On the Challenges and Opportunities of Physically Situated Dialog

Dan Bohus and Eric Horvitz

Microsoft Research
One Microsoft Way
Redmond, WA, 98052
{dbohus,horvitz}@microsoft.com

Abstract

We outline several challenges and opportunities for building physically situated systems that can interact in open, dynamic, and relatively unconstrained environments. We review a platform and recent progress on developing computational methods for situated, multiparty, open-world dialog, and highlight the value of representations of the physical surroundings and of harnessing the broader situational context when managing communicative processes such as engagement, turn-taking, language understanding, and dialog management. Finally, we outline an open-world learning challenge that spans these different levels.

Introduction

Most research to date in spoken dialog systems has focused on dyadic interactions over a speech- or text-only channel, with the telephony-based interactive voice response system (IVR) as the prototypical application. Various interactional problems have been investigated in this context, and advances in areas like speech recognition and synthesis, language understanding and generation, and dialog management have led to wide-scale deployments and use of IVR and multimodal mobile systems in daily life.

At the same time, the goal of developing autonomous systems (such as robots) that act and interact in the open world via spoken language is still in its early stages and raises significant research challenges. Interactions in the open-world are characterized by several aspects which represent key departures from assumptions typically made in spoken dialog systems. First open-world interactions are *physically situated*: the surrounding environment provides a rich, continuously streaming physical context often critical to understanding and organizing communications. Open world interactions are also typically *multiparty* and *dynamic*: the world contains not just one, but multiple actors that may be relevant to the computational system, and each actor has their own set of evolving goals, desires and intentions.

The *physically situated*, *dynamic*, and *multiparty* nature of open-world interactions brings to fore new challenges to the traditional spoken dialog processing stack. Managing communicative processes like engagement, turn-taking, language understanding, and dialog management in an open-world setting requires integrative reasoning that goes beyond the confines of language problems and takes into account the broader *situational context*: the who, where, what, and why of the scene (Figure 1.) At the low level, this includes basic physical awareness and reasoning about relevant actors, objects, and events in the environment, their location, physical characteristics and relationships, topologies and trajectories, etc. At a higher level, it includes semantic context about the (changing) activities that the human and computational actors are involved in, and about the long-term goals, plans, desires, knowledge, and intentions that are driving these activities. As an example, consider the problem of establishing an open communication channel—the first step in any successful language interaction. In traditional dialog systems, this problem is simply resolved via unambiguous signals such as a call being received in a telephony IVR system, or a push-to-talk button. These solutions are however inadequate for systems that operate continuously in open, dynamic environments, where participants might come and go, initiate interactions from a distance, interact with each other and with the system and interleave their

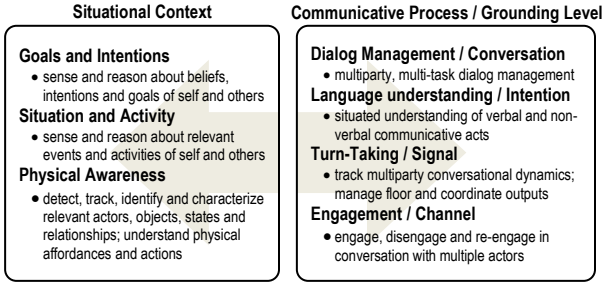


Figure 1. Components for reasoning in support of dialog for communication and joint activity in the open world.

communications with other activities. New models that explicitly reason about spatiotemporal trajectories, proxemics and geometric relationships in formations of people, non-verbal behaviors, body pose and the dynamics of gaze and eye contact, as well as higher level inferences about the long-term goals and activities of each agent are required in order to fluidly manage the engagement process in an open-world setting. Similar challenges in integrating the streaming physical and semantic context are raised for the other conversational competencies like turn-taking, language understanding, and dialog management.

In this paper, we review some of these challenges and opportunities in more detail, summarize our initial research efforts in this space, and outline directions for future work. We begin by describing a set of hardware and software components, and a number of applications that provide the test-bed for the work described in the sequel.

