Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
horvitz@microsoft.com

Abstract

We consider the challenge of predicting the engagement of people with an open-world dialog system, where one or more participants may establish, maintain, and break the communication frame. We show in particular how a system can learn to predict an intention to engage from multiple observations that are extracted from a visual analysis of people coming into the proximity of a system.

1 Introduction

We address the challenge of predicting the forthcoming engagement of people with conversational systems in relatively unconstrained environments, where multiple participants might come and go, establish, maintain and break the communication frame, and simultaneously interact with a system and with others. Examples of such *open-world* dialog (Bohus and Horvitz, 2009a) include people passing by an interactive billboard in a mall, robots in a home environment, intelligent home control systems, interactive systems that offer assistance and provide support during procedural tasks, etc.

With traditional *closed-world* dialog systems the engagement problem is generally resolved via simple, unambiguous signal. For example, engagement can be assumed in such systems when a phone call is answered by a telephony dialog system. Similarly, a push-to-talk button is a clear engagement signal in speech enabled mobile applications. These solutions are typically inappropriate however for systems that must operate continuously in open environments, working to engage and support different people and groups over time. Such systems should ideally be ready to initiate dialog in a fluid, natural manner and work to understand and engage users who are both close by and at a distance. Participants include both people with a standing plan to interact with a system, and those whom opportunistically decide to engage with a dialog system, in-stream with their other ongoing activities. They need to minimize false positives of engaging someone who may come into the proximity of a system or just be passing by the system, while also minimizing the unnatural delays and discontinuities that come with false negatives about engagement intentions.

(Bohus and Horvitz, 2009b) describes a computational model for supporting fluid engagement in open-world contexts. The proposed situated, multiparty engagement model harnesses components for sensing the engagement state, actions, and intentions of multiple participants in the scene, for making high-level engagement control decisions, and for rendering these decisions using appropriate, coordinated low-level behaviors, such as the changing pose and expressions of the face of an embodied agent. We shall focus on the sensing component of this larger model and describe an approach for automatically learning to detect engagement intentions from interaction.

2 Related Work

The challenges of engagement among people, and between people and computational systems, have received attention in several communities, including sociolinguistics, conversational analysis, and in the human-computer interaction communities. In an early treatise, Goffman (1963) discusses how people use cues to detect engagement in an effort to avoid the social costs of engaging in interaction with an unwilling participant. In later work, Kendon (1990a) presents a detailed investigation of greetings in human-human interaction, based on video sequences. Several stages of complex coordinated action (*pre-sighting, sighting, distance salutation, approach, close salutation*) are identified and discussed, together with the head and body gestures that they typically involve. In (1990b), Kendon introduces the notion of an *F-formation*, a pattern said to arise when "two or more people sustain a spatial and orientational relationship in which they have equal, direct, and exclusive

access," and discusses the role of F-formations in establishing and maintaining social interactions. Argyle and Cook (1976) as well as others (Duncan, 1972; Vertegaal et al., 2001) have identified and discussed the various functions of eye gaze in communication and in maintaining social engagement. Overall, this body of work suggests that engagement is a rich, mixed-initiative, and well-coordinated process that relies on non-verbal cues and signals, such as spatial trajectory and proximity, gaze and mutual attention, head and hand gestures, as well as verbal greetings.

More recently, a number of researchers have investigated issues of engagement in human-computer and human-robot interaction contexts. Sidner et al. (2004; 2005) define engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake," and conduct a user study that explores the process of maintaining engagement. They show that people directed their attention to a robot more often when the robot makes engagement gestures throughout an interaction, *i.e.* tracked the user's face, and pointed to relevant objects at appropriate times in the conversation.

Peters et al (2005a; 2005b) use an alternative definition of engagement as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction," and present the high-level schematics for an algorithm for establishing and maintaining engagement. The proposed algorithm highlights the importance of mutual attention and eye gaze in this process and relies on a heuristically computed *interest level* to decide when to start a conversation.

