Dialog in the Open World: Platform and Applications

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+(01) 425 706 5880
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+(01) 425 706 2127
horvitz@microsoft.com

ABSTRACT

We review key challenges of developing spoken dialog systems that can engage in interactions with one or multiple participants in relatively unconstrained environments. We outline a set of core competencies for *open-world dialog*, and describe three prototype systems. The systems are built on a common underlying conversational framework which integrates an array of predictive models and component technologies, including speech recognition, head and pose tracking, probabilistic models for scene analysis, multiparty engagement and turn taking, and inferences about user goals and activities. We discuss the current models and showcase their function by means of a sample recorded interaction, and we review results from an observational study of open-world, multiparty dialog in the wild.

Categories and Subject Descriptors

H.1.2 **[Models and Principles]**: User/Machine System – *Human Information Processing*; H.5.2 **[Information Interfaces and Presentation]** User Interfaces – *Natural Language*; I.4.8 [**Scene Analysis**]: Tracking, Sensor Fusion

General Terms

Algorithms; Human Factors

Keywords

Spoken dialog; open-world models; multimodal; multiparty interaction; situated interaction; engagement; turn-taking; floor management.

1. INTRODUCTION

Most spoken dialog systems research to date can be characterized as the study and support of interactions between a single human and a computing system within a constrained, predefined communication context. Efforts in this realm have led to significant progress culminating in wide-scale deployments that now make telephony-based spoken dialog systems commonplace in the lives of millions of people. Nevertheless, numerous and important challenges remain with enabling computational systems to engage in fluid conversations in open, unconstrained environments, where multiple people with different and varying intentions enter and leave, and communicate and coordinate with each other and with interactive systems. We focus in this paper on these challenges.

We begin by reviewing several aspects of open-world interaction that represent key departures from assumptions typically made in traditional spoken dialog systems and we highlight a set of related research challenges and opportunities in Section 2. Then, in Sections 3 and 4, we present details of a framework for dialog systems that addresses several of these challenges. The approach integrates several core technologies, including speech recognition, machine vision, probabilistic models for scene analysis, multiparty engagement, turn-taking, and behavioral models for controlling an avatar, to support fluid dialog in open, dynamic environments.

We have explored three different applications on this platform, allowing us to investigate differences and similarities in open-world dialog across different domains. We discuss these different conversational agents in Section 5. We showcase by means of a recorded interaction how the different models work together to support mixed-initiative engagement and dialog with multiple parties. We also review results from an initial in situ observational study of multiparty interaction performed with one of these systems. Finally, in Section 6 we conclude and outline current and future planned research in this realm.

2. DIALOG IN THE OPEN WORLD

Interaction in open, unconstrained environments can be characterized as making two key departures from assumptions typically made in traditional spoken dialog systems. The first difference is the *dynamic, multiparty* nature of the interaction, *i.e.*, the world typically contains not just one, but multiple agents who may be relevant to the computational system. Furthermore, interactions in the open world are often dynamic and asynchronous, *i.e.* relevant agents may enter and leave the observable world at any time, may interact with the system and with others, and their goals, plans, and needs may change over time.

A second departure from traditional spoken dialog systems is that the interactions are *situated*, *i.e.* the surrounding physical environment provides rich, streaming context that is relevant for conducting and organizing the interactions. Situated interactions among people often hinge on shared information about physical details and relationships, including structures, geometric

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11...$10.00.

relationships and pathways, objects, topologies, and communication affordances. Like the multi-participant aspect, the often implicit, yet powerful *physicality* of situated interaction, provides opportunities for making ongoing inferences in open-world dialog systems, and challenges system designers to innovate across a spectrum of complexity and sophistication.

Specifically, we note that the dynamic, multiparty, and situated aspects of open-world interaction frame new challenges in areas like engagement, turn-taking, language understanding, and dialog management. As an example, simple approaches for regulating engagement, such as push-to-talk buttons, are sufficient in closed-world contexts where there is an assumed single user. However, these solutions are not appropriate for systems that must operate in open environments, such as robots, interactive billboards, and embodied conversational agents. New models that can leverage the physical details of the scene (*e.g.,* spatiotemporal trajectories, geometric relationships in formations of people, and objects being carried or pointed at) as well as additional communication affordances (*e.g.,* the dynamics of gaze among multiple people and system) are required for enabling computational systems to regulate engagement in an open-world, multi-participant setting.

Once engaged, a natural language interactive system must be able to coordinate with the other participants on the presentation and recognition of communicative signals, in a process known as turn-taking [11, 18]. Computational models for turn taking have been proposed and evaluated in prior work [16, 21, 22]. However, most models developed to date operate in a single-user setting. Open-world dialog requires the development of situated multiparty turn-taking models that would allow a system to continuously track who is speaking to whom and who has the conversational floor, in order to seamlessly coordinate its outputs with others.

At the higher levels, such systems must be able to correctly decode the meaning of the received communicative signals. Interesting challenges arise here in terms of integrating continuously streaming context into the language understanding and intentions recognition process. These challenges extend beyond the utterance, to the discourse and dialog level. With the exception of a few incipient efforts [13, 23], most current models for discourse understanding and dialog management [4, 5, 6, 14, 17] make an implicit single-user assumption and do not represent nor leverage the situated nature of the interactions.

