2008) to cover the more interactive aspects of dialogue. This would also help us understand the advantages and potential disadvantages of more interactive systems: more interactivity is not always better, and we need models to help us predict when it is and isn't.

Challenge J: Engaging Social Scientists

The behaviors in today's dialogue systems are seldom based on the findings of social scientists, and conversely, the results of dialog systems research are rarely noticed by them.

Yet how people manage joint action is an important scientific question. There is growing interest in models of joint action, social action, and neural coupling in communication (Sebanz, Bekkering, and Knoblich 2006; Marsh, Richardson, and Schmidt 2009; Stephens, Silbert, and Hasson 2010). Language often plays an important role; indeed, echoing the old yo-he-ho theory of language origin, Bangerter and Clark (2003) have argued that "dialogue has its origins in joint activities, which it serves to coordinate." Unfortunately spoken dialogue research so far has produced scant findings about language behavior that are interesting to nonengineers. But some of the challenges discussed here involve fundamental scientific questions about the nature of human communication, so we see an opportunity for the community to adopt one or two as high-profile "grand" challenges, ideally formulated so that they can be addressed, empirically or theoretically, without requiring researchers to develop end-to-end systems (Raux et al. 2006).

There is the related challenge of supporting non-systems dialogue research. Although many people are fascinated by language and dialogue, spoken dialogue systems research has only sporadically tapped this enthusiasm. For example, researchers in the conversation analysis tradition and teachers of foreign languages love to explore patterns of dialogue. The opportunity here is in creating tools to support nontechnical people in discovering things themselves. Even middle-school science fairs should feature studies satisfying curiosity or evaluating hunches about dialogue behaviors. Our community ought to be producing tools and tool sets that support the complete workflow in such studies — eclectically supporting tagging, searching, juxtaposing clips, and so on — and supporting both perceptually based and quantitative analysis in an integrated way. This would not be purely altruistic. While deep learning approaches currently seem to be driving out other methods, the field needs to retain and enhance a wider portfolio of approaches, valuing also modeling techniques that are intrinsically more understandable.

Doing so will ultimately help with major societal challenges. The growing deployment of AI is raising general concerns, mostly about safety and about putting people out of work. While these are also relevant here, dialogue systems bring a further worry: when our machines become truly wonderful dialogue partners, will we still want to talk to other people? Estrangement, alienation, and societal fragmentation are real problems in our society, and technology has been part of the problem (Putnam 2001). But we also see it as a solution; if the fruits of our research include an improved understanding of how good conversationalists communicate well, we can help everyone become more effective in dialogue.

Prospects

Today spoken dialogue systems are in a funny place. On the one hand, most commercial state-of-the-art dialogue systems are an embarrassment for AI. On the other hand, research has produced compelling demonstrations of highly interactive and highly effective behavior. We need to bridge the gap, to produce systems that are more robust, more capable, and more useful.

Figure 3. Erica, An Android Who Ought to Be Able to Have Real Conversations.

We are not the first to list challenges for dialogue systems (Cohen 1997; Zue and Glass 2000; Allen et al. 2001; Bohus, Kamar, and Horvitz 2012; Lison and Meena 2014). But today, the field has advanced beyond the basic challenges in getting things to work. Although the component technologies — speech recognition, understanding, synthesis — still need improvement to support real-time interaction (Buss and Schlangen 2010; Baumann 2013), these are not the main limiting factors. Now the challenge lies in synthesizing component technologies into highly interactive systems.

Dialogue and interactive behaviors are notoriously difficult to visualize, but figures 3 and 4 may help illustrate our points. The android, Erica, is a mechanical marvel with a strong suite of technologies for user tracking, gesture production, speech recognition, speech synthesis, and so on. In contrast to the advanced nature of these technologies, dialogue technology has lagged: making her capable of real dialogues — at the level suggested to users by her appearance, voice, and movements — is clearly a long-term challenge. The brain image highlights the cerebellum (although AI has always preferred to focus on the cerebral) because its functions — including sequencing, prediction, cross-modal binding, temporal coordination, and the integration of "internal representations with external stimuli and self generated responses" — are all things that dialogue systems need to do better (Marien et al. 2014).