Michalowski et al (2006) propose and conduct experiments with a spatial model of engagement. The model is grounded in proxemics (Hall, 1966) and classifies relevant agents in the scene in four different categories based on their distance to the robot: *present* (standing far), *attending* (standing closer), *engaged* (next to the robot), and *interacting* (standing right in front of the robot). The robot's behaviors are conditioned on the four categories above: it turns towards attending people, greets engaged people and verbally prompts interacting people for input if they are not typing. The authors discuss several important lessons learned from an observational study conducted with the robot in a building lobby. They find that the fast-paced movements of people in the environment pose a number of challenges: often the robot greeted people that passed by too late (earlier anticipation was needed), or greeted people that did not intend to engage (more accurate anticipation was needed). The authors recognize that these limitations stem in part from their reliance on static models, and hypothesize that temporal information such as speed and trajectory may provide additional cues regarding a person's future engagement with the robot.

We expand in this paper our previous work on a situated multiparty engagement model (Bohus and Horvitz, 2009b), and focus our attention on a key issue in managing the engagement process: detecting whether or not a user intends to engage in an interaction with a system. We introduce an approach that significantly improves upon existing work (Peters 2005a, 2005b; Michalowski et. al, 2006) in several ways. Most importantly, we construct data-driven, predictive models from an array of observations that includes temporal features. The use of machine learning techniques allows a system to adapt to the specific characteristics of its physical location and to the behaviors of the surrounding population of potential participants. Finally, no developer supervision is required, as the supervision signal is extracted automatically, in-stream with the interactions.

3 Situated Multiparty Engagement Model

To set the broader context for the experiments on engagement, we shall briefly review the overall framework for managing engagement in an open-world setting. The engagement model outlined in (Bohus and Horvitz, 2009b) 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 in time. The system is actively engaged in at most one interaction at a time (with one or multiple participants), but it can simultaneously keep track of additional, suspended interactions. In this context, engagement is viewed as the process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume, or terminate an interaction.

Successfully managing this process requires that the system (1) senses and reasons about the engagement state, actions and intentions of multiple agents in the scene, (2) makes high-level engagement control decisions (*i.e.* about whom to engage or disengage with, and when) and (3) executes and signals these decisions to the other participants in an appropriate manner (*e.g.* via a set of coordinated behaviors such as gestures, greetings, etc.) The proposed model, illustrated in Figure 1, subsumes these three components.

The sensing subcomponent in the model tracks the engagement state, engagement actions, and engagement intention 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*. The state is updated based on the joint actions of the system and the agent.

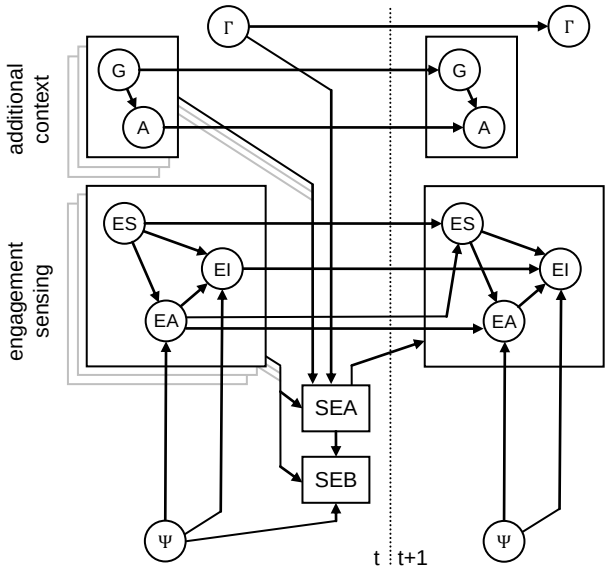

Figure 1. Graphical model showing key variables and dependencies in managing engagement.