Beyond adding new dimensions to existing dialog problems, the open-world setting also raises new fundamental research challenges. Interacting successfully in open environments requires that information from multiple sensors is fused to detect, identify, track and characterize the relevant agents and entities in the scene, as well as the evolving relationships between them. Models for inferring and tracking the activities, goals, and long-term plans of these agents can provide additional context for reasoning and providing assistance within and beyond the confines of a given interaction. Furthermore, goals and plans may lay outside the scope of the current models used by system to understand human intentions. A system may need to recognize the prospect that it does not understand something about a situation that people might easily interpret in human-human conversation. Such awareness and readiness for addressing the likelihood that a system's models are incomplete is important in grounding with people in a natural manner. More generally, the ability to make inferences about the inadequacy of current models and to activate measures to extend them, are important aspects of open-world intelligence.

Developing end-to-end, open-world interactive systems hinges therefore on the successful integration of a number of different technologies. Some of the sub-problems, such as tracking, activity recognition, and the identification of the sources and targets of speech have already received significant amounts of attention in different research communities, and specialized solutions have been developed. Open-world dialog tests the boundaries of these solutions and poses new challenges in combining existing and new technologies in support of seamless interaction. It also highlights new opportunities; for instance, within an interactive setting, there are opportunities for engaging people to assist with learning so as to increase the robustness of components and models over time.

Our long-term research goal is to construct computational models that provide the core skills needed for handling open-world dialog with the etiquette, fluidity, and social awareness expected in human-human interactions. In order to provide an ecologically valid basis for investigating these challenges, we have brought together a number of technologies into a reusable framework for open-world interaction, and we have used this framework to construct systems that provide a real-world experimental test bed for research. In the sequel, we describe this platform and review the component technologies and an initial set of models that provide core competencies for open-world dialog.

3. SYSTEM ARCHITECTURE

Figure 1 provides a high-level overview of the current underlying hardware and software architecture. Although the three systems we have developed to date take the form of static multimodal kiosks, the methods extend to other form factors, such as for instance mobile robots. The sensory apparatus currently used in these systems includes the following components:

- a wide-angle AXIS 212 camera with a 140° field of view and a resolution of 640x480 pixels; the camera also supports pan-tilt-zoom in software, and we are currently exploring a foveal vision solution using a combination of two of these cameras;
- a 4-element linear microphone array that captures the audio signal, performs acoustic echo cancellation, and provides sound-source localization information in 10° increments;
- a 19" touch-screen that displays a talking avatar, at times complemented by a graphical user interface; the touch screen can be used as an additional input channel;
- an RFID badge reader that can provide identification information for employees at our organization.

Data gathered by the sensors is preprocessed and forwarded to a scene-analysis module that fuses the incoming streams and constructs in real-time a coherent picture of the dynamics in the surrounding environment (illustrated in Figure 1). The analysis includes detecting and tracking the location of multiple agents in the scene, reasoning about their attention, activities, goals and relationships (*e.g.,* which people are in a group), and tracking the conversational context at different levels (*e.g.,* who is currently engaged, or waiting to engage in a conversation, who has the conversational floor, who is currently speaking to whom, etc.). The models are discussed in more detail in the next section.

The conversational scene analysis results are forwarded to the control level, which is structured in a two-layer reactive-deliberative architecture. The reactive layer implements and coordinates low-level reactive behaviors (*e.g.* for engagement and turn taking, for coordinating spoken and gestural outputs, etc.)

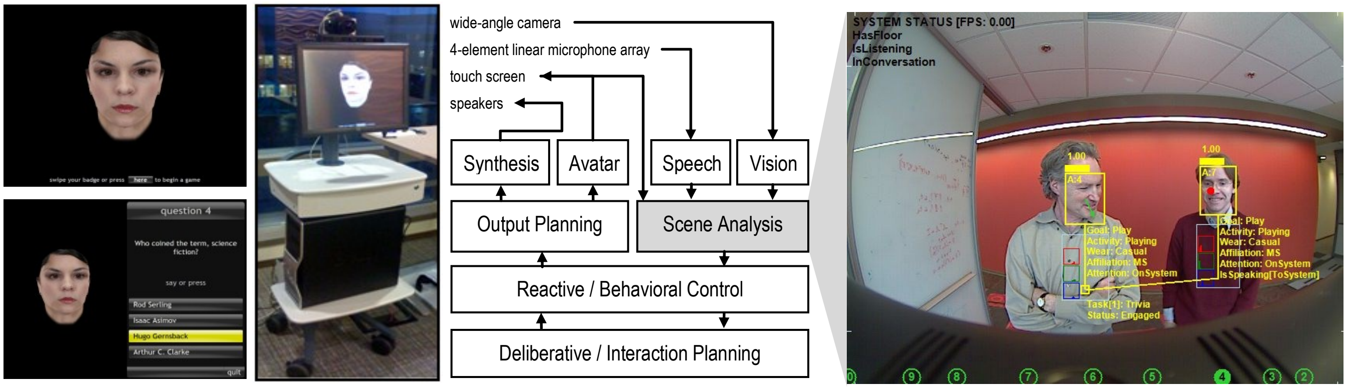


Figure 1. Hardware and software components within the overall architecture, and an illustration of scene analysis results

The deliberative layer makes conversation control decisions, and plans the system's responses.

4. CORE COMPETENCIES

4.1 Situational Awareness

Conducting interaction in the open world requires a minimal set of situational awareness capabilities. Higher-level interaction processes and inferences are, to a large extent, predicated on the ability to detect, track, identify, and characterize relevant agents and entities in the scene. Below, we describe the set of physical awareness components currently implemented in our framework.