Dialogue systems are a very active research area and the field is rapidly maturing. Addressing the

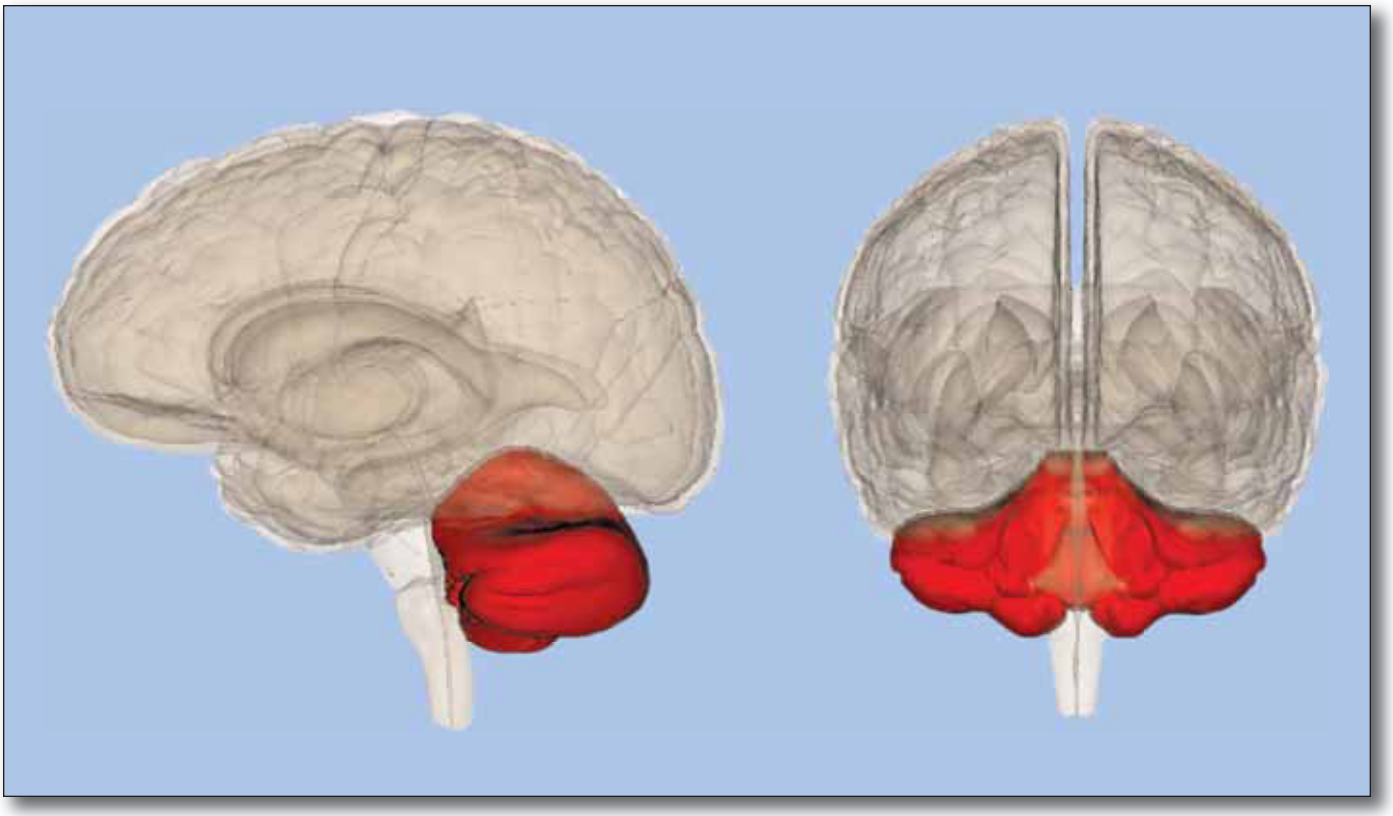

Figure 4. The Cerebellum, an Often-Forgotten Part of the Brain.

challenges noted here will help lead to a future of highly interactive, engaging, and truly effective dialogue systems.

Acknowledgments

We thank Jason D. Williams, Gabriel Skantze, David Novick, Tatsuya Kawahara, Divesh Lala, Pierrick Milhorat, and David Suendermann-Oeft for discussion.

This work was supported by the NSF (IIS-1449093 and IIS-1219253), by the U.S. Army Research, Development, and Engineering Command (RDECOM), and by the Fulbright Program. Figure 3 is courtesy of the JST ERATO Ishiguro Project and Kyoto University's Tatsuya Kawahara, and figure 4 is adapted from the Wikimedia Commons/Life Science Databases. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, position, or policy of the National Science Foundation or the United States government, and no official endorsement should be inferred.

