Open-World Dialog: Challenges, Directions, and Prototype

Dan Bohus and Eric Horvitz

Microsoft Research
One Microsoft Way
Redmond, WA, 98052

Abstract

We present an investigation of *open-world dialog*, centering on building and studying systems that can engage in conversation in an open-world context, where multiple people with different needs, goals, and long-term plans may enter, interact, and leave an environment. We outline and discuss a set of challenges and core competencies required for supporting the kind of fluid multiparty interaction that people expect when conversing and collaborating with other people. Then, we focus as a concrete example on the challenges faced by receptionists who field requests at the entries to corporate buildings. We review the subtleties and difficulties of creating an automated receptionist that can work with people on solving their needs with the ease and etiquette expected from a human receptionist, and we discuss details of the construction and operation of a working prototype.

1. Introduction

Most spoken dialog 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 space have led to the development and wide-scale deployment of telephony based, and more recently multimodal mobile applications. At the same time, numerous and important challenges in the realm of situated and open-world communication remain to be addressed.

In this paper, we review challenges of dialog in *open-world* contexts, where multiple people with different and varying intentions enter and leave, and communicate and coordinate with each other and with interactive systems. We highlight the opportunity to develop principles and methods for addressing these challenges and for enabling systems capable of supporting natural and fluid interaction with multiple parties in open worlds—behaviors and competencies that people simply assume as given in human-human interaction. We begin by reviewing the core challenges of moving from *closed-world* to *open-world* dialog systems, and outline a set of competencies required for engaging in natural language interaction in open, dynamic, relatively unconstrained environments. We ground this discussion with the review of a real-world trace of human-human interaction. Then, we present details of a prototype open-world conversational system that harnesses multiple component technologies, including speech recognition, machine vision, conversational scene analysis, and probabilistic models of human behaviour. The system can engage in interaction with one or more participants in a natural manner to perform tasks that are typically handled by receptionists at the front desk of buildings. We describe the set of models and inferences used in the current system and we highlight, via review of a sample interaction, how these components are brought together to create fluid, mixed-initiative, multiparty dialogs.

2. Open-World Dialog

To illustrate several challenges faced by open-world dialog systems, we shall first explore real-world human-human interactions between a front-desk receptionist and several people who have arrived in need of assistance. We focus on a representative interaction that was collected as part of an observational study at one of the reception desks at our organization. The interacting parties and physical configuration are displayed in the video frame in Figure 1.

At the beginning of the segment, the receptionist is on the phone, handling a request about scheduling a conference room, viewing availabilities of rooms and times on her computer in support of the request. Participant 1 (P_1) is an external visitor who the receptionist has just finished speaking with; he is currently filling in a visitor registration form. As P_1 is completing the form, the receptionist answers the telephone and engages in a phone conversation with participant 4 (P_4). During this time, participant 2 (P_2) enters the lobby from inside the building, approaches the reception desk, and makes eye contact with the receptionist. The receptionist, knowing that P_1 needs additional time to complete the registration form, and that the conversation can continue with P_4 while she engages in a fast-paced interaction with P_2, moves to engage with P_2. Apparently relying on inferences from the observation that P_2 came from inside the building, the receptionist guesses that P_2 most likely needs a shuttle to another building on the corporate campus. She lifts her gaze towards P_2 and asks P_2 softly (while moving her mouth away from the phone microphone), "Shuttle?" P_2 responds with a building number.

Figure 1. Video frame from a multiparty interaction.

While the receptionist continues on the phone with P_4 on options for arranging a meeting room in the building, she interacts with a shuttle ordering application on the computer. Soon, participant 3 (P_3) approaches the reception desk. At this time, P_2 re-establishes eye contact with the receptionist and indicates with a quick hand gesture and a whisper that the shuttle is for two people. The receptionist now infers that P_2 and P_3—who have not yet displayed obvious signs of their intention to travel together—are actually together. The receptionist whispers the shuttle identification number to P_2 and continues her conversation with P_4, without ever directly addressing P_3. Later, once P_1 completes the form, the receptionist re-engages him in conversation to finalize his badge and contact his host within the building.

The interaction described above highlights two aspects of open-world dialog that capture key departures from the assumptions typically made in traditional dialog systems. The first one is the *dynamic, multiparty* nature of the interaction, *i.e.*, the world typically contains not just one, but multiple agents who are relevant to a computational system, each with their own goals and needs. The second departure from traditional dialog systems is that the interaction is *situated*, *i.e.*, that the surrounding physical environment, including the trajectories and configuration of people, provides rich, relevant, streaming context for the interaction. Our long-term goal is to construct computational models that can provide the core skills needed for handling such situated interaction in dynamic multiparty settings, and work with people with the etiquette, fluidity and social awareness expected in human-human interactions.

In the following two subsections, we discuss the multiparty and situated aspects of open-world interaction in more detail, and we identify the challenges and opportunities that they frame. In Section 3, we review these challenges and outline a set of core competencies required for open-world dialog. Then, in Sections 4 and 5, we describe a prototype situated conversational agent that implements multiple components of an open-world dialog and review their operation in the receptionist setting.

2.1. Multiparty Aspect of Open-World Dialog

The assumption in spoken dialog research to date that only one user interacts with the system is natural for telephony-based spoken dialog systems and is reasonable for a large class of multimodal interfaces. In contrast, if we are interested in developing systems that can embed their input and interaction into the natural flow of daily tasks and activities, the one-user assumption can no longer be maintained.

The open world typically contains more than one relevant agent. Each agent may have distinct actions, goals, intentions, and needs, and these may vary in time. Furthermore, the open world is dynamic and asynchronous, *i.e.*, agents may enter or leave the observable world at any point in time, and relevant events can happen asynchronously with respect to current interactions.