Face detection and tracking. A detector and tracker for multiple faces [25] are used to track the location $x_a(t)$ of each agent a. The detector runs at every frame and is used to initialize a mean-shift tracker. The frame-to-frame face correspondence problem is resolved by a proximity based algorithm. The algorithms run on a scaled-up image (1280x960 pixels), allowing us to detect frontal faces up to a distance of about 20 feet. Apart from the face locations $x_a(t)$ and sizes $w_a(t)$, the tracker also outputs a confidence score $fc_a(t)$, which is used to prune false detections and to infer focus of attention (described later.)

Pose tracking. While an agent is engaged in a conversation with the system, a face-pose tracking algorithm [24] runs on a cropped region of interest encompassing the agent's face. During group interactions, multiple instances of this algorithm run in parallel on different regions of interest. The pose tracker provides 3D head orientation information for each engaged agent $\overline{\omega_a}(t)$, which is in turn used to infer the focus of attention (see below.)

Focus of attention. At every frame, a probabilistic model is used to infer whether the attention of each agent in the scene is oriented towards the system or not: $\mathrm{p}(foa_a(t)|fc_a(t),\overline{\omega_a}(t))$. This inference is currently based on a maximum entropy model trained using a hand-labeled dataset. The features used are the confidence score from the face tracker $fc_a(t)$ (this is close to 1 when the face is frontal), and the 3D head orientation generated by the pose tracker $\overline{\omega_a}(t)$, when available (recall that the pose tracker runs only for engaged agents.) We are currently exploring models that leverage additional high-level interaction features to jointly track the attention of multiple agents during multiparty interactions.

Agent characterization. Apart from detecting and tracking relevant agents in the scene, we have also implemented a simple model that performs a basic visual analysis of the clothing of each detected agent. The color variance in a rectangular patch below the face is currently used to infer whether the agent is dressed casually or formally (*e.g.,* if a person is wearing a suit, this often leads to high variance in this image patch), and to re-identify people that leave the visible scene for a short period of time. The clothing information is further used to infer the agent's likely affiliation; at our organization, casually dressed agents are more likely to be employees and formally dressed ones are likely to be visitors. We are currently exploring more robust models for agent characterization, with a focus on person and gender identification.

Group inferences. Beyond characterizing single agents, we have also implemented models for inferring group relationships among agents in the scene. The probability of two agents being in a group together $\mathrm{p}(\mathrm{group}(a_1,a_2))$ is computed by a maximum entropy model that was trained on a small hand-labeled dataset. The model currently uses as features the size, location and proximity of the faces, but can also leverage observations collected through interaction. For instance, the system might ask a clarification question like, "Are the two of you together?" Positive or negative responses to this question are also used as evidence.

4.2 Situated, Multiparty Engagement

As a prerequisite for open-world interaction, a dialog system must be able to coordinate its actions with other participants in the scene to initiate, maintain, and terminate *engagement* [15, 19]. Observational studies have revealed that humans negotiate engagement via a mixed-initiative, coordinated process in which non-verbal cues and signals, such as spatial trajectory and proximity, gaze and mutual attention, head and hand gestures, and verbal greetings all play essential roles [1, 7, 12]. Successfully modeling this process requires that the system (1) senses and reasons about the engagement actions, state and intentions of multiple agents in the scene, (2) makes high-level engagement control decisions (*i.e.* whom to engage with and when), and (3) renders these decisions via low-level coordinated behaviors and outputs. The engagement model that we implemented subsumes these three components. A full description of this model is available in [2, 3]. Here, we provide a brief overview.

The model is centered on a reified notion of *interaction*, defined as a basic unit of sustained, interactive problem-solving. Each interaction involves two or more participants, and this number may vary in time; new participants may join and current participants may leave an existing interaction at any point. The system is actively engaged in at most one interaction at a time, but it can simultaneously track additional, suspended interactions.

The sensing subcomponent in the model tracks the *engagement state*, *engagement actions*, and *engagement intentions* for each agent in the visual scene. The engagement state, $ES_a^i(t)$, denotes whether an agent a is engaged in interaction i and is modeled as a deterministic variable with two possible values: *engaged* and *not-engaged*. Since engagement is a collaborative, coordinated process, the state is updated based on the joint actions of the system and the agent; for instance, if both the system and the agent take an engaging action, the state is updated to *engaged*.

A second engagement variable, $EA_a^i(t)$, models the actions that an agent takes to initiate, maintain or terminate engagement. Four possible actions are defined in our model: *engage*, *no-action*, *maintain*, and *disengage*. These actions are tracked by means of a conditional probabilistic model that takes into account the current engagement state, the previous agent and system actions, as well as additional sensory evidence capturing committal engagement signals. These include salutations (*e.g.* "Hi!"); calling behaviors (*e.g.* "Laura!"); the establishment or the breaking of an F-formation [12] (*e.g.,* a user standing in front of the system); expected opening dialog moves (*e.g.* "Come here!"), etc.

A third variable in the proposed model, $EI_a^i(t)$, tracks whether or not each agent intends to be engaged in a conversation with the system. Like the engagement state, the intention can either be *engaged* or *not-engaged*. Intentions are tracked separately from actions since an agent might intend to engage or disengage the system, but not yet take an explicit engagement action. For instance, consider the case in which the system is already engaged in an interaction and another agent is waiting in line to interact with the system: although the waiting agent does not take an explicit, committed engagement action, she might signal (*e.g.*, via a glance that makes brief but clear eye contact between the agent and the system) that her intention is to engage in a new conversation as soon as the opportunity arises. More generally, the engagement intention captures whether or not an agent would respond positively should the system initiate engagement. In that sense, it roughly corresponds to Peters' [15] "interest level," *i.e.*, to the value the agent attaches to being engaged in a conversation. Like engagement actions, engagement intentions are inferred based on probabilistic models that take into account the current engagement state, the previous agent and system actions, the previous engagement intention, as well as additional evidence that captures implicit engagement cues, *e.g.* the spatiotemporal trajectory of the participant, the level of sustained mutual attention, etc. We describe in [2] an approach for automatically learning such models directly from interaction data.