References

Acosta, J. C., and Ward, N. G. 2011. Achieving Rapport with Turn-By-Turn, User-Responsive Emotional Coloring. *Speech Communication* 53(9–10): 1137–1148. dx.doi.org/10.1016/j.specom.2010.11.006

Allen, J. F.; Byron, D. K.; Dzikovska, M.; Ferguson, G.; Galescu, L.; and Stent, A. 2001. Toward Conversational Human-Computer Interaction. *AI Magazine* 22(4): 27–37.

Andrist, S.; Mutlu, B.; and Gleicher, M. 2013. Conversational Gaze Aversion for Virtual Agents. In *Intelligent Virtual Agents: 13th International Conference* (IVA 2013). Lecture Notes in Computer Science 8108. Berlin: Springer. dx.doi.org/10.1007/978-3-642-40415-3_22

Bailly, G.; Mihoub, A.; Wolf, C.; and Elisei, F. 2015. Learning Joint Multimodal Behaviors for Face-To-Face Interaction: Performance and Properties of Statistical Models. Paper presented at the Workshop on Behavior Coordination Between Animals, Humans, and Robots, March 2, Portland, OR.

Bangerter, A., and Clark, H. H. 2003. Navigating Joint Projects with Dialog. *Cognitive Science* 27(2): 195–225. dx.doi.org/10.1207/s15516709cog2702_3

Baumann, T. 2013. Incremental Spoken Dialogue Processing: Architecture and Lower-Level Components. Ph.D. Dissertation, Faculty of Linguistics and Literature, Universität Bielefeld, Bielefeld, Germany.

Bohus, D., and Horvitz, E. 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions. In *Proceedings of the SIGDIAL 2011 Conference, The 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsberg, PA: Association for Computational Linguistics.

Bohus, D.; Kamar, E.; and Horvitz, E. 2012. Towards Situated Collaboration. In *Proceedings of the NAACL-HLT Work-*

shop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data, 13–14. Stroudsburg, PA: Association for Computational Linguistics.

Bunt, H. 2011. Multifunctionality in Dialogue. *Computer Speech and Language* 25(2): 222–245. dx.doi.org/10.1016/j.csl.2010.04.006

Buss, O., and Schlangen, D. 2010. Modelling Sub-Utterance Phenomena in Spoken Dialogue Systems. Paper presented at the 14th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2010), 16–18 June, Poznan, Poland.

Chao, C., and Thomaz, A. L. 2012. Timing in Multimodal Turn-Taking Interactions: Control and Analysis Using Timed Petri Nets. *Journal of Human-Robot Interaction* 1(1): 4–25. dx.doi.org/10.5898/JHRI.1.1.Chao

Clark, H. H. 2002. Speaking in Time. *Speech Communication* 36(1–2): 5–13. dx.doi.org/10.1016/S0167-6393(01)00022-X

Clark, H. H. 2014. Spontaneous Discourse. In *The Oxford Handbook of Language Production,* ed. M. Goldrick, V. S. Ferreira, and M. Miozzo, 292–306. Oxford, UK: Oxford University Press.

Cohen, M. H.; Giangola, J. P.; and Balogh, J. 2004. *Voice User Interface Design.* Boston: Addison-Wesley.

Cohen, P. 1997. Dialogue Modeling. In Cole, R., ed., *Survey of the State of the Art in Human Language Technology,* 204–210. Cambridge, UK: Cambridge University Press.

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.; and Morency, L.-P. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*. New York: Association for Computing Machinery.

DeVault, D.; Sagae, K.; and Traum, D. 2009. Can I Finish? Learning When to Respond to Incremental Interpretation Results in Interactive Dialogue. In *Proceedings of the SIGDIAL 2010 Conference, The 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsberg, PA: Association for Computational Linguistics.

Forbes-Riley, K., and Litman, D. 2011. Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor. *Speech Communication* 53(9–10): 1115–1136. dx.doi.org/10.1016/j.specom.2011.02.006

Foster, M. E.; Keizer, S.; and Lemon, O. 2014. Towards Action Selection Under Uncertainty for a Socially Aware Robot Bartender. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI'14), 158–159. New York: Association for Computing Machinery. dx.doi.org/10.1145/2559636.2559805

Fujie, S.; Fukushima, K.; and Kobayashi, T. 2005. Back-Channel Feedback Generation Using Linguistic and Nonlinguistic Information and Its Application to Spoken Dialogue System. In *Interspeech 2005: Eurospeech, 9th European Conference on Speech Communication and Technology,* 889–892. Grenoble, France: International Speech Communication Association.