The flow of considerations from single-user, closed-world systems to increasingly open worlds is highlighted graphically in Figure 2. Systems providing service in the open world will often have to have competencies for working with multiple people, some of whom may in turn be coordinating with others within and outside an agent's frame of reference. Such a competency requires the abilities to sense and track people over time, and to reason jointly about their goals, needs, and attention. We can categorize interactive systems based on the assumptions they make regarding the number and dynamics of relevant agents and parties involved in the interaction as follows:

- *Single-user interactive systems* engage in interaction with only one user at a time. Traditional telephony spoken dialog systems, as well as most multimodal interfaces such as multimodal mobile systems, *e.g.* [1, 26], multi-modal kiosks *e.g.* [9, 13], or embodied conversational agents *e.g.* [5] fall into this category.
- *Fixed multi-participant interactive systems* can interact with one or more participants at a given time. The number of participants in a given interaction is known in advance.
- *Open multi-participant interactive systems* can interact with one or more participants. Participants may leave or join an interaction at any given time.
- *Open multiparty interactive systems* further extend the class of open multi-participant systems in that they can engage in, pursue, and interleave multiple parallel interactions with several different parties. The receptionist interaction discussed earlier falls into this last category, as does the prototype system we shall discuss later, in Sections 4 and 5.

The pursuit of multi-participant and multiparty interactive systems brings to fore several research challenges. First, the multi-participant aspect adds a new dimension to several core dialog system problems like dialog management, turn taking, and language understanding. Current solutions for these problems typically rely on the single-user assumption and do not generalize easily to the multi-participant case. We also face entirely new types of prob-

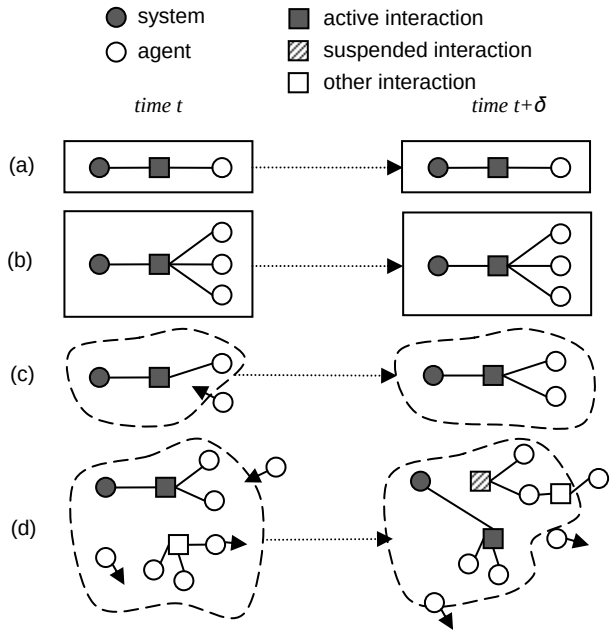


Figure 2. Conversational dynamics in: (a) single-user system; (b) a fixed multi-participant system; (c) an open multi-participant system, (d) an open multiparty system

lems, such as identifying the source and the target for each communicative signal in a multi-participant interaction, or handling engagement and disengagement in dynamic multi-participant settings. Moving from multi-participant to multiparty systems raises additional problems with respect to maintaining multiple interaction contexts, and triaging attention between multiple goals, parties and conversations. We shall discuss these new challenges in more detail in Section 3. Before that, we turn our attention to a second central feature of open-world dialog: the situated nature of the interaction.

2.2. Situated Aspect of Open-World Dialog

Dialog systems developed to date operate within narrow, predefined communication contexts. For example, in telephony-based spoken dialog systems, the audio-only channel limits the available context to the information that can be gained through dialog. In some cases, a stored user profile might provide additional information. Multimodal mobile systems might also leverage additional context from simple sensors like a GPS locator.

In contrast, systems designed to be effective in the open world will often need to make inferences about multiple aspects of the context of interactions by considering rich streams of evidence available in the surrounding environment. Such evidence can be observed by standing sensors or actively collected to resolve critical uncertainties. People are physical, dynamic entities in the world, and the system must reason about them as such, and about the conversational scene as a whole, in order to successfully and naturally manage the interactions. Concepts like presence, identity, location, proximity, trajectory, attention, and inter-agent relationships all play fundamental roles in shaping natural, fluid interactions, and need to become first-order objects in a theory of open-world dialog.

Like the multiparty aspect of open-world dialog, the situated nature of the interaction raises a number of new research challenges and brings novel dimensions to existing problems. One challenge is creating a basic set of physical and situational awareness skills. Interacting successfully in open environments requires that information from multiple sensors is fused to detect, identify, track and characterize the relevant agents in the scene, as well as the relationships between these agents. At a higher level, models for inferring and tracking the activities, goals, and long-term plans of these agents can provide additional context for reasoning within and beyond the confines of a given interaction, and optimizing assistance to multiple parties. Finally, new challenges arise in terms of integrating this streaming context in various interaction processes, like the engagement or disengagement process, turn taking, intention recognition, and multiparty dialog management.

3. Core Competencies for Open-World Dialog

We anchor our discussion of challenges for open-world dialog in Clark's model of language interaction [7]. With this model, natural language interaction is viewed as a joint activity in which participants in a conversation attend to each other and coordinate their actions on several different levels to establish and maintain mutual ground. Components of Clark's perspective are displayed in Figure 3. At the lowest level (*Channel*), the participants coordinate their actions to establish, maintain or break an open communication channel. At the second (*Signal*) level, participants coordinate the presentation and recognition of various communicative signals. At the third (*Intention*) level, participants coordinate to correctly interpret the meaning of these signals. Finally, at the fourth (*Conversation*) level, participants coordinate and plan their overall collaborative activities and interaction.

Successfully engaging in dialog therefore requires a minimal set of competencies at each of these levels. And indeed, most spoken dialog systems are organized architecturally in components that closely mirror Clark's proposed model: a voice activity detector and speech (and/or gesture) recognition engine identify the communicative signals, a language understanding component which extracts a corresponding semantic representation, and a dialog management component which plans the interaction.

We review in the rest of this section challenges raised by the multiparty and situated aspects of open-world dialog in each of these areas. We begin at the *Channel* level.