Based on the inferred engagement state, actions, and intentions of the agents in the scene, as well as additional high-level evidence such as the agents' inferred goals, activities and relationships, the proposed model uses an engagement control policy to make engagement decisions. The system's action-space consists of the same four actions previously discussed: *engage*, *disengage*, *maintain,* and *no-action*. Finally, at the lower level, these engagement actions are rendered into a set of coordinated low-level behaviors such as head gestures, establishing and breaking eye contact, issuing greetings and salutations, interjections, etc.

The lower-level engagement sensing and behavioral components are domain-independent, and are reused across different systems. However, the engagement control policy can be tuned to the characteristics of a particular application. In Section 5, we illustrate how these models work together and also leverage domain-specific information to regulate the engagement process in an open-world, multiparty setting, and we review results from a preliminary in-the-wild study of multiparty engagement.

4.3 Situated, Multiparty Turn-Taking

Once engaged in a conversation, a system must coordinate with the other participants on the presentation and recognition of various communicative signals, in a process known as *turn-taking* [11, 18]. Our current framework implements a multi-participant turn taking model, which we briefly review below. The model allows the system to track the speech source and target for each utterance, to identify who currently holds the conversational floor, and to coordinate its outputs with the other participants.

A voice activity detector is used to identify and segment out spoken utterances from background noise. The speaker S_u for each utterance u is identified by a model that fuses throughout the duration of the utterance the sound source localization information provided by the microphone array with information from the vision subsystem, specifically the location of the agents in the scene. For each identified utterance, the system infers whether the utterance was addressed to it or not. This is accomplished by means of a model that integrates over the user's inferred focus of attention throughout the duration of the spoken utterance $\mathrm{p}(T_u = \mathrm{system}|foa_{S_u}(t))$. If the user's focus of attention stays on the system, the utterance is assumed to be addressed to it; otherwise, the utterance is assumed to be directed towards the other participants engaged in the conversation.

In addition, the multi-participant turn-taking model tracks whether or not each agent currently holds the conversational floor $FS_a(t)$ (*i.e.,* has the right to speak), and what the floor management actions each agent takes at any point in time $FA_a(t)$: *no-action*, *take-floor*, *hold-floor*, *release-to-system*, or *release-to-other*. These actions are currently inferred based on a set of rules that leverage information about the current state of the floor $\{FS_a(\mathrm{t})\}_\mathrm{a}$, the current utterance u, its speaker S_u and its addressees T_u. For instance, a *take-floor* action is detected if a participant does not currently hold the floor but starts speaking or interacts with the GUI; a *release-to-system* action is detected when a participant finishes speaking, and the utterance was addressed to the system; and so on. The floor state for each agent $FS_a(t)$ is updated based on the joint floor-management actions of the system and engaged agents. For instance if a user currently holds the floor and performs a *release-to-system* action, immediately afterwards the floor is assigned to the system.

Based on who is currently speaking to whom and on who holds the floor, the system triggers its own floor actions to coordinate its outputs with the other conversational participants. For instance, the system behavior that generates spoken utterances verifies first that the system currently holds the floor. If this is not true, a *take-floor* action is invoked. The behavioral layer renders this action by coordinating the avatar's gaze, gesture and spoken signals (*e.g.,* "Excuse me!," if the system is trying to take the floor but a participant is holding it while speaking to another participant).

The current multi-participant turn-taking model is an initial iteration on a path to more sophisticated models. It employs a combination of handcrafted heuristic rules and limited evidential reasoning, treats each participant independently, and does not explicitly take into account the rich temporality of interactions. We are exploring the construction and use of more sophisticated data-driven models for jointly tracking over time the speech source S_u, target T_u, focus of attention $foa_a(t)$ and floor state $FS_a(t)$ and actions $FA_a(t)$ in multi-participant conversation, by fusing through time audio-visual information with additional information about the system actions (*e.g.,* its pose and gaze trajectory, etc.), and the history of the conversation: $\mathrm{p}(S_u, T_u, foa_{\{a\}}(t), FS_{\{a\}}(t), FA_{\{a\}}(t) | \Psi(t))$

4.4 Situated Intention Recognition

Once communicative signals have been segmented and identified, the system must correctly interpret their meaning and recognize the underlying user intentions. The methodology provides support for defining domain- and application-specific goal and activity models, and for building hybrid belief updating models that leverage both streaming context and information collected via dialog to infer goals, activities, and intentions.

For instance, in the Receptionist system described in the next section, the goal inference models take into account the estimated actor affiliation and whether or not the actor is part of a larger group (*e.g.*, people at our organization are more likely to want shuttles than to register as visitors, people in a group are more likely to register as visitors, etc.). If the probability of the most likely goal does not exceed a grounding threshold, the system collects additional evidence through interaction, by directly asking or confirming the speculated goal. Similarly, in case an agent's goal is to make a shuttle reservation, the number of people for the reservation is inferred by a model that integrates information from the scene (*e.g.*, how many people are present) with data gathered through dialog. The activity model allows the system to track the long term plans of individual agents and provides support for assistance beyond the temporal confines of a single conversation.