A Research Platform for Situated Interaction

Our hardware platform consists of a custom-assembled multi-modal kiosk shown in Figure 2. The sensory apparatus of this prototype includes a wide-angle camera with 140° field of view and a resolution of 640x480 pixels; a 4-element linear microphone array that can provide sound-source localization information in 10° increments; a 19" touch-screen; and a RFID badge reader. The output consists of a talking avatar head with controllable pose, synchronized lip movements and limited facial gestures.

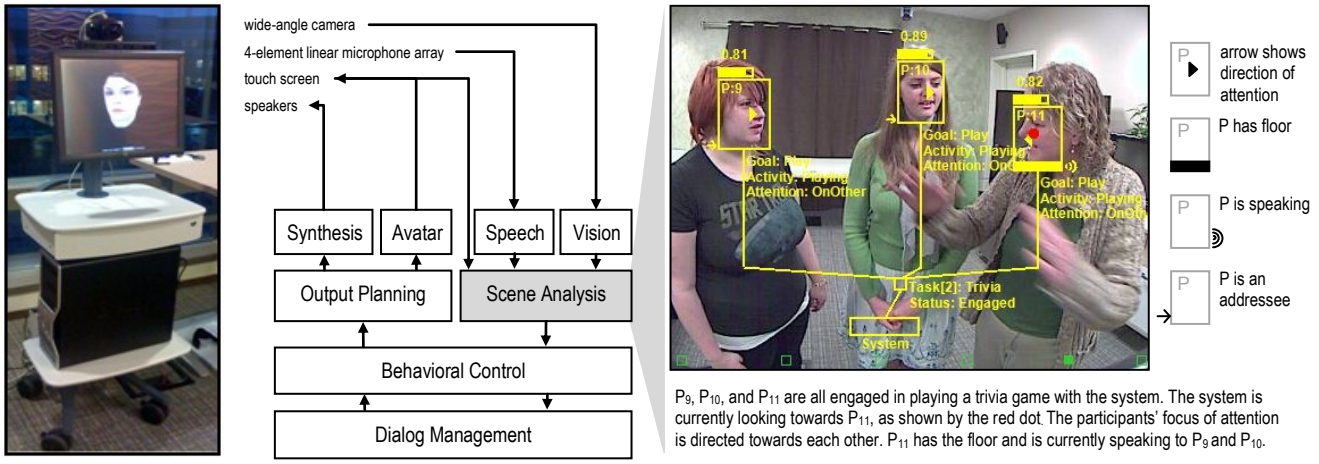
Data gathered by the sensors is forwarded to a scene analysis module that fuses the incoming streams and constructs (in real-time) a coherent picture of what is happening in the surrounding environment. This includes models for detecting and tracking the location of multiple actors in the scene, reasoning about their attention, activities, goals and relationships (*e.g.* which people are in a group together), reasoning about engagement (*e.g.* tracking engagement states, actions and intentions), and turn-taking (e.g. tracking who is talking to whom, who has the floor, etc.) The results of this real-time scene analysis (some of them illustrated in Figure 3) are forwarded to a reactive control layer, which orchestrates the avatar's behaviors based on the semantic outputs planned by a multiparty dialog management component. A more detailed description of these components is available in [4].

These software components are implemented on top of the Microsoft Robotics Studio platform [14], which facilitates concurrent, coordinated computation. Given its degree of modularization and abstraction, we expect that our software stack can be easily adapted to different hardware platforms, including robotic systems.

To date, we have developed several applications using this framework [4]. Videos of recorded interactions are available in [21]. The first application we have developed, called *Receptionist*, implements a situated conversational agent that makes shuttle reservations, a task performed by front-desk receptionists at our campus. The system can handle multi-participant interactions and can manage interleaved engagements with multiple parties *e.g.* by interrupting a conversation to address a waiting customer and let them know they will be attended to momentarily.

Another application, the *Questions Game,* implements a situated conversational agent that can engage with one or multiple participants and challenge them to answer questions on a variety of topics. In multi-participant situations, the system can monitor side conversations, and, once it receives an answer, also seeks confirmation from other engaged participants before moving on. In addition, the system can attract bystanders and engage them in an already running game. This application has served as the primary platform for the research described below.