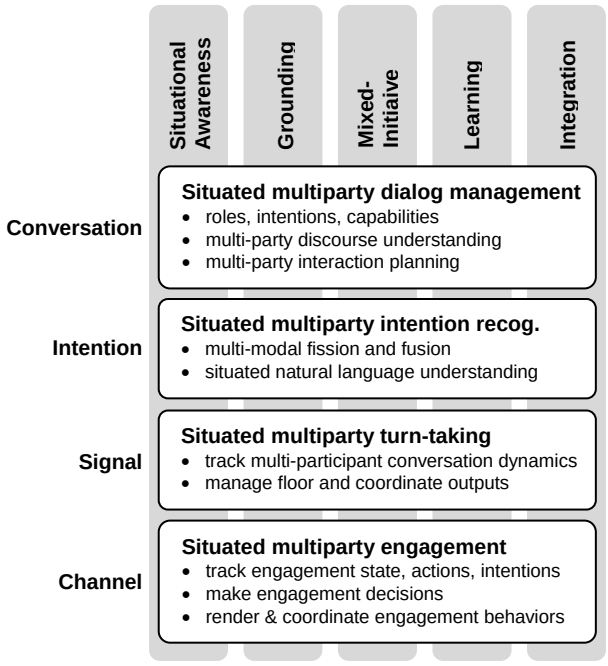


Figure 3. Core competencies for open-world dialog

3.1. Situated Multiparty Engagement

As a prerequisite for interaction, participants in a dialog must coordinate their actions to establish and maintain an open communication channel. In single-user systems this problem is often solved in a trivial manner. For instance, in telephony-based spoken dialog systems the channel is assumed to be established once a call has been received. Similarly, multimodal mobile applications oftentimes resolve the channel problem by using a push-to-talk solution.

Although these solutions are sufficient and perhaps natural in closed, single-user contexts, they become inappropriate for systems that must operate continuously in open, dynamic environments. We argue that such systems should ideally implement a *situated multiparty engagement model* that allows them to fluidly engage, disengage and re-engage in conversations with one or more participants.

Observational studies have revealed that humans negotiate conversational engagement via a rich, 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 [2, 3, 14]. Successfully modeling this coordinated process requires that the system (1) can sense and reason about the engagement actions, state and intentions of multiple agents in the scene, (2) can make high-level engagement control decisions (such as whom to engage with and when), and (3) can render engagement decisions in low-level coordinated behaviors and outputs.

Models for sensing the engagement state, actions, and intentions of various agents in the scene are, to a large extent, predicated on the system's capabilities to understand the physical environment in which it is immersed, *i.e.* to detect, identify and track multiple agents, including their location, trajectory, focus of attention, and other engagement cues. Higher-level inferences about the long-term goals, plans and activities of each agent can also provide informative priors for detecting engagement actions.

Beyond the engagement sensing problem, at a higher level, the system must reason about the boundaries of each conversation and make real-time decisions about whom to engage (or disengage) with, and when. In a dynamic multiparty setting these decisions have to take into account additional streams of evidence, and optimize tradeoffs between the goals and needs of the multiple parties involved (*e.g.*, interrupting a conversation to attend to a more urgent one). In making and executing these decisions, the system must consider social and communicative expectations and etiquette. Finally, such high-level engagement decisions must be signalled in a meaningful, understandable manner to the relevant participants. For instance, in an embodied anthropomorphic agent, engagement actions have to be rendered into a set of corresponding behaviors (*e.g.*, establishing or breaking eye contact, changing body posture, generating subtle facial expressions, or issuing greetings) that must often be coordinated at the millisecond scale.

3.2. Situated Multiparty Turn Taking

Going one level up in Clark's model, at the *Signal* level, the system must coordinate with other participants in the conversation on the presentation and recognition of communicative signals (both verbal and non-verbal, *e.g.*, gestures and emotional displays.) The coordinated process by which participants in a conversation take turns to signal to each other is known as turn-taking and has been previously investigated in the conversational analysis and psycholinguistics communities, *e.g.* [12, 18]. While computational models for turn-taking [19, 23, 24] have also been proposed and evaluated to date, most current systems make simplistic one-speaker-at-a-time assumptions and have relied on voice activity detectors to identify when the user is speaking. Phenomena like interruptions or barge-ins are often handled using ad-hoc, heuristic solutions, which can lead to turn-overtaking issues and ultimately to complete interaction breakdowns even in single-user systems [6].

Open-world dialog requires the development of a computational, *situated multiparty turn-taking model*. On the sensing side, such a model should be able to track the multi-participant conversational dynamics in real time by fusing lower-level evidence streams (*e.g.*, audio and visual). The model should be able to identify the various communicative signals as they are being produced, and, in a multi-participant setting, identify the sender, the addressees (and potentially the over-hearers) for each signal. In addition, the model should be able to track who has the conversational floor, i.e., the right to speak, at any given point in time. On the control side, a multiparty situated turn-taking model should make real-time decisions (that

are in line with basic conversational norms) about when the system can or should start or stop speaking, take or release the conversational floor, etc. Finally, the model must coordinate the system's outputs and render them in an appropriate manner. For instance, in an embodied conversational system, speech, gaze, and gesture must be tightly coordinated to signal that the system is addressing a question to two conversational participants, or to indicate that the system is trying to currently acquire the floor.

3.3. Situated Multiparty Intention Recognition

At the *Intention* level, a dialog system must correctly interpret the meaning of the identified communicative signals. In traditional dialog systems this is the realm of the language understanding component. Given the static, relatively limited communication context, the language understanding challenges tacked in traditional dialog systems have been typically limited to generating an appropriate semantic representation for the hypotheses produced by a speech recognizer, and integrating this information with the larger dialog context. In certain domains, issues like ellipsis and anaphora resolution also have played an important role. Systems that use multiple input modalities (*e.g.*, speech and gesture) face the problem of multi-modal fusion at this level: signals received from the lower levels must be fused based on content and synchronicity into a unified semantic representation of the communicative act.

The physically situated nature of open-world dialog adds new dimensions to each of these problems. In situated interactions, the surrounding environment provides rich streaming context that can oftentimes be leveraged for intention recognition. For instance, in the receptionist domain, an interactive system might be able to infer intentions based on identity (John always needs a shuttle at 3pm on Wednesday), spatiotemporal trajectories (people entering the lobby from inside the building are more likely to want a shuttle reservation than people entering the lobby from outside the building), clothing and props (a formally-dressed person is more likely a visitor who wants to register than an internal employee), and so on. Novel models and formalisms for reasoning about the streaming context and fusing it with the observed communicative signals to decode intentions and update beliefs are therefore required.