A second engagement variable, $EA_a^i(t)$, models the actions that an agent takes to initiate, maintain or terminate engagement. There are four possible engagement actions: *engage*, *no-action*, *maintain*, *disengage*. These actions are tracked by means of a conditional probabilistic model that takes into account the engagement state $ES_a^i(t)$, the previous agent and system actions, as well as additional sensory evidence Ψ capturing committed engagement actions, such as: salutations (*e.g.* "Hi!"); calling behaviors (*e.g.* "Laura!"); the establishment or the breaking of an F-formation (Kendon, 1990b); 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, let us 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 dialog systems) that her intention is to engage in a new conversation once 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' (2005; 2005b) "interest level", *i.e.* to the value the agent attaches to being engaged in a conversation with the system. 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 pant, the level of sustained mutual attention, etc.

Based on the inferred engagement state, actions, and intentions of the agents in the scene, as well as other additional high-level evidence such as the agents' inferred goals (G), activities (A) and relationships (Γ), the proposed model outputs engagement actions – denoted by the SEA decision node in Figure 1. The action-space consists of the same four actions previously discussed: *engage*, *disengage*, *maintain* and *no-action*. At the lower level, the engagement decisions taken by the system are translated into a set of coordinated lower-level behaviors (SEB) such as head gestures, making eye contact, facial expressions, salutations, interjections, etc.

In related work (Bohus and Horvitz, 2009a; 2009b), we have demonstrated how this model can be used to effectively create and support multiparty interactions in an open-world context. We focus here our attention on one specific subcomponent in this framework: the model for detecting engagement intentions.

4 Approach

To illustrate the problem of detecting engagement intentions, consider for instance a situated conversational agent that examines through its sensors the scenes from Figure 3. How can the system detect whether the person in the image intends to engage in a conversation or is just passing-by? Studies of human-human conversational engagement (Goffman, 1963; Argyle and Cook, 1976; Duncan, 1972; Kendon, 1990, 1990b) indicate that people signal and detect engagement intentions by producing and monitor for a variety of cues, including sustained attention, trajectory and proximity, head and hand gestures, etc.

We shall investigate the value of employing machine learning to enable an open-world interactive system to learn to detect the specific patterns that characterize an engagement intention and thus, forthcoming engagement, directly from interaction. In general, as discussed in the previous section, the engagement intentions of an agent may evolve temporally under the proposed model, as a function of the various system actions and behaviors (*e.g.* an embodied system that makes eye contact, or smiles, or moves toward a participant might alter the engagement intention of that participant). In this work we concentrate on a simplified problem, in which the system's behavior is fixed (*e.g.* system always tracks people that pass by), and the engagement intention can be assumed constant within a limited time window.

The central idea of the proposed approach is to start by using a very conservative (*i.e.*, low false-positives) detector for engagement intentions, such as a push-to-engage button, and automatically gather sensor data surrounding the moments of engagement, together with labels that indicate whether someone actually engaged or not. In most cases the system eventually finds out if a person becomes engaged with it. If we assume that an intention to engage existed for a limited window of time prior to the moment of engagement, the collected data can then be used to learn a model for predicting this intentions ahead of the actual moment of engagement.

Previous work on detecting engagement intentions has focused on static heuristic models that leverage proximity and attention features (Peters, 2005, 2005b; Michalowski, 2006). The use of machine learning allows us to consider such observations as trajectory, speed, and attention of a potential participant over time. As previously discussed, psychologists have shown the important role of geometric relationships and trajectories in signaling and detecting engagement intentions. The patterns of engagement can therefore be highly dependent on the physical surroundings and on the placement of the system. The data-driven approach we propose enables a system to learn how to predict forthcoming engagement from interactions in any new environment, without any developer supervision.

5 Experimental Setup

To provide an ecologically valid basis for data collection and for evaluating the proposed approach, we developed a situated conversational agent and deployed it in the real-world. The system, illustrated in Figure 2, is an interactive multimodal kiosk that displays a realistically rendered avatar head. The avatar can engage and interact via natural language with one or more participants, and plays a simple game in which the users have to respond to multiple-choice trivia questions. The system, and sample interactions are described in more detail in (Bohus and Horvitz, 2009.)