Gasic, M.; Mrksic, N.; Su, P.; Vandyke, D.; Wen, T.-H.; and Young, S. J. 2015. Policy Committee for Adaptation in Multi-Domain Spoken Dialogue Systems. In *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU 2015) 806–812. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/asru.2015.7404871

Geiger, J. T.; Eyben, F.; Schuller, B.; and Rigoll, G. 2013. Detecting Overlapping Speech with Long Short-Term Memory Recurrent Neural Networks. In *Interspeech 2013: 14th Annual Conference of the International Speech Communication Association,* 1668–1672. Grenoble, France: International Speech Communication Association.

Ghigi, F.; Eskenazi, M.; Torres, M. I.; and Lee, S. 2014. Incremental Dialog Processing in a Task-Oriented Dialog. In *Interspeech 2014: 15th Annual Conference of the International Speech Communication Association,* 308–312. Grenoble, France: International Speech Communication Association.

Goldwasser, D., and Daume III, H. 2014. I Object: Modeling Latent Pragmatic Effects in Courtroom Dialogues. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* 655–663. Stroudsburg, PA: The Association for Computer Linguistics

Gratch, J.; Wang, N.; Okhmatovskaia, A.; Lamothe, F.; Morales, M.; van der Werf, R.; and Morency, L.-P. 2007. Can Virtual Humans Be More Engaging Than Real Ones? In *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments: 12th International Conference,* Lecture Notes in Computer Science 4552, 286–297. Berlin: Springer. dx.doi.org/10.3115/v1/E14-1069

Grothendieck, J.; Gorin, A. L.; and Borges, N. M. 2011. Social Correlates of Turn-Taking Style. *Computer Speech and Language* 25(4): 789–801. dx.doi.org/10.1016/j.csl.2011.01.002

Huang, C.-M., and Mutlu, B. 2014. Learning-Based Modeling of Multimodal Behaviors for Humanlike Robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (HRI'14), 57–64. New York: Association for Computing Machinery. dx.doi.org/10.1145/2559636.2559668

Huang, L.; Morency, L.-P.; and Gratch, J. 2010. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multi-Agent Systems*. New York: Association for Computing Machinery.

Kalchbrenner, N., and Blunsom, P. 2013. Recurrent Continuous Translation Models. In *Proceedings of the Empirical Methods in Natural Language Processing* (EMNLP 2013). Stroudsberg, PA: Association for Computational Linguistics.

Kawahara, T.; Uesato, M.; Yoshino, K.; and Takanashi, K. 2015. Toward Adaptive Generation of Backchannels for Attentive Listening Agents. Paper presented at the 2015 International Workshop on Spoken Dialogue Systems Technology, 11–13 January, Busan, South Korea.

Kim, D.; Breslin, C.; Tsiakoulis, P.; Gasic, M.; Henderson, M.; and Young, S. 2014. Inverse Reinforcement Learning for Micro-Turn Management. In *Interspeech 2014: 15th Annual Conference of the International Speech Communication Association*. Grenoble, France: International Speech Communication Association.

Lee, Y.; Wampler, K.; Bernstein, G.; Popovic, J.; and Popovic, Z. 2014. Motion Fields for Interactive Character Locomotion. *Communications of the ACM* 57(6): 101–108. dx.doi.org/10.1145/2602758

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. Unpublished preprint. Technical Report arXiv:1603.06155 [cs.CL]. Ithaca, NY: Cornell University Library.

Li, L.; He, H.; and Williams, J. D. 2014. Temporal Supervised

Learning for Inferring a Dialog Policy from Example Conversations. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop* (SLT 2014), 312–317. Piscataway, NJ: Institute for Electrical and Electronics Engineers. dx.doi.org/10.1109/SLT.2014.7078593

Lison, P., and Meena, R. 2014. Spoken Dialogue Systems: The New Frontier in Human-Computer Interaction. *XRDS: Crossroads, The ACM Magazine for Students* 21(1): 46–51. dx.doi.org/10.1145/2659891

Marien, P.; Ackermann, H.; Adamaszek, M.; Barwood, C. H.; Beaton, A.; Desmond, J.; De Witte, E.; Fawcett, A. J.; Hertrich, I.; Küper, M.; Leggio, M.; Marvel, C.; Molinari, M.; Murdoch B. E.; Nicholson, R. I.; Schmahmann, J. D.; Stoodley, C. J.; Thürling, M.; Timmann, D.; Wouters, E.; and Ziegler, W. 2014. Consensus Paper: Language and the Cerebellum: An Ongoing Enigma. *Cerebellum* 13(3): 386–410.