An additional challenge for open-world dialog is that of situated language understanding. Physically situated systems might often encounter deictic expressions like "Come here!" "Bring me the red mug," and "He's with me", etc. Resolving these referring expressions requires a set of language understanding skills anchored in spatial reasoning and a deep understanding of the relevant entities in the surrounding environment and of the relationships between these entities. The same holds true for pointing gestures and other non-verbal communicative signals.

3.4. Situated Multiparty Dialog Management

At the fourth level, referred as the *Conversation* level, participants coordinate the high-level planning of the interaction. This is the realm of dialog management, a problem that has already received significant attention in the spoken dialog systems community, *e.g.* [4, 6, 8, 16, 17, 20]. However, with the exception of a few incipient efforts [15, 25], current models make an implicit single-user assumption, and do not deal with the situated nature of the interactions.

One of the main challenges for open-world spoken dialog systems will be the development of models for *mixed-initiative, situated multiparty dialog management*. To illustrate the challenges in this realm, consider the situation in which a visitor, accompanied by her host, engages in dialog with a receptionist to obtain a visitor's badge. In order to successfully plan multi-participant interactions, the dialog manager must model and reason about the goals and needs of different conversational partners (*e.g.* get a badge versus accompany the visitor), their particular roles in the conversation (*e.g.* visitor versus host), their different knowledge and capabilities (*e.g.* only the visitor knows the license plate of her car). Individual contributions, both those addressed to the system, and those that the participants address to each other, need to be integrated with a larger multi-participant discourse and situational context.

Mixed-initiative interaction [10] with multiple participants requires that the system understands how to decompose the task at hand, and plan its own actions accordingly (*e.g.* directing certain questions only to certain participants, etc.) All the while, the dialog planning component must be able to adapt to the dynamic and asynchronous nature of the open-world. For instance, if the visitor's host disengages momentarily to greet a colleague in the lobby, the system must be able to adjust its conversational plans on-the-fly to the current situation (*e.g.* even if it was in the middle of asking the host a question at that point)

Handling multiparty situations (*e.g.* a third participant appears and engages on a separate topic with the host) requires that the system maintain and track multiple conversational contexts, understand potential relationships between these contexts, and is able to switch between them. Furthermore, providing long-term assistance requires that the system is able to reason about the goals, activities and long-term plans of individual agents beyond the temporal confines of a given conversation. To illustrate, consider another example from the receptionist domain: after making a reservation, a user goes outside to wait for the shuttle. A few minutes later the same user re-enters the building and approaches the reception desk. The receptionist infers that the shuttle probably did not arrive and the user wants to recheck the estimated time of arrival or to make another reservation; she glances towards the user and says "Two more minutes." Inferences about the long-term plans of various agents in the scene can provide valuable context for the streamlining the interactions.

3.5. Other Challenges

So far, we have made use of Clark's four-level model of grounding to identify and discuss a set of four core competencies for open-world spoken dialog systems: multiparty situated engagement models, multiparty situated turn-taking models, situated intention recognition, and mixed-initiative multiparty dialog management. However, developing an end-to-end system requires more than a set of such individual models. A number of additional challenges cut across each of these communicative processes. In the remainder of this section, we briefly review five challenges: situational awareness, robustness and grounding, mixed-initiative interaction, learning, and integration.

Given the situated aspect of open-world interaction, a major overarching challenge for open-world spoken dialog systems is that of *situational awareness*. As we have already seen, the ability to fuse multiple sensor streams and construct a coherent picture of the physical surrounding environment and of the agents involved in the conversational scene plays a fundamental role in each of the conversational processes we have previously discussed. Open-world systems should be able to detect, identify, track and characterize relevant agents, events, objects and relationships in the scene. Models for reasoning about the high-level goals, intentions, and long-term plans of the various agents can provide additional information for establishing rapport and providing long-term assistance. In contrast to traditional work in activity recognition (*e.g.,* in the vision or surveillance community), interactive systems also present opportunities for eliciting information on the fly and learning or adapting such models through interaction.

A second major challenge that spans the communicative processes discussed above is that of dealing with the uncertainties resulting from sensor noise and model incompleteness. Uncertainties abound even in human-human communication, but we are generally able to monitor the conversation and re-establish and maintain mutual ground. Open-world dialog systems can benefit from the development of similar *grounding models* that explicitly represent and make inferences about uncertainties at different levels and, when necessary, take appropriate actions to reduce the uncertainties and re-establish mutual ground.

A third important overall challenge is that of *mixed-initiative interaction***.** So far, we have discussed the notion of mixed-initiative in the context of the dialog management problem. It is important to notice though that, like situational awareness and grounding, the notion of mixed-initiative pervades each of the communicative processes we have discussed. At each level, the system's actions need to be tightly coordinated with the actions performed by the other agents involved in the conversation. Examples include the exchange of cues for initiating or breaking engagement, or "negotiating" the conversational floor. Mechanisms for reasoning about and managing initiative will therefore play a central role in each of these layers.

A fourth important challenge that cuts across the four competencies discussed above is that of *learning*. Given the complexities involved, many of the models we have discussed cannot be directly authored but must be learned from data. Ideally, we would like to build systems that learn throughout their lifetimes, directly from interaction, from their experience, without explicit supervision from their developers. Furthermore, such systems should be able to share the knowledge they acquire with each other.

Finally, another challenge not be underestimated is that of *system integration*, of weaving together all these different components into an architecture that is transparent, modular, and operates asynchronously and in real-time to create a seamless natural language interaction.

4. A Prototype System

We now describe a concrete implementation of a prototype system, named the Receptionist. The Receptionist is a situated conversational agent that can fluidly engage with one or more people and perform tasks typically handled by front-desk receptionists (*e.g.,* making shuttle reservations, registering visitors, providing directions on campus, etc.) at our organization. In previous work in this domain [11], we have investigated the use of a hierarchy of Bayesian models and decision-theoretic strategies for inferring intentions and controlling question asking and backtracking in dialog. Here, we focus on exploring the broader challenges of open-world dialog.

The front-desk assistance domain has several properties that make it a valuable test-bed for this endeavor. The interactions happen in an open, public space (building lobbies) and frequently involve groups of people. The complexity of the tasks involved ranges from the very simple, like making shuttle reservations, to more difficult ones requiring complex collaborative problem solving skills. Finally, a deployed system could provide a useful service and its wide adoption would create a constant stream of ecologically-valid real-world interaction data.

