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 framework 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 component 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

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

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.)

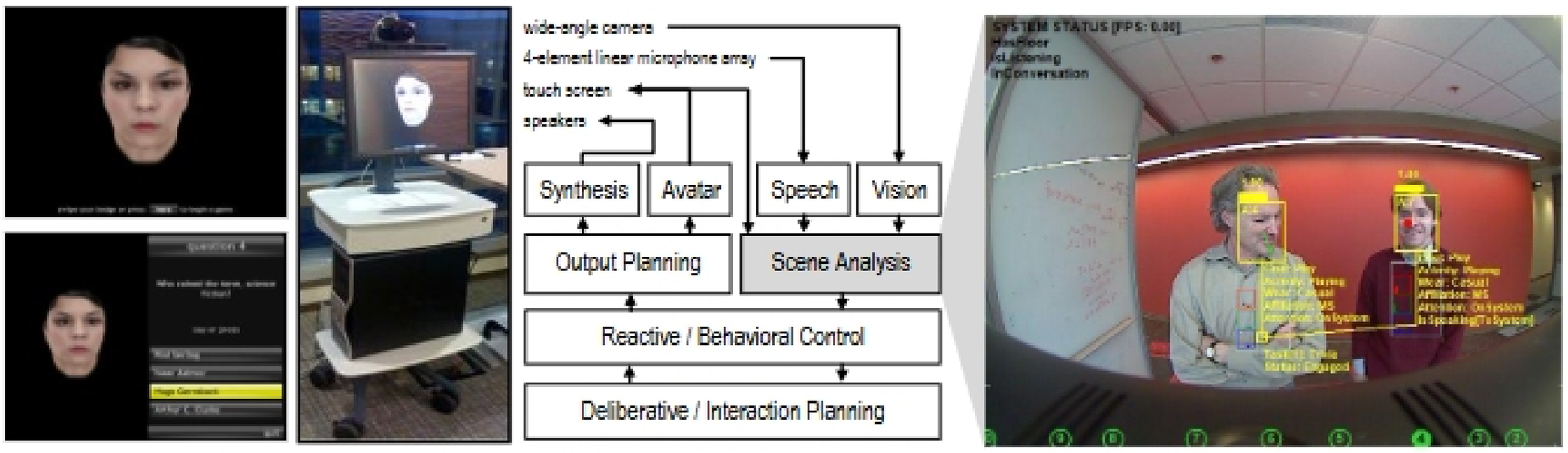

Figure 1. Hardware and software components within the overall architecture, and an illustration of scene analysis results

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.