A third application, *PASS*, or the *Personal Assistant and Scheduling System* is designed to act as an administrator with the ability to handle messaging and scheduling tasks. The system is deployed outside its owners' office and has access to their calendar and to components that learn and reason about the owner's presence and availability. By leveraging this back-end information, *PASS* can engage people that approach the office when the owner is not present and provide assistance with scheduling meetings and relaying messages.



P_9, P_{10}, and P_{11} are all engaged in playing a trivia game with the system. The system is currently looking towards P_{11}, as shown by the red dot. The participants' focus of attention is directed towards each other. P_{11} has the floor and is currently speaking to P_9 and P_{10}.

Figure 2. Prototype, software components, and automated annotation of conversational scene analysis

Research Challenges

We now turn our attention to the set of core conversational competencies for enabling physically situated interaction. We review our previous efforts on modeling engagement [2, 3] and turn-taking [5, 6] in multiparty, open-world settings and we discuss lessons learned and future work. We then review challenges at the higher levels in the stack of conversational competencies, in the areas of situated language understanding and dialog management.

Engagement

We follow the definition by Sidner et al. [20] and view *engagement* as "the process by which […] participants establish, maintain and end their perceived connection during interactions they jointly undertake". Research in sociolinguistics and conversational analysis [9,11,12,13] has shown that this is a mixed-initiative, coordinated process, managed via a multitude of verbal, non-verbal and spatial cues such as trajectory and proximity, gaze and mutual attention, head and hand gestures, salutations etc.

In contrast, traditional spoken dialog systems rely on simple solutions like push-to-talk buttons to resolve this problem. Such solutions are inadequate for systems that need to seamlessly engage, disengage, re-engage with multiple parties in an open-world setting. To address this challenge, we have outlined a modeling framework [2] that enables an embodied interactive system to explicitly represent and reason about engagement. The proposed framework harnesses components for sensing the engagement state, actions and intentions of various actors in the scene (*e.g.* who is engaged, who is trying to engage, etc.), for making engagement control decisions (*e.g.*, whom should a system engage with, and when?) and for controlling the verbal and non-verbal behaviors of the agent such as to convey its own engagement intentions.

These models were implemented and evaluated in the context of the *Questions Game* application. We deployed the application near a kitchenette in our building and conducted a 4-week in-the-wild experiment in which the system could make eye contact and then invite the passer-bys that slowed and approached it to play the game; no instructions for how to interact with the system were provided. Additionally, activity models based on the spatiotemporal trajectory of actors in the scene were used to determine whether bystanders were present in the scene while a game was in progress. If bystanders were detected, the system would temporarily suspend the existing interaction and create a side-engagement with the bystander to get them to join the game (*e.g.* towards engaged participant: *'It looks like you could use some help. Excuse me for one second'*, then towards bystander: *'Hi, would you like to join in?'*). If the bystander agreed, the system continued playing the game with both participants. Experimental results [2] indicate that the proposed models enabled the system to successfully create multi-participant engagements in this setting. Overall, bystanders successfully recognized that they are being engaged and solicited by the system and responded (either positively or negatively) in 87% of cases. The side comments produced by the participants around the moments of engagement indicated surprise and excitement at the system's multiparty capabilities.

In [3] we proposed and evaluated an approach for learning to make inferences about when an actor might be interested in engaging with the system by leveraging the temporal structure of spatial and focus-of-attention features for that particular actor. The proposed approach does not require any explicit supervision, and allows a system to learn from interaction, in a previously unseen environment. The central idea is to start with a conservative detector for engagement intentions, such as a push-to-engage button, and automatically gather sensor data surrounding the moments of engagement, together with labels that indicate whether someone actually engaged or not (the system eventually finds out if an actor becomes engaged: the actor will either start playing the game, or will eventually disappear from view). Experiments were conducted in two different spatial orientations (see Figure 3), which create considerable differences in the relative trajectories of people that go by (dashed lines) and people that engage with the system (continuous lines). The results indicate that the system was able to learn to predict engagement