In the rest of this section, we describe the Receptionist system, and discuss an initial set of models that address the core competencies for open-world dialog we have previously outlined. In particular, we focus our attention on the situational awareness, engagement, and multi-participant turn-taking capabilities of this system. Despite the preliminary and sometimes primitive nature of these models (they represent only a first iteration in this long-term research effort), as we shall see in Section 5, when weaved together, they showcase the potential for seamless natural language interaction in open, dynamic environments.

We begin with a high-level overview of the hardware and software architecture. The current prototype takes the form of an interactive multi-modal kiosk, illustrated in Figure 4. On the input side, the system uses four sensors: a wide-angle camera with 140° field of view and a resolution

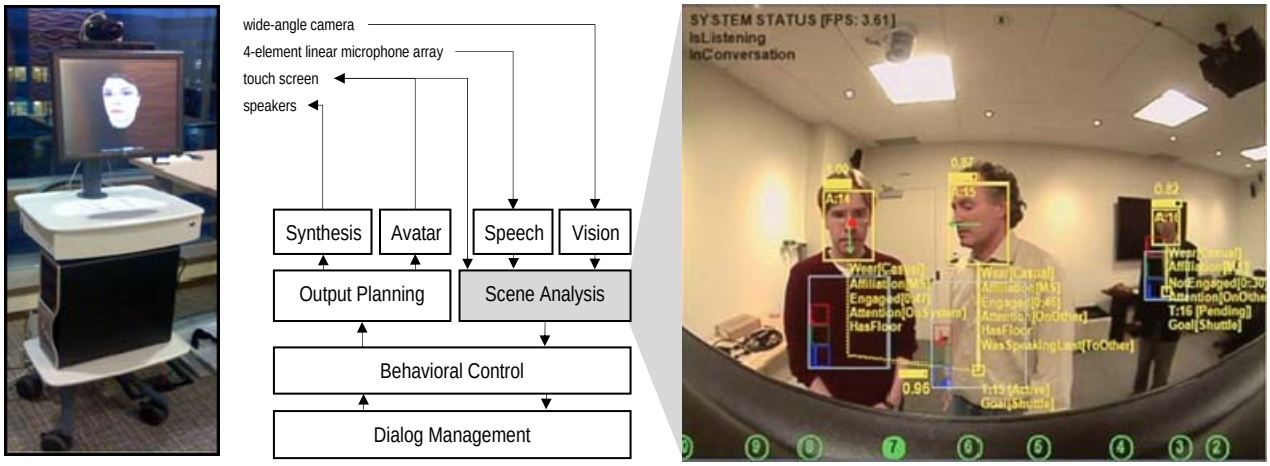


Figure 4. Receptionist system: (a) prototype, (b) architectural overview, and (c) runtime conversational scene analysis

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. As output, the system displays a realistic talking avatar head, which is at times complemented by a graphical user interface (*e.g.* when speech recognition fails the GUI is displayed and users can interact via the touch-screen – see Figure 5.c). The system currently runs on a 3.0GHz dual-processor Intel Xeon machine (total 8 cores).

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 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 together), and tracking the current conversational context at different levels (*e.g.* who is currently engaged in a conversation, who is waiting to engage, who has the conversational floor, who is currently speaking to whom, etc.) The individual models that implement these functions are described in more detail in the sequel.

The conversational scene analysis results are then forwarded to the control level, which is structured in a two-layer reactive-deliberative architecture. The lower-level, reactive layer implements and coordinates various low-level behaviors (*e.g.* for engagement and conversational floor management, for coordinating spoken and gestural outputs, etc). The higher-level, deliberative layer makes conversation control decisions, planning the system dialog moves and high-level engagement actions.

4.1. Situational Awareness

The system currently implements the following situational awareness capabilities.

Face detection and tracking. A multiple face detector and tracker are used to detect and track the location $x_a(t)$ of each agent a in the scene. The face 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. These vision algorithms run on a scaled-up image (1280x960 pixels), which allows 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 face confidence score $fc_a(t)$, which is used to prune out false detections but also 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 runs on a cropped region of interest encompassing the agent's face. In group conversations, 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 direct conditional model is used to infer whether the attention of each agent in the scene is oriented towards the system or not: $P(foa_a(t)|fc_a(t), \overline{\omega_a}(t))$. This inference is currently based on a logistic regression model that was trained using a hand-labelled 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.)

Agent characterization. In addition to face detection and tracking, the system also performs a basic visual analysis of the clothing for each detected agent. The probability that the agent is formally or casually dressed $P(formal_a(t))$ is estimated based on the color variance in a rectangular patch below the face (*e.g.* if a person is wearing a suit, this typically leads to high variance in this image patch). This information is further used to infer the agent's likely affiliation, based on a simple conditional model $P(affiliation_a(t)|formal_a(t))$. Casually dressed agents are more likely to be Microsoft employees; formally dressed ones are more likely to be visitors.

Group inferences. Finally, the Receptionist system also performs a pairwise analysis of the agents in the scene to infer group relationships. The probability of two agents being in a group together $P(group(a_1, a_2))$ is computed by a logistic regression model that was trained on a hand-labelled dataset. The model uses as features the size, location and proximity of the faces, but also observations collected through interaction. For instance, the system might ask a clarification question like "Are the two of you together?" A positive or negative response to this question is also used as evidence by the group inference model.

4.2. A Multiparty Situated Engagement Model

We now turn our attention to the problem of *engagement* [21], the process by which participants in a conversation establish, maintain and terminate their interactions (corresponding to the first level of coordinated action in Clark's language interaction model).

The engagement model currently used in the Receptionist system is centered on a reified notion of *interaction*, defined here 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 an existing interaction, and current participants may leave an interaction. The system is actively engaged in at most one interaction at a time, but it can simultaneously keep track of additional, suspended interactions. Engagement is then viewed as the joint activity of the system and its users by which interactions are initiated, terminated, suspended, resumed, joined or abandoned.

To manage this coordinated process, the system: (1) constantly monitors the engagement state, actions and intentions of surrounding agents, (2) makes high-level decisions about whom to engage (or disengage) with and when, and (3) renders these decisions via behaviors such as establishing or breaking eye contact, issuing and responding to verbal greetings, etc. In the following subsections, we discuss each of these components in more detail.