The hardware and software architecture is also illustrated in Figure 2. Data gathered from a wide-angle camera, a 4-element linear microphone array, and a 19" touch-screen is 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. The system detects and tracks the location of multiple agents in the scene, tracks the head pose for engaged agents, and infers the focus of attention, activities, goals and (group) relationships among different agents in the scene. An in-depth description of these scene analysis components falls beyond the scope of this paper; more details are available in (Bohus and Horvitz, 2009). The scene analysis results are forwarded to the control level, which is structured in a two-layer

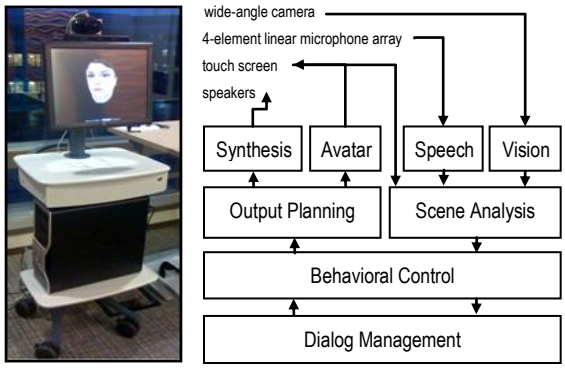

Figure 2. System prototype and architectural overview.

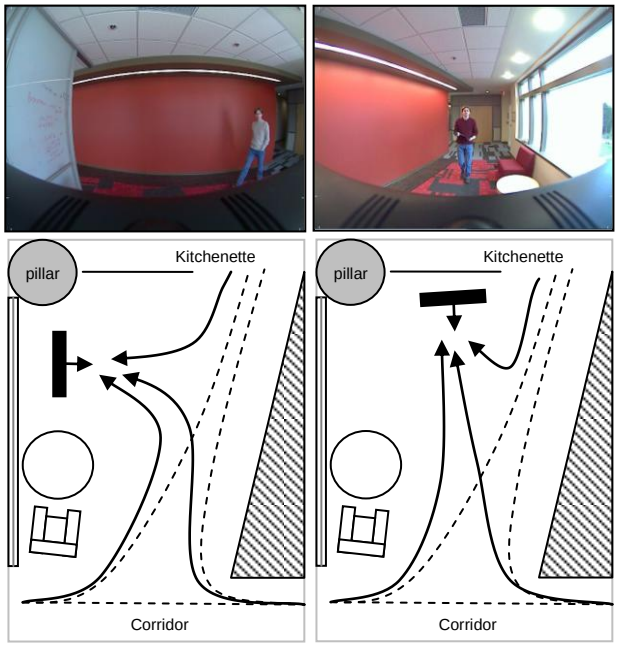

Figure 3. Placement and visual fields of view for *side* (right) and *front* (left) orientations.

reactive-deliberative architecture. The reactive layer implements and coordinates low-level behaviors, including engagement, conversational floor management and turn-taking, and coordinating spoken and gestural outputs. The deliberative layer plans the system's dialog moves and high-level engagement actions.

We deployed the system described above in an open-space near the kitchenette area in our building. As we were interested in exploring the influence of the spatial setup on the engagement models, we deployed the system in two different spatial orientations, illustrated together with the resulting visual fields of view in Figure 3. Even though the location is similar, the two orientations create considerable differences in the relative trajectories of people that go by (dashed lines) and people that engage with the system (continuous lines). In the *side* orientation, people typically enter the system's field of view and approach it from the sides. In the *front*

	Side	Front	Total
Size (hours:minutes)	83:16	75:15	158:32
# face traces	2025	1249	3274
# engaged	72	74	146
% engaged	3.55%	5.92%	4.46%
# false-positive engaged	1	5	6
% false-positive engaged	0.04%	0.40%	0.18%
# not-engaged	1953	1175	3128
% not-engaged	96.45%	94.08%	95.54%

Table 1. Corpus statistics.

orientation, people enter the field of view and approach either frontally, or from the immediate right side.