5. SYSTEMS

Having discussed the core set of component technologies, we now describe three different systems that we have developed using this platform, and that currently serve as an experimental test-bed for investigating the challenges of open-world dialog. The first system, described in subsection 5.1, is a situated conversational agent that performs tasks typically handled by front-desk receptionists at our organization. We use a sample interaction with this system to showcase how the models described earlier work together to support fluid, multiparty dialog in open environments. Then, in subsection 5.2, we describe a second system that plays an educational questions game. We review results and lessons learned from an initial in-the-wild deployment of this system, focusing on the multiparty engagement aspects. Finally, in subsection 5.3, we discuss a third prototype designed to serve as personal assistant.

5.1 Receptionist

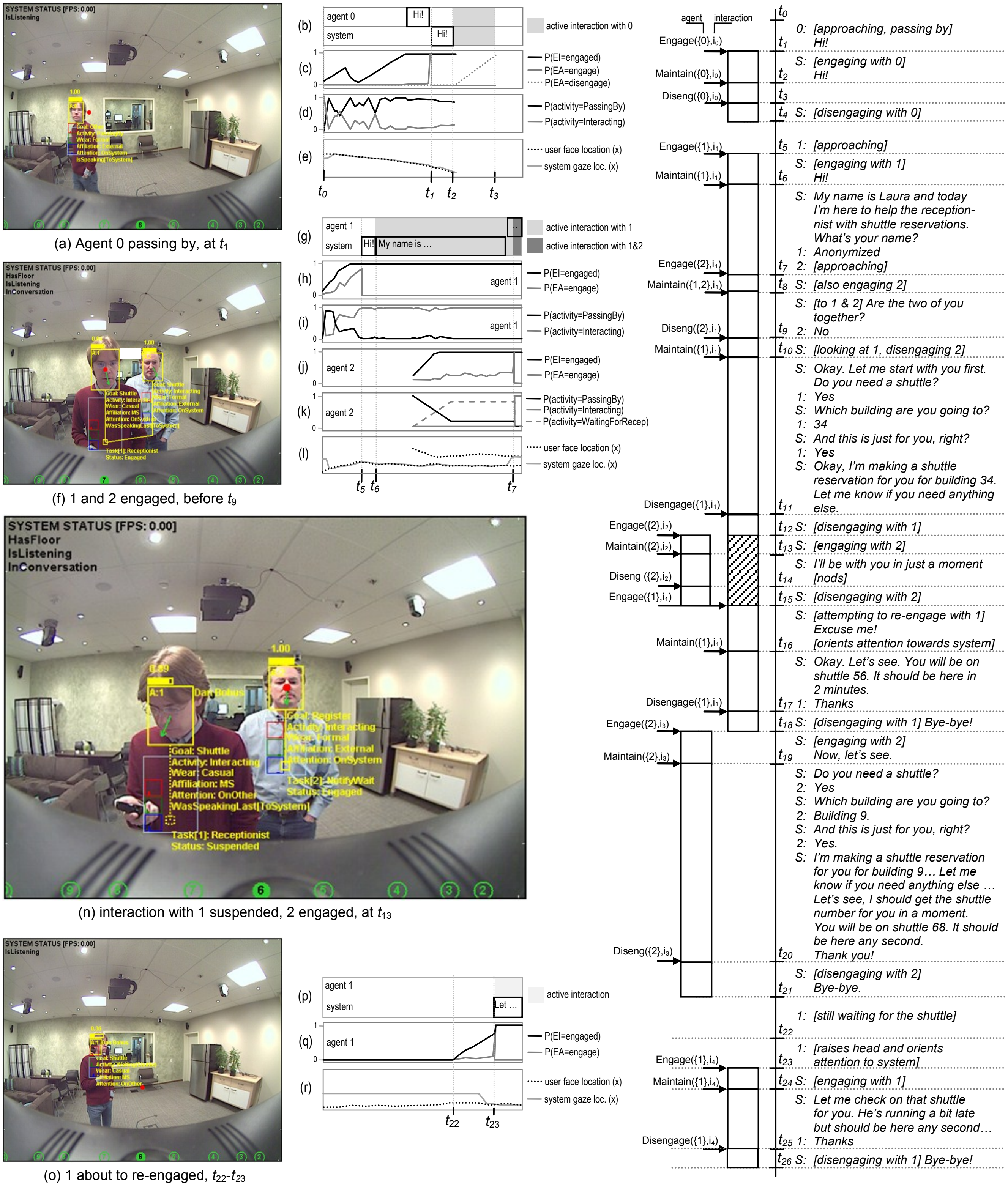
Receptionist is a situated conversational agent that performs some of the tasks typically handled by front-desk receptionists, such as making shuttle reservations, registering visitors, providing directions on campus, etc. Figure 2 illustrates a conversation with this system, showing several successive snapshots from the interaction together with the runtime annotations created by the various models, as well as plots of the temporal evolution of key variables in the underlying models. Full video captures of this, as well as additional interactions with Receptionist and the other systems described in the sequel are available online [20].