4.2.1. Engagement State, Actions, and Intentions

The basis for making engagement decisions is provided by a model that tracks the engagement state $ES_a(t)$, actions $EA_a(t)$ and intentions $EI_a(t)$ for each agent in the scene.

The engagement state of an agent $ES_a(t)$ is modeled as a deterministic variable with two possible values: *engaged* and *not-engaged*, and is updated based on the joint actions of the agent and the system. The state transitions to *engaged* when both the system and an agent take an engaging action. On the other hand, disengagement can be a unilateral act: if either the system or an engaged agent take a disengaging action, the state transitions to *not-engaged*.

A second engagement variable, $EA_a(t)$, models the actions that an agent takes to initiate, maintain and terminate engagement (i.e. to transition between engagement states). There are four possible engagement actions: *engage*, *no-action*, *maintain*, *disengage*. An agent can take the first two actions only from the *not-engaged* state and the last two only from the *engaged* state. Currently, a direct conditional model $P(EA_a(t)|ES_a(t), \Psi(t))$ is used to estimate an agent's engagement action based on the current engagement state and additional evidence $\Psi(t)$ gathered from various sensors and processes in the system. Examples include the detection of greetings or calling behaviors (*e.g.* "Hi!" or "Laura!"), the establishment or the breaking of a conversation frame (*e.g.* the agent approaches and positions himself in front of the system; or the agent departs), continued attention (or lack thereof) to the system, etc.

Apart from the engagement state and actions, the system also keeps track of a third variable, the engagement intention $EI_a(t)$ of each agent in the scene; this can be *engaged* or *not-engaged*. Intentions are tracked separately from actions since an agent might intend to engage the system, but not take a direct, explicit engagement action. A typical case is that in which the system is already engaged in an interaction and the participant is simply waiting in line. More generally, the engagement intention corresponds to whether or not the user would respond positively should the system initiate engagement. Currently, the engagement intention is inferred using a handcrafted direct conditional model $P(EI_a(t)|ES_a(t), EA_a(t), \Psi(t))$ that leverages information about the current engagement state and action, as well as additional evidence gleaned from the scene including the spatiotemporal trajectory of the participant, the level of sustained mutual attention, etc.

While the current models for sensing engagement actions and intentions are handcrafted, we are also investigating data-driven approaches for learning these models.

4.2.2. Engagement Decisions

Based on the inferred state, actions and intentions of the agents in the scene, as well as other additional evidence, the system makes high-level decisions about when and with whom to engage in interaction. The system's engagement action-space at contains the same four actions previously discussed. The actual surface realization of these actions in terms of low-level behaviors, such as greetings, making or breaking eye contact, etc. is discussed in more detail in the following subsection.

As the Receptionist system operates in an open, multiparty environment, the engagement decisions can become quite complex. For instance, new participants might arrive and wait to engage while the system is already engaged in an interaction; in some cases, they might even actively try to barge-in and interrupt the current conversation. In such cases, the system must reason about the multiple tasks at hand, and balance the goals and needs of multiple participants in the scene and resolve various trade-offs, for instance between continuing the current interaction and temporarily interrupting it to address a new (perhaps shorter and more urgent task).

Currently, a simple heuristic model is used for making these decisions. If the system is not currently engaged in an

interaction, it conservatively waits for a user to initiate engagement (*e.g.* $EA_a(t)$=*engage*), before making the decision to engage. In addition, if the system is currently engaged in a conversation interaction, but other agents are present and waiting to engage (*e.g.* $EI_a(t)$=*engaged*, $EA_a(t)$=*no-action*), the system may suspend the current interaction to momentarily engage a waiting agent to either let them know that they will be attended to momentarily, or to inquire about their goals (this is illustrated in more detail in Section 5.) This decision is made by taking into account the appropriateness of suspending the current conversation at that point, and the waiting time of the agent in the background. We are currently exploring more principled models for optimizing the scheduling of assistance to multiple parties under uncertainties about the estimated goals and needs, the duration of the interactions, time and frustration costs, social etiquette, etc.

4.2.3. Engagement Behaviors

Each high-level engagement decision (*e.g. Engage* / *Disengage*) is rendered into a set of coordinated lower-level behaviors, such as making and breaking eye contact, issuing greetings, etc.

The sequencing of these lower-level behaviors is highly dependent on the current situation in the scene, including the estimated engagement state, actions and intentions for each agent, the evolving state of the environment and system (*e.g.* is the system in a conversation or not, are there other agents in the scene, what is their focus of attention, etc.) For instance, consider the case when the system is not yet engaged in any conversations and a high-level decision is made to engage a certain agent. If mutual attention has already been established, the *engage* behavior triggers a greeting. In contrast, if the agent's focus of attention is not on the system, the *engage* behavior attempts to draw the agent's attention by gazing towards him or her and saying "Excuse me!" in a raised voice. After the initial salutation the system monitors the spatiotemporal trajectory of the agent, and, if the agent approaches the system, establishes or maintains mutual attention, the *engage* behavior completes successfully; the agent's engagement state is updated to *engaged*. Alternatively if a period of time elapses and the agent does not establish mutual attention (or leaves the scene), the *engage* behavior completes with failure (which is signalled to the higher engagement control layer). The system implements several other engagement and disengagement behaviors dealing with agents joining or leaving an existing conversation. While a full description of these behaviors is beyond the scope of this paper, instances of various engagement behaviors are illustrated in the example discussed in Section 5.

4.3. Multi-Participant Turn Taking

While engaged in a conversation, the system coordinates with other conversational participants on the presentation and recognition of various communicative signals. Our current prototype attends to verbal signals (*i.e.*, spoken utterances) and to signals received from the graphical user interface, which can be accessed via the touch-screen. On the output side, the system coordinates spoken outputs with gaze and various gestures such as smiles, and furrowed or questioning eye-brows.

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 integrates 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 the system 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 $P(T_u = system|foa_{S_u}(t))$. If the user's focus of attention stays on the system, the utterance is assumed to be addressed to the system; otherwise, the utterance is assumed to be directed towards the other participants engaged in the conversation. Touch events detected by the graphical user interface are assumed to be generated by the closest agent, and addressed to the system.