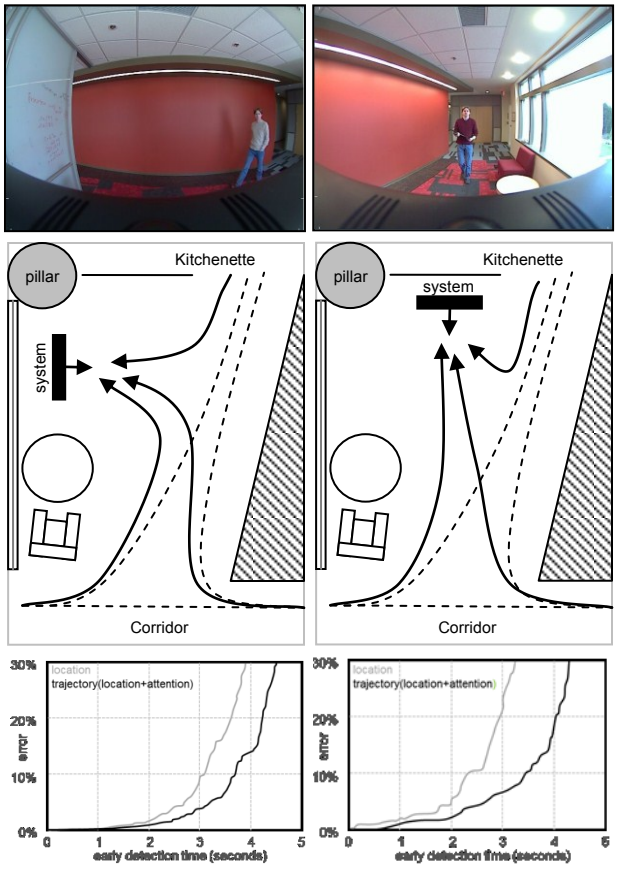


Figure 3. Placement, fields of view, and error rates for detecting engagement intentions in two spatial orientations.

intentions at low error rates up to 3 seconds prior to the actual moment of engagement. Figure 3 shows the error rate as a function of how early the predictions are being made. The learned models were different across the two orientations, as they were adapted to the *specifics of the spatial configuration* the system was placed in. These results highlight the importance of physical context in situated interaction, and raise questions regarding the development of representations that best capture this context in a manner that generalizes across situations.

Similar models are required for detecting whether engaged participants are actively maintaining the conversational engagement or are disengaging. Apart from tracking and reasoning about physical space as well as verbal and non-verbal cues and gestures, such as affirmations, gaze and attention, body position, etc., these inferences could also significantly benefit from a tighter integration with higher level context. This includes information about goals, intentions and activities (*e.g.*, what role does the participant play in the current collaboration? is the participant involved in another non-communicative, but still on-task activity?) as well as information from the higher levels in the dialog processing stack. Examples of the latter include the turn-taking context (*e.g.,* was the last utterance produced by the participant addressed to the system or to someone going by?), language understanding (*e.g.*, did the last utterance bring a contribution to the current interaction?), and dialog management (*e.g.*, what is the expected rhythm of contributions and the pace of the interaction at this point?).

Important challenges also remain in making engagement control decisions in open-world settings. Consider for instance the problem of optimizing engagements with multiple parties, who all desire access to a single point-of-service system (*e.g.*, *Receptionist*). Or consider the problem in which a system has to engage with a participant already engaged in another task or conversation. We believe such problems require decision-theoretic solutions that draw on a deep understanding of the current context, including the goals and tasks at hand, and take into account the underlying uncertainties, the costs of continuing versus interrupting collaborations, as well as notions of conversational etiquette, expectations of fairness, etc.

Interesting challenges lay ahead on the creation of accurate low-level behavioral models, including the fine-grained control of pose, gesture, facial expressions, etc. Mobility adds yet another dimension to the problem, as the behavioral control components need to also reason deeply about trajectories, proxemics and the structure of *f-formations*, and more generally about the etiquette of space and movement in interaction. Developing such methods will likely have subtle, yet powerful influences on the effectiveness and fluidity of the engagement process.

Multi-Participant Turn-Taking

Once participants are engaged in a conversation, given the serial nature of the verbal channel, they have to coordinate with each other on the presentation and recognition of various communicative signals. This happens in a process known as *turn-taking*, which, like engagement, is regulated through a rich set of verbal and non-verbal cues, such as establishing eye contact, head and hand gestures, changes in prosody and verbal affirmations [1,8,10,17,18,19].