At time t_0, a first participant (A_0) enters the visual field of the system, which immediately detects and tracks the participant as he approaches (see illustrated system gaze in Figure 2.e). When idle, Receptionist uses a conservative engagement policy: it waits until a participant performs an explicit engagement action. In this case, as A_0 passes by, he says "Hi!" This is recognized as an engagement action (see also Figure 2.c) and the system responds by triggering its own *engage* action. At the behavioral level, mutual attention has already been established, and the system starts a greeting behavior "Hi!" at time t_1 (see Figure 2.a,b). A_0's state transitions to *engaged*, and the interaction becomes active at time t_2. The higher level activity and goal models, which also leverage the streaming context of A_0's trajectory and attention, indicate that his most likely current activity is *Passing-By* (Figure 2.d) and the most likely goal is *Other* (Receptionist uses a goal-activity model with 3 goals - *Shuttle, Register, Other*, and 4 activities - *Interacting, Waiting-For-Receptionist, Waiting-For-Shuttle,* and *Passing-By*). Leveraging this inference, the dialog manager decides not to start a dialog just yet. As A_0 leaves the field of view, the system recognizes a disengagement action at t_3 (Figure 2.d). The system disengages and goes back to idle at t_4.

Shortly thereafter, right before time t_5 the participant approaches again (A_1). Based on attention and trajectory, the system again recognizes the intention to engage (Figure 2.h) and waits for an explicit engagement action. In this case A_1 enters in an *F-formation* with the system by standing right in front of it at time t_5 (see Figure 2.h). The system responds by triggering an *engage* action. This time, the goal model indicates that A_1's most likely current activity is *Interacting* (Figure 2.i). The dialog manager asks for the participant's name. The difference between the two mini-interactions described so far shows how, by separately tracking engagement state, actions, and intentions, as well as higher level goals and activities, the system can implement a policy that allows is to engage and interact to different degrees, according to the inferred participants goals and needs.

As the system begins its interaction with A_1, another participant (A_2) also approaches. Based on the level of sustained mutual attention, the system infers that A_2 also wants to engage even though no explicit engagement action can yet be observed with high probability (see $\mathrm{p}(EI = \text{engaged})$, $\mathrm{p}(EA = \text{engage})$ in Figure 2.i, prior to time t_7). By leveraging proximity and the times of arrival, the group inference model indicates there is a significant likelihood that A_1 and A_2 are together. In light of this inference and of the recognized engagement intention, once the system finishes its current prompt, the engagement control policy decides to engage this A_2 and allow him to join the current interaction. The engagement action is rendered as a simple glance to A_2, shown in Figures 2.f and 2.l, around time t_7. After time t_7, A_1 and A_2 are both engaged, as illustrated in Figure 2.f and 2.g.

Next, the dialog manager tries to identify the roles and goals of the current participants. Since the group relationship is not yet grounded, a clarification action is taken – the system asks "Are the two of you together?", while glancing at both A_1 and A_2. A_2 responds "No." The system adjusts its interaction plans and decides to disengage A_2 (at time t_9) and continue the interaction only with A_1 from time t_{10}. The interaction continues and shuttle arrangements are made for A_1. While waiting for the shuttle reservation backend component to respond, the system decides to engage again with A_2 temporarily to let him know that he will be attended to shortly. This is again accomplished by means of several successive disengagement and engagement actions, occurring over the time period t_{11} to t_{15}.



(a) Agent 0 passing by, at t_1

(f) 1 and 2 engaged, before t_9

(n) interaction with 1 suspended, 2 engaged, at t_{13}

(o) 1 about to re-engaged, t_{22}-t_{23}

t_0
0: *[approaching, passing by]*
t_1 *Hi!*
S: *[engaging with 0]*
t_2 *Hi!*
t_3
t_4 S: *[disengaging with 0]*
t_5 1: *[approaching]*
S: *[engaging with 1]*
t_6 *Hi!*
S: *My name is Laura and today I'm here to help the receptionist with shuttle reservations. What's your name?*
1: *Anonymized*
t_7 2: *[approaching]*
t_8 S: *[also engaging 2]*
S: *[to 1 & 2] Are the two of you together?*
t_9 2: *No*
t_{10} S: *[looking at 1, disengaging 2]*
S: *Okay. Let me start with you first. Do you need a shuttle?*
1: *Yes*
S: *Which building are you going to?*
1: *34*
S: *And this is just for you, right?*
1: *Yes*
S: *Okay, I'm making a shuttle reservation for you for building 34. Let me know if you need anything*
t_{11} *else.*
t_{12} S: *[disengaging with 1]*
t_{13} S: *[engaging with 2]*
S: *I'll be with you in just a moment*
t_{14} *[nods]*
t_{15} S: *[disengaging with 2]*
S: *[attempting to re-engage with 1] Excuse me!*
t_{16} *[orients attention towards system]*
S: *Okay. Let's see. You will be on shuttle 56. It should be here in 2 minutes.*
t_{17} 1: *Thanks*
t_{18} S: *[disengaging with 1] Bye-bye!*
S: *[engaging with 2]*
t_{19} *Now, let's see.*
S: *Do you need a shuttle?*
2: *Yes*
S: *Which building are you going to?*
2: *Building 9.*
S: *And this is just for you, right?*
2: *Yes.*
S: *I'm making a shuttle reservation for you for building 9… Let me know if you need anything else … Let's see, I should get the shuttle number for you in a moment. You will be on shuttle 68. It should be here any second.*
t_{20} *Thank you!*
S: *[disengaging with 2]*
t_{21} *Bye-bye.*
1: *[still waiting for the shuttle]*
t_{22}
1: *[raises head and orients*
t_{23} *attention to system]*
t_{24} S: *[engaging with 1]*
S: *Let me check on that shuttle for you. He's running a bit late but should be here any second…*
t_{25} 1: *Thanks*
t_{26} S: *[disengaging with 1] Bye-bye!*

Figure 2. Sample interaction with Receptionist. Left side shows stills from a recorded interaction, annotated with the scene analysis results (red dot indicates the target of the system's gaze), as well as the temporal evolution of key variables for engagement and activity models. Right side shows a transcript of this multiparty interaction, and illustrates the various engagement actions taken by the system (parameterized by the agent and the interaction).