In order to fluidly coordinate its own outputs (*e.g.* spoken utterances, gestures, GUI display) with the other agents engaged in the conversation, the system implements a simple multiparty situated turn-taking model. The model tracks whether or not each engaged agent currently holds the conversational floor $FS_a(t)$ (*i.e.* has the right to speak), and what the floor management actions each engaged agent takes at any point in time $FA_a(t)$: *No-Action*, *Take-Floor*, *Release-to-System*, *Release-to-Other*, *Hold-Floor*. These actions are inferred based on a set of handcrafted rules that leverage information about the current state of the floor $\{FS_a(t)\}_a$, the current utterance u, its speaker S_u and its addressees T_u. For instance, a *Take-Floor* action is detected when 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 coordinates its output 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 floor management action is invoked for acquiring the floor. The lower level behaviors render this action by coordinating the avatar's gaze, gesture and additional spoken signals (*e.g.* "Excuse me!", if the system is trying to take

the floor but a participant is holding it and speaking to another participant).

The current multi-participant turn-taking model is an initial iteration. It employs 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 through 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: $P(S_u, T_u, foa_{\{a\}}(t), FS_{\{a\}}(t), FA_{\{a\}}(t)|\Psi(t))$

4.4. Situated Intention Recognition

To infer user goals and intentions, the Receptionist system makes use of several hybrid belief updating models that integrate streaming evidence provided by the situational context, with evidence collected throughout the dialog. For instance, the system relies on a conditional goal inference model $P(G_a|affiliation_a, group(a, a_i), SG_a)$ that currently takes that takes into account the estimated actor affiliation and whether or not the actor is part of a larger group (e.g. Microsoft employees 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 - SG_a - 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 runtime behavior of these models is illustrated in more detail in the following section.

5. A Sample Interaction

We now illustrate how the models outlined in the previous section come together to create a seamless multiparty situated interaction, by describing a sample interaction with the receptionist system. Figure 5 shows several successive snapshots from a recorded interaction, with the runtime annotations created by the various models, as well as a capture of the system's display and a transcript of the conversation. A full video capture is available online [22].

Initially two participants are approaching the system (A14 and A15 in Figure 5). The system detects and tracks their location. As the users get closer and orient their attention towards the system, the engagement model indicates that they are performing an engaging action. In response, the avatar triggers an engaging behavior, greets them and introduces itself (line 3 in Figure 5).

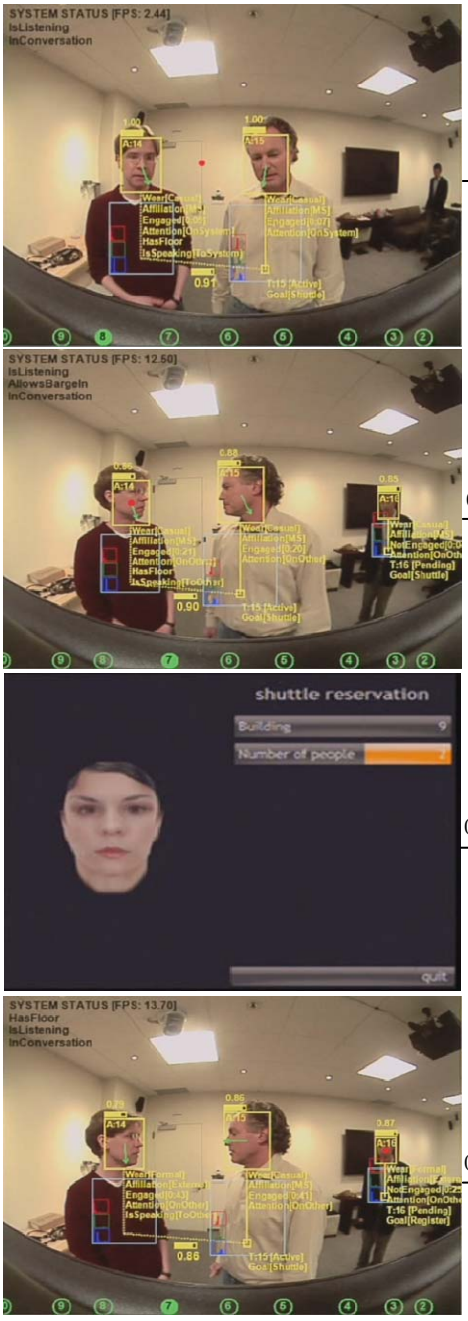
After the initial greeting, the system attempts to ground the goals of the two participants. The group inference model indicates that, with high likelihood (0.91 in Figure 5.a) the two participants are in a group together. The clothing and affiliation models indicate that the two participants and dressed casually, and therefore most likely Microsoft employees. Based on this information, the system infers that the participants most likely want a shuttle. Since the likelihood of the shuttle goal does not exceed the grounding threshold, the system confirms this information through dialog, by glancing at the two participants and asking: "Do you need a shuttle?" A14 confirms.

Next, the system asks "Which building are you going to?" At this point (see also Figure 5.b) the first participant (A14) turns towards the second one (A15) and initiates a side conversation (lines 8-12). By fusing information from the microphone array, the face detector and pose tracker, the multiparty turn-taking model infers that the two participants are talking and releasing the floor to each other. Throughout this side conversation (lines 8-12) the avatar's gaze follows the speaking participant. In addition, the recognition system is still running and the system overhears the building number from this side conversation. When the two participants turn their attention again towards the system, the turn-taking model identifies a *Release-To-System* floor action. At this point, the system continues the conversation by confirming the overheard information: "So you're going to 9, right?" A14 confirms again.

Next, the system grounds how many seats are needed for this reservation. Here, a belief updating model fuses information gathered from the scene analysis with information collected through interaction. Based on the scene, the system infers that most likely this shuttle reservation is for two people (A14 and A15). The likelihood however does not exceed a grounding threshold (since at this point a third agent has already appeared in the background – A16). The system therefore confirms the number of seats through dialog, by asking "And this is for both of you, right?" Once the number of people is grounded, the system notifies A14 and A15 that it is currently making a reservation for them.

As we have already noted, while A14 and A15 where engaged in the side conversation (lines 8-12), a new participant (A16) entered the scene – see Figure 5.b. When the new participant appears, the system glances for a fraction of a second at him (this is a hard-coded reactive behavior). The group models indicate that A16 is most likely not in a group with A14 and A15. The clothing and affiliation models for A16 indicate that this participant is dressed formally and therefore most likely to be an external visitor. As a consequence, the activity and goal models indicate that A16 is waiting for the receptionist with the intention to register.