Marsh, K. L.; Richardson, M. J.; and Schmidt, R. C. 2009. Social Connection Through Joint Action and Interpersonal Coordination. *Topics in Cognitive Science* 1(2): 320–339. dx.doi.org/10.1111/j.1756-8765.2009.01022.x

Meena, R.; Skantze, G.; and Gustafson, J. 2014. Data-Driven Models for Timing Feedback Responses in a Map Task Dialogue System. *Computer Speech and Language* 28(4): 903–922. dx.doi.org/10.1016/j.csl.2014.02.002

Möller, S., and Ward, N. 2008. A Framework for Model-Based Evaluation of Spoken Dialog Systems. In *Proceedings of the SIGDIAL 2006 Workshop, The 7th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsberg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/1622064.1622099

Möller, S.; Engelbrecht, K.-P.; and Schleicher, R. 2008. Predicting the Quality and Usability of Spoken Dialog Services. *Speech Communication* 50(8–9): 730–744. dx.doi.org/10.1016/j.specom.2008.03.001

Nass, C., and Brave, S. 2007. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA: MIT Press.

Nijholt, A.; Reidsma, D.; van Welbergen, H.; op den Akker, R.; and Ruttkay, Z. 2008. Mutually Coordinated Anticipatory Multimodal Interaction. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference,* Lecture Notes in Computer Science Volume 5042, 70–89. Berlin: Springer.

Paetzel, M.; Manuvinakurike, R.; and DeVault, D. 2015. So, Which One Is It? The Effect of Alternative Incremental Architectures in a High-Performance Game-Playing Agent. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsberg, PA: Association for Computational Linguistics.

Pickering, M. J., and Garrod, S. 2013. An Integrated Theory of Language Production and Comprehension. *Behavioral and Brain Sciences* 36(4): 329–347. dx.doi.org/10.1017/S0140525X12001495

Pieraccini, R.; Suendermann, D.; Dayanidhi, K.; and Liscombe, J. 2009. Are We There Yet? Research in Commercial Spoken Dialog Systems. In *Proceedings of Text, Speech and Dialogue, 12th International Conference* (TSD 2009). Lecture Notes in Computer Science 5729, 3–13. Berlin: Springer. dx.doi.org/10.1007/978-3-642-04208-9_3

Putnam, R. D. 2001. *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Ranganath, R.; Jurafsky, D.; and McFarland, D. 2013. Detecting Friendly, Flirtatious, Awkward, and Assertive Speech in Speed-Dates. *Computer Speech and Language* 27(1): 89–115. dx.doi.org/10.1016/j.csl.2012.01.005

Raux, A., and Eskenazi, M. 2012. Optimizing the Turn-Taking Behavior of Task-Oriented Spoken Dialog Systems. *ACM Transactions on Speech and Language Processing* (TSLP) 9(1): 1–23. dx.doi.org/10.1145/2168748.2168749

Raux, A., and Nakano, M. 2010. The Dynamics of Action Corrections in Situated Interaction. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (SIGDIAL '10), 165–174. Stroudsberg, PA: Association for Computational Linguistics.

Raux, A.; Bohus, D.; Langner, B.; Black, A. W.; and Eskenazi, M. 2006. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go Experience. In *Interspeech 2006: Ninth International Conference on Spoken Language Processing*. Grenoble, France: International Speech Communication Association.

Schaal, S. 1999. Is Imitation Learning the Route to Humanoid Robots? *Trends in Cognitive Sciences* 3(6): 233–242. dx.doi.org/10.1016/S1364-6613(99)01327-3

Schröder, M.; Bevacqua, E.; Cowie, R.; Eyben, F.; Gunes, H.; Heylen, D.; ter Maat, M.; Gary; Pammi, S.; Pantic, M.; Pelachaud, C.; Schuller, B.; de Sevin, E.; Valstar, M.; and Wollmer, M. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing* 3(2): 165–183. dx.doi.org/10.1109/T-AFFC.2011.34

Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; and Narayanan, S. 2013. Paralinguistics in Speech and Language: State-of-the-Art and the Challenge. *Computer Speech and Language* 27(1): 4–39. dx.doi.org/10.1016/j.csl.2012.02.005

Sebanz, N.; Bekkering, H.; and Knoblich, G. 2006. Joint Action: Bodies and Minds Moving Together. *Trends in Cognitive Sciences* 10(2): 70–76. dx.doi.org/10.1016/j.tics.2005.12.009

Skantze, G.; Hjalmarsson, A.; and Oertel, C. 2014. Turn-Taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication* 65(Nov.–Dec.): 50–66. dx.doi.org/10.1016/j.specom.2014.05.005

Stephens, G. J.; Silbert, L. J.; and Hasson, U. 2010. Speaker-Listener Neural Coupling Underlies Successful Communication. *Proceedings of the National Academy of Sciences* 107(32): 14425–14430. dx.doi.org/10.1073/pnas.100866 2107

Ward, N., and Tsukahara, W. 1999. A Responsive Dialog System. In *Machine Conversations, The Springer International Series in Engineering and Computer Science,* volume 511, ed. Y. Wilks, 169–174. Dortrecht, The Netherlands: Kluwer. dx.doi.org/10.1007/978-1-4757-5687-6_14

Ward, N. G. 2014. Automatic Discovery of Simply-Composable Prosodic Elements. In *Proceedings of the 7th Speech Prosody Conference,* 915–919. Stroudsburg, PA: Association for Computational Linguistics.

Ward, N. G., and Abu, S. 2016. Action-Coordinating Prosody. In *Proceedings of the 9th Speech Prosody Conference*. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.21437/speechprosody.2016-129

Watanabe, S.; Hershey, J. R.; Marks, T. K.; Fujii, Y.; and Koji, Y. 2014. Cost-level Integration of Statistical and Rule-Based Dialog Managers. In *Interspeech 2014: Proceedings of the 15th Annual Conference of the International Speech Communication Association,* 323–327. Grenoble, France: International Speech Communication Associationn.

Weyand, T.; Kostrikov, I.; and Philbin, J. 2016. PlaNet: Photo Geolocation with Convolutional Neural Networks. Unpublished preprint arXiv:1602.05314 [cs.CV]. Ithaca, NY: Cornell University Library.

Williams, J. D.; Henderson, M.; Raux, A.; Thomson, B.; Black, A.; and Ramachadran, D. 2014. The Dialog State Tracking Challenge Series. *AI Magazine* 35(4): 121–123.

Xu, Y. 2011. Speech Prosody: A Methodological Review. *Journal of Speech Sciences* 1(1): 85–115.

Young, S.; Gasic, M.; Thomson, B.; and Williams, J. D. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE* 101(5): 1160–1179. dx.doi.org/10.1109/JPROC.2012.2225812

Yu, Z.; Bohus, D.; and Horvitz, E. 2015. Incremental Coordination: Attention-Centric Speech Production in a Physically Situated Conversational Agent. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Stroudsberg, PA: Association for Computational Linguistics.

Zue, V. W., and Glass, J. R. 2000. Conversational Interfaces: Advances and Challenges. *Proceedings of the IEEE* 88(8): 1166–1180. dx.doi.org/10.1109/5.880078

Nigel G. Ward is a professor of computer science at the University of Texas at El Paso. He received his Ph.D. from the University of California at Berkeley in 1991. Ward's research areas are at the intersection of spoken language and human-computer interaction. He is currently working on modeling the prosodic patterns of interaction in dialogue, and on applying these models to improve turn-taking in dialogue systems, to support pragmatics-aware information retrieval, to make telecommunications more efficient, and to help reduce cross-cultural misunderstandings. His research has been supported by the NSF, DARPA, the Fulbright Program, and the Japanese Ministry of Education, among other industrial and government sponsors.

David DeVault is a research assistant professor at the USC Institute for Creative Technologies and the Department of Computer Science at the University of Southern California. He received his Ph.D. from the Department of Computer Science at Rutgers University in 2008. His research lies in the area of natural language dialogue systems. He has research interests in incremental speech processing, turn-taking, natural language understanding, and dialogue management. His work aims to provide dialogue systems with the ability to understand, predict, and respond to human speech in real-time, and to do so in a way that is robust to the inevitable uncertainties of communication. His research has been funded by the National Science Foundation (NSF) and the Army Research Office (ARO). He is the author of more than 50 technical articles.