With a few exceptions, *e.g.* [15,16,22,23], most spoken dialog systems have been designed for closed-world dyadic interactions and make a simplifying "you speak then I speak" assumption. This can often lead to breakdowns, even in dyadic interactions, and various heuristics are used to handle departures from this expected volley of interaction, such as user barge-ins or time-outs. The inadequacy of simple heuristics for guiding turn-taking is even more salient in multiparty settings, where multiple participants vie for the floor and may address contributions to the system or to each other, and where events external to the conversation can impinge on the urgency of a participants' need to make a contribution.

In [5], we outlined a computational framework for modeling and managing turn-taking in open-world spoken dialog systems. We take the view that turn-taking is a collaborative, interactive process by which participants in a conversation monitor each other and take coordinated actions in order to ensure that (generally) only one participant speaks at a given time—that participant is said to have the *conversational floor*. Furthermore, we assume floor shifts from one participant to another emerge as a result of joint, coordinated *floor management actions* performed by the participants: *Hold*, *Take*, *Release* and *Null*. The proposed framework subsumes models for tracking the conversational dynamics in multiparty interaction, for making floor management decisions, and for rendering these decisions into appropriate behaviors.

The sensing subcomponent in the proposed framework is responsible for tracking the conversational dynamics, *i.e.* identifying spoken signals, their source and target (the framework represents various addressee roles as per [7] – see Figure 4 and Table 1), and the floor state, actions and intentions for each engaged participant. The decision component decouples input processing from response generation and floor control decisions: all inputs are processed as soon as they are detected. However the decisions to generate a new contribution and the selection of the floor management actions to be performed by the system are made separately, based on rules that take into account the larger turn-taking and dialog context (*i.e.*, floor state, actions and intentions for each participant in the scene, set of planned outputs, etc.). Finally, the system's floor management actions are rendered into a set of accurately timed verbal and non-verbal behaviors (*e.g.*, establishing and breaking eye contact, lifting eyebrows, etc.) that convey the system's turn-taking intentions.

The proposed models were implemented and evaluated via a set of multi-participant interaction experiments with the *Questions Game* [5]. Results indicate that the proposed framework can indeed enable the conversational agent to

Role	Description
Addressee	participant that utterance is addressed to
Side participant	participant that utterance is not addressed to
Overhearer	others known to the speaker who are not participants in conversation but will hear the utterance
Eavesdropper	others not known to the speaker who are not participants in the conversation but will hear the utterance

Table 1. Addressee roles in multiparty interaction.

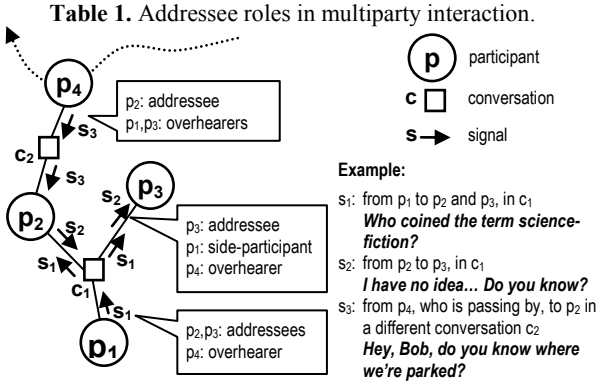


Figure 4. Sample multiparty interaction with illustrated addressee roles and three different signals.

participate in multi-participant interactions, and to handle a diversity of naturally occurring turn-taking phenomena, including multi-participant floor management, barge-ins, restarts, and continuations. Users rated the system's multiparty turn-taking abilities favorably in a post-experiment subjective assessment questionnaire [5].

The data collected in these experiments also shows that the current behavioral models allow the avatar to effectively shape turn allocation and convey addressee roles in multi-participant interactions [6]. For instance, in interactions involving two participants and the system, during verbally produced *RequestConfirmation* dialog acts that were addressed to a single participant (*e.g. 'Is that correct?'*) the addressee designated by the system (via gaze) was the one to respond in 86.2% of cases. Even when the *RequestConfirmation* dialog act was performed in an entirely non-verbal manner (*i.e.*, by simply gazing towards the addressee and lifting eyebrows), the designated addressee was the one to respond in 78.6% of cases. A detailed discussion of these results, including an analysis of various contextual factors that also impact the system's ability to shape turn-allocation is available in [6].