After the avatar notifies A14 and A15 that it is making their shuttle reservation, these two participants turn again to each other and begin another side conversation. The system decides to temporarily suspend its conversation with



(a)

(b)

(c)

(d)

1	A14, A15:	*[approaching, talking to each other]*
2	A14, A15:	*[getting close and turning their attention to the system]*
3	S:	Hi, my name is Laura and today I'm here to help the receptionist with shuttle reservation
4	A14:	I'm Dan and this is Eric
5	S:	*[glancing at both participants]* Nice to meet you. Do you need a shuttle?
6	A14:	Yes
7	S:	*[turns to A14]* Which building are you going to?
8	A14:	*[turns to A15]* I forget where are we going?
9	S:	*[quickly glances at A16, who just entered the scene; then follows the side conversation that A14 and A15 are having, by turning its attention towards the speaking participant]*
10	A15:	*[turns to A14]* I think it's building 9?
11	A14:	*[still looking at A15]* Are you sure?
12	A15:	*[still looking at A14]* Yes
13	A14,A15:	*[both turn back towards the system]*
14	S:	*[glancing at both A14 and A16]* So you're going to building 9, right?
15	A14:	Yes
16	S:	And this if for both of you, right?
17	A14:	Yes
18	S:	I'm making a shuttle reservation for building 9 for 2 people. In case you want to correct anything, say or press start over.
19	S:	*[turning to A16]* Are you visiting someone?
20	A16:	Yes
21	S:	I'll help you register in just a moment.
22	S:	*[turning back to A14 and A15]* Excuse me.
23	A14,A15:	*[turning their attention back to the system]*
24	S:	You will be on shuttle 53. It should be here any second.
25	A14:	All right, thank you. Bye
26	S:	*[glancing at A14 and A15]* Bye-bye
27	A16:	*[approaching]* Hi

Figure 5. Sample interaction with Receptionist.

A14 and A15 and engages with A16 by asking "Are you visiting someone?" A16 confirms, and the system notifies A16 that it will help with the registration momentarily. The decision to suspend the conversation with A14 and A15 is taken by the high-level engagement control model, which is aware of the fact that the current conversation is interruptable at this point (the system is waiting for the shuttle backend to respond and A14 and A15 are talking to each other), and that, with high likelihood, there is an additional participant in the scene (A16) waiting for assistance.

After the system notifies A16 that it would attend to his needs momentarily (line 22), the shuttle backend responds with the initial reservation. The system turns its attention again at A14 and A15, and attempts to resume that conversation, by invoking a corresponding engagement behavior. Since the two participants are still talking to each other and not paying attention to the system, the *Resume-Conversation* behavior triggers an "Excuse me!" prompt (line 22). As soon at A14 and A15's attention turns back to the system, the avatar provides the information about the shuttle number and estimated time of arrival (line 24). The

two participants then disengage and the system turns its attention back to and engages with A16.

Conclusion and Future Work

We have outlined a research agenda aimed at developing computational systems that can interact naturally and provide assistance with problem-solving needs over extended periods of time in open, relatively unconstrained environments. We first introduced the pursuit and challenges of developing systems competent in *open-world dialog*—with the ability to support conversation in an open-world context, where multiple people with different needs, goals, and long-term plans may enter, interact, and leave an environment, and where the physical surrounding environment typically provides streaming evidence that is important for organizing and conducting the interactions.

The dynamic, multiparty and situated nature of open-world dialog brings new dimensions to traditional spoken dialog problems, like turn-taking, language understanding and dialog management. We found that existing models are limited in that they generally make an implicit single-user assumption and are not equipped to leverage the rich streaming context available in situated systems. Open-world settings pose new problems like managing the conversation engagement process in a multiparty setting, scheduling assistance to multiple parties, and maintaining a shared frame that includes inferences about the long-term plans of various agents--inferences that extend beyond the confines of an acute interaction.

To provide focus as well as an experimental testbed for the research agenda outlined in this paper, we have developed a prototype system that displays several competencies for handling open-world interaction. The prototype weaves together a set of early models addressing some of the open-world dialog challenges we have identified, and showcases the potential for creating systems that can interact with people on problem-solving needs with the ease and etiquette expected from a human.

We take the research agenda and the prototype described in this paper as a starting point. We plan to investigate the challenges we have outlined, and to develop and empirically evaluate computational models that implement core competencies for open-world dialog. We hope others will join us on the path towards a new generation of interactive systems that will be able embed interaction and computation deeply into the natural flow of daily tasks, activities and collaborations.

ACKNOWLEDGMENTS

We would like to thank George Chrysanthakopoulos, Zicheng Liu, Tim Paek, Qiang Wang, Cha Zhang for their contributions, useful discussions, and feedback.

REFERENCES

[1] A. Acero, N. Bernstein, R. Chambers, Y-C Ju, X. Li, J. Odell, P. Nguyen, O. Scholtz, G. Zweig. Live Search for Mobile: Web Services by Voice on the Cellphone. in Procs ICASSP'08. Las Vegas (2008)

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

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

[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] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, and H. Yan. Embodiment in Conversational Interfaces: Rea, in Procs of CHI'99, Pittsburgh, PA, (1999).

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

[7] H.H. Clark, and E.F. Schaefer. Contributing to Discourse. Cognitive Science. 13. (1989)

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

[9] J. Gustafson, N. Lindberg, and M. Lundeberg. The august spoken dialogue system, in Procs. Eurospeech'99, Budapest, Hungary (1999).

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

[11] E. Horvitz and T. Paek. A Computational Architecture for Conversation, in Procs of 7th International Conference on User Modeling, Banff, Canada (1999)

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

[13] M. Johnston, S. Bangalore. MATCHKiosk: a multimodal interactive city guide, in Procs of ACL'04, Barcelona, Spain (2004).

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

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

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

[17] M. McTear. Spoken dialogue technology: enabling the conversational user interface, in ACM Computing Surveys 34(1):90-169.

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

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

[21] 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)

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

[23] 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)

[24] 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)

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

[26] V-Lingo Mobile - http://www.vlingomobile.com/downloads.html