After the system disengages A_2 at time t_{15}, it attempts to re-engage A_1 to finalize his reservation. The system's gaze moves back towards A_1. However, in the meantime the A_1's attention has drifted to his mobile device (Figure 2.n). In an effort to re-establish engagement, as part of its *engage* action, the system triggers an "Excuse me, sir!" behavior at time t_{15}. As P_2 lifts his gaze towards the gaze of the embodied agent, mutual attention is re-established and an engage action is recognized. The participant transitions to an engaged state, the system's *engage* action completes successfully, and the interaction continues.

Once the interaction is completed, the system re-engages A_2 and assists him with a shuttle reservation (from time t_{18} to t_{21}.) In the meantime, A_1 has been pacing around in the background, waiting for the shuttle, while checking his mobile device. After A_2 leaves, at time t_{22}, the A_1 lifts his gaze from the mobile device towards the system (see Figure 2.o), and the system recognizes an intention to engage (see Figure 2.q), but not a direct engagement action yet. The engagement control policy however leverages high-level information about the long-term goals and activities of this participant *i.e.* the system knows that the participant's current activity is *Waiting-For-Shuttle*, and decides to initiate engagement in order to reassure the participant about the shuttle's arrival.

Although the current models are still embryonic, the sample interaction we discussed illustrates how multiple components can be weaved together to provide support for fluid, multi-participant engagement and dialog in an open-world context. We are working on improving the robustness of these models and on developing more comprehensive task, goal and activity models prior to a real-world deployment of this system. In the next subsection, we discuss a second system in a different domain, which has already been deployed.

5.2 Questions Game

This second prototype, named the *Questions Game*, can engage one or multiple participants to play a game, in which users are challenged to answer multiple choice questions on a range of topics. In each game, the avatar goes through four multiple choice questions, one at a time. The possible answers are displayed on the screen after each question (as shown in Figure 1) and users can respond by either speaking an answer or by touching it. If the answer is incorrect, the system provides a short explanation regarding the correct answer before moving on to the next question. Like Receptionist, the Questions Game supports multi-participant interactions. When engaged in a multiplayer game, the system leverages the multi-participant turn-taking model to continuously track who the current speaker is, and who has the conversational floor. At the behavioral level, the avatar orients its head pose and gaze towards the current speaker, or towards the addressee(s) of its own utterances. After a response is received from one of the users, the avatar confirms the answer with the other user(s) who have been engaged and have agreed to play over the course of the session, *e.g.* "Do you agree with that?" A sample interaction with the system (video) is available online [20].

The Questions Game uses an engagement policy designed to foster multi-participant game situations, by attracting bystanders to join games that are already in progress. In making its engagement decisions, the system leverages higher level inferences about the activity of various agents in the surrounding environment. Specifically, the activity inference model uses information about attention and trajectory to classify the agents in the scene into: *Passing-By*, *Interacting*, *Playing*, *Watching*, and *Departing*. If a *not-engaged*, *Watching* agent (*i.e.,* a bystander) is detected while someone else is engaged in the game, the system attempts to attract the bystander in the interaction by suspending the existing game temporarily and engaging the bystander: "Hi. Would you like to join in?"

We conducted an initial, in situ observational study with this system to investigate whether it can effectively create and support such multi-participant interactions. The system was deployed for a period of 20 days near a kitchenette in our building. Throughout this time, 49 single user and 18 multiple user interactions were recorded. Most people who interacted with the system did so for the first time. No instructions were provided.

The full details of this observational study are described in more detail in [3]. In summary, the results show that the system can effectively initiate and maintain multi-participant interactions. Bystanders recognized when they were engaged and solicited by the system; they responded in the large majority of cases (87%), either by joining the game (35%) or refusing to do so (52%). In addition, the study also showed that a number of bystanders answered questions (either directly towards the system or towards the engaged participant) prior to the moment the system engaged them, highlighting novel challenges in detecting engagement intentions, as well as opportunities for designing more flexible, mixed initiative engagement policies. Additional lessons learned include the importance of robust face tracking in the presence of occlusions (41% of multiparty interactions where affected to various degrees by vision failures), and of the models for detecting the speech source and target in complex scenes.

5.3 Personal Assistant for Scheduling

A third prototype we have constructed is the *Personal Assistant for Scheduling* system (PASS). PASS "lives" outside the door of an employee's office at our organization, and is designed to act as a personal administrator with the ability to handle the challenges of coordination among people at an organization (*e.g.* scheduling and rescheduling meetings, passing messages, etc.)

At the core of the system's domain expertise are subsystems that provide expert knowledge about the availability and presence of the person it is serving. The system has access to the owner's online calendar which contains information about current and forthcoming meetings, as well as a long history of past meetings. Deeper knowledge about the forecasting of future presence and availability comes from a system that has been deployed and has been in use by some employees at our organization [9]. The system continuously acquires data about a user's locations and activities over time, across multiple computers and devices. It can consider desktop and mobile activities as well as calendar information. The system constructs query-specific case libraries and performs real-time Bayesian learning and reasoning to generate forecasts about location and communication presence in response to custom-tailored or standing queries. For example, the system can report the amount of time until a user will be present in their office for at least 15 minutes given the time of day, day of week, the meetings listed on their calendar, and the observation that they have been away from their office for 45 minutes. A second subsystem [10] learns via supervised learning to predict the cost of interrupting a user based on desktop activity, location, time of day and day of week, and properties of a meeting that is currently in progress. The two systems endow PASS with the ability to engage people who approach the system with rich and informative dialog, and provide assistance with scheduling or rescheduling meetings, and relaying messages between the owner and people that stop by.