As with engagement, numerous challenges remain with respect to turn-taking. Perhaps key among them is developing robust models for tracking conversational dynamics. We initially used handcrafted heuristic models for making inferences about turn-taking from audiovisual information, but have been exploring the promise of building richer models from case libraries of data that are used to perform joint inferences about all participants in the scene. We expect that richer audio-visual and physical context (*e.g.*, prosody, head and body pose), temporal context (*e.g.* who spoke last, how long ago), as well as high-level interactional context (*e.g.*, what turn-taking expectations does the current dialog state create, where is the avatar looking, etc.) all carry relevant information for these inferences. In addition, given the key role of timing in turn-taking, we believe that inference models that not only track, but also anticipate and predict floor and turn-taking events, can significantly improve performance and enable more fluid turn-taking.

Given the underlying uncertainties in the signal and floor inference models, we believe that utility-theoretic methods that resolve trade-offs, between timely actions and greater accuracies promised by delays to collect additional audio-visual evidence, can provide more robust performance. Detection and recovery from turn-taking errors is another important area of future research. We note that turn-taking errors generally manifest themselves as overlaps or long unfilled pauses. Coupled with appropriate blame-assignment models, these signals may provide useful online cues for learning or adaptation.

Finally, important challenges remain in developing more refined behavioral models for signaling turn-taking actions and intentions, *e.g.*, modulating gaze and prosody on the fly, producing backchannels, etc.

Language Understanding and Dialog Management

Numerous challenges remain also in the areas of spoken language understanding and dialog management. Novel mechanisms for integrating the streaming physical context into the typically discrete language understanding and dialog planning processes are required. For instance, the *Receptionist* uses information about the number of actors present in the scene and their relationships (*i.e.* who is in a group together) to make inferences about the number of people that a particular shuttle reservation should be made for. The system's belief updating process fuses such continuously streaming information with discrete discourse contributions from the participants. Similarly, resolving deictic expressions like 'Come here!' requires language understanding models anchored in spatial reasoning and a deep understanding of the relevant entities (and their relationships) in the surrounding environment.

New formalisms may be required for dialog management of open-world, mixed-initiative, multiparty interactions. Handling such interactions requires discourse understanding models that reason more deeply about addressee roles, and about how contributions from multiple participants can be interleaved. These models in turn can be anchored in an analysis of the roles played by each participant in the interaction, as well as their knowledge, goals, and intentions. Finally, open-world systems must be able to reason beyond the confines of any single interaction in order to provide continuous, long-term assistance.

Towards Open-World Learning

We have highlighted some of the specific challenges of integrating streaming situational context with various conversational competencies in support of fluid spoken language interaction in the open-world. We conclude by

outlining a more generic challenge that cuts across all these different components: *open-world learning*.

We believe that developing deeper competencies with open-world dialog will hinge on the incorporation of deeper domain-specific skills as well as key aspects of cross-domain commonsense knowledge about intentions, goals, activities, and about objects in the physical world. Higher-level processes like spoken language recognition, understanding, and dialog management are tightly anchored in both general commonsense and domain-specific competencies, as captured by lexicons and grammars, knowledge about ontologies of objects and affordances in the world, and the abilities to infer goals, perform plan recognition, and to create dialog plans. Typically, interactive applications are developed by carefully defining boundaries and by engineering the knowledge and models required to provide good coverage in a particular domain. The engineering costs can be simplified to some degree by creating reusable components that decouple domain-specific from domain-independent aspects: the latter can be reused across applications. Nevertheless, this approach tends to produce brittle solutions susceptible to "out-of-domain" problems (*e.g.*, out-of-vocabulary, out-of-grammar, out-of-understanding, etc.). These problems become even more acute for systems that operate in the open, unconstrained world.

We envision a two-step solution to this problem. The first step involves the development of models for detecting, explicitly reasoning about, and diagnosing out-of-domain situations. The second step involves learning to extend models and domains with new knowledge in a lifelong, ongoing manner. This can be accomplished in small increments, by special machinery for autonomously probing and learning about situations and phenomena that are noted as poorly understood, as well as via leveraging and seeking help through interactions, or from a domain expert. Open problems include modeling uncertainties about "unknown unknowns," developing representations that are expressive yet easily support extensions, developing models for eliciting domain knowledge through interaction, and developing models for sharing and fusing the knowledge learned online by different systems. We believe that taking steps to solve these hard problems can lead to increased robustness and lower engineering costs, and ultimately move us closer to a long-standing dream in the AI community: systems that can continuously learn and improve themselves through experience. We challenge the research community to innovate in these areas.

References

[1] M. Argyle and M. Cook, 1976, Gaze and Mutual Gaze, Cambridge University Press, New York

[2] D. Bohus, and E. Horvitz, 2009. *Models for Multiparty Engagement in Open-World Dialog*, in Proc. of SIGdial'09, London, UK.

[3] D. Bohus, and E. Horvitz, 2009. *Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings*, in Proc. SIGdial'09, London, UK

[4] D. Bohus, and E. Horvitz, 2009. *Dialog in the Open-World: Platform and Applications*, in Proceedings of ICMI'09, Boston, MA

[5] D. Bohus, and E. Horvitz, 2010. *Computational Models for Multiparty Turn-Taking*, Technical Report, Microsoft Research.

[6] D. Bohus, and E. Horvitz, 2010. *Facilitating Multiparty Dialog with Gaze, Gesture and Speech*, in Proceedings of ICMI'10, Beijing, China.

[7] H. Clark, and T. Carlson, 1982. *Hearers and speech acts,* Language, 58, 332-373.

[8] S. Duncan, 1972. Some Signals and Rules for Taking Speaking Turns in Conversation, Journal of Personality and Social Psychology 23, 283-292.

[9] E. Goffman, 1963, Behaviour in public places: notes on the social order of gatherings, The Free Press, New York

[10] C. Goodwin, 1980. Restarts, pauses and the achievement of mutual gaze at turn-beginning, Sociological Inquiry, 50(3-4), 272-302.

[11] E.T. Hall, 1966, The Hidden Dimension: man's use of space in public and private, New York: Doubleday.

[12] A. Kendon, 1990a, A description of some human greetings, Conducting Interaction: Patterns of behavior in focused encounters, Cambridge University Press

[13] A. Kendon, 1990b, Spatial organization in social encounters: the F-formation system, Conducting Interaction: Patterns of behavior in focused encounters, Studies in International Sociolinguistics, Cambridge University Press

[14] Microsoft Robotics Studio web-site – http://msdn.microsoft.com/en-us/robotics/default.aspx

[15] A. Raux, and M. Eskenazi, 2008. Optimizing end-pointing thresholds using dialogue features in a spoken dialogue system, in Proc. SIGdial-2008, Columbus, OH.

[16] A. Raux, and M. Eskenazi, 2008. *A Finite-State Turn-Taking Model for Spoken Dialog Systems,* in Proc. HLT-2009, Boulder, CO.

[17] H. Sacks, E. Schegloff, and G. Jefferson, 1974. A simplest systematics for the organization of turn-taking in conversation, Language, 50, 696-735.

[18] E. Schegloff, 2000. *Accounts of Conduct in Interaction: Interruption, Overlap and Turn-Taking,* The handbook of sociological theory, 287-321, New York: Plenum.

[19] E. Schegloff, 2000. *Overlapping talk and the organization of turn-taking in conversation*, Language in Society, 29, 1-63.

[20] C.L. Sidner, C. Lee, C.D. Kidd, N. Lesh, and C. Rich, 2005, Explorations in engagement for humans and robots, Artificial Intelligence, 166 (1-2), pp. 140-164

[21] Situated Interaction, project web-page - http://research.microsoft.com/~dbohus/research_situated_interaction.html

[22] K.R. Thorisson, 2002. *Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perceptions to Action*, Multimodality in Language and Speech Systems, Kluwer Academic Publishers, 173-207.

[23] D. Traum, and J. Rickel, 2002. Embodied Agents for Multi-party Dialogue in Immersive Virtual World, in Proc. AAMAS-2002, 766-773.

[24] J. Wiemann, and M. Knapp, 1975. *Turn-taking in conversation*, Journal of Communication, 25, 75-92.