6. CONCLUSION

We have described a research platform for dialog systems that can interact naturally and provide assistance in open, relatively unconstrained environments, where multiple people with different needs, goals, and long-term plans may enter, interact, and leave the observable world, and where the physical surroundings provide rich streaming evidence relevant for organizing and conducting the interactions. We discussed several challenges that are highlighted by moving from the implicit assumptions made in traditional spoken dialog systems to open-world interaction. We described how we can leverage a set of sensory, learning, and reasoning components within an overall architecture that can address several of these challenges. Finally, we presented three applications developed on this platform that showcase the potential for creating systems that can interact in the open-world, with the ease, social skills and etiquette expected from a human.

Together with these systems, the framework described in this paper provides a real-world test-bed for investigating the challenges of open-world interaction. For instance, in an initial set of experiments briefly reviewed here but reported in detail in [2, 3], we have developed models for automatically learning to recognize engagement intentions, and for managing the engagement process in an open-world setting. Moving forward, we will investigate the many remaining challenges of open-world interaction, from situated multiparty turn-taking models, to multi-party dialog management, to open-domain learning and knowledge acquisition from interaction. We believe solutions to these problems can pave the way towards interactive systems that can seamlessly embed computation and interaction deeply into the natural flow of our daily activities and collaborations.

7. ACKNOWLEDGMENTS

We would like to thank G. Chrysanthakopoulos, Z. Liu, T. Paek, Q. Wang, C. Zhang for useful discussions and feedback.

8. REFERENCES

[1] M. Argyle. Bodily Communication, International University Press, Inc, New York (1975).

[2] D. Bohus and E. Horvitz, Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings, in Proceedings of SIGdial'09, London, UK (2009)

[3] D. Bohus and E. Horvitz, Models for Multiparty Engagement in Open-World Dialog, in Proceedings of SIGdial'09, London, UK (2009)

[4] D. Bohus and A. Rudnicky. The RavenClaw Dialog Management Framework: Architecture and Systems, Computer Speech and Language, DOI:10.1016/j.csl.2008.10.001

[5] R. Cole. Tools for Research and Education in Speech Science, in Proceedings of International Conference of Phonetic Sciences, San Francisco, CA (1999)

[6] G. Ferguson, and J. Allen. TRIPS: An Intelligent Integrated Problem-Solving Assistant, in Proceedings of AAAI'98, Madison, WI (1998)

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

[8] E. Horvitz. Reflections on Challenges and Promises of Mixed-Initiative Interaction, in AI Magazine vol. 28, Number 2 (2007)

[9] E. Horvitz, P. Koch, C.M. Kadie, and A. Jacobs. Coordinate: Probabilistic Forecasting of Presence and Availability, in Proceedings of UAI '02, Edmonton, Canada (2002).

[10] E. Horvitz, J. Apacible, and P. Koch. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption, in Proceedings of CSCW, ACM Press, (2004).

[11] J. Jaffe and S. Feldstein. Rhythms of Dialogue, Academic Press (1970)

[12] A. Kendon. Conducting Interaction: Patterns of Behavior in Focused Encounters, Studies in International Sociolinguistics, Cambridge University Press (1990)

[13] F. Kronlid. Steps towards Multi-Party Dialogue Management, Ph.D. Thesis, University of Gothenburg (2008)

[14] S. Larsson. Issue-based dialog management, Goteborg University, Ph.D. Thesis (2002)

[15] C. Peters, C. Pelachaud, E. Bevacqua, and M. Mancini, "A model of attention and interest using gaze behavior", *Lecture Notes in Computer Science,* pp. 229-240, 2005.

[16] A. Raux and M. Eskenazi. Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System, in Procs SIGdial'08, Columbus, OH (2008)

[17] C. Rich, C. Sidner, and N. Lesh. COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction, in AI Magazine. 22:15-25 (2001)

[18] H. Sacks, A. Schegloff, G. Jefferson. A simplest systematic for the organization of turn-taking for conversation. *Language*, 50(4):696-735 (1974).

[19] C. Sidner and C. Lee. Engagement rules for human-robot collaborative interactions, in IEEE International Conference on Systems, Man and Cybernetics, Vol 4, 3957-3962, (2003)

[20] Situated Interaction Project page: http://research.microsoft.com/en-us/um/people/dbohus/research_situated_interaction.html

[21] K. R. Thórisson. A Mind Model for Multimodal Communicative Creatures and Humanoids, in International Journal of Applied Artificial Intelligence, 13(4-5): 449-486 (1999)

[22] K. R. Thórisson. Natural Turn-Taking Needs No Manual: Computational Theory and Model, From Perception to Action, in Multimodality in Language and Speech Systems, 173-207, Kluwer Academic Publishers (2003)

[23] D. Traum and J. Rickel. Embodied Agents for Multi-party Dialogue, in Immersive Virtual Worlds, AAMAS'02, pp 766-773 (2002)

[24] Q. Wang, W. Zhang, X. Tang and H. Shum. Real-Time Bayesian 3-D Pose Tracking, in IEEE Trans. CSVT, vol. 16, no.12, pp. 1533-1541 (2006)

[25] C. Zhang, and Y. Rui. Robust Visual Tracking via Pixel Classification and Integration, in ICPR'2006, Hong Kong, China (2006)