6 Data and Modeling

The system was deployed during regular business hours for 10 days in each of the two orientations described above, for a total of 158 hours and 32 minutes. No instructions where provided and most people that interacted with the system did so for the first time.

6.1 Corpus and Implicit Labels

Throughout the data collection, the system used a conservative heuristic to detect engagement intentions: it considered that a user wanted to engage when they approached the system and entered in an F-formation (Kendon, 1990b) with it. Specifically, if a sufficiently large (close by) frontal face was detected in front of it, the system triggered an engaging action and started the interaction. We found this F-formation heuristic to be fairly robust, having a false-positive rate of 0.18% (6 false engagements out of 3274 total faces tracked). In 2 of these cases the face tracker committed an error and falsely identified a large nearby face, and in 4 cases a person passed by very close to the system but without any visible intention to engage.

Although details on false-negative statistics have not yet been calculated (this would require a careful examination of all 158 hours of data), our experience with the face detector suggests this number is near 0. In months of usage, we never observed a case where the system failed to detect a close by, frontal face. At the same time, we note that there is an important distinction between people who *actually engage* with the system, and people who *intend to engage*, but perhaps not come in close-enough proximity for the system to detect this intention (according to the heuristic described above). In this sense, while our heuristic can detect people who engage at a 0 false-negative rate, the false-negative rate with respect to engagement intentions is non-zero. Despite these false-negatives, we found that the proposed heuristic still represents a good starting point for learning to detect engagement intentions. As we shall see later, empirical results indicate that, by learning to detect who actually engages, the system can learn to also detect people who might intend to engage, but who ultimately do not engage with the dialog system.

In the experiments described here, we focus on detecting engagement intentions for people that approached while the system was idle. We therefore automatically eliminated all faces that were temporally overlapping with the periods when the system was already engaged in an interaction. For the remaining face traces, we automatically generate labels as follows:

- if a person entered in an F-formation and became engaged in interaction with the system at time t_e, the corresponding face trace was labeled with a positive engagement intention label from t_e-20sec; until t_e; the initial portion of the trace, from the moment it was detected until t_e-20sec was marked with a negative engagement intention label. Finally, the remainder of the trace (from t_e until the face disappeared) was discarded, as the user was actively engaged with the system during this time.
- if the face was never engaged in interaction (*i.e.* a person was just passing by), the entire trace was labeled with a negative engagement intention.

Note that in training the models described below we used these automatic labels, which are not entirely accurate: they include a small number of false-positives, as discussed above. However, for evaluation purposes, we used the corrected labels (no false-positives).

6.2 Models

To review, the task at hand is to learn a model for predicting engagement intentions, based on information that can be extracted at runtime from face traces, including spatiotemporal trajectory and cues about attention. We cast this problem as a frame-by-frame binary classification task: at each frame, the model must classify each visible face as either intending to engage or not. We used a maximum entropy model to make this prediction:

$$P(EI|X) = \frac{1}{Z(X)} exp\left(\sum_i \lambda_i \cdot f_i(X)\right)$$

The key component in the proposed maximum entropy model is the set of features $f_i(X)$, which must capture cues that are relevant for detecting an engagement intention. We designed several subsets of features, summarized in Table 2. The location subset, loc, includes the x and y location of the detected face in the visual scene, and the width and height of the face region, which indirectly capture proximity information. The second feature subset, loc+ff, also includes a (continuous and binarized) score produced by the face detector which reflects the confidence that the face is frontal and thus provides an automatic measure of the focus-of-attention of the agent. Apart from these automatically generated attention features, we also experimented with

Feature sets	Description **[total # of features in set]**
Loc	location features: x, y, width and height **[4]**
loc+ff	location features plus a confidence score indicating whether the face is frontal (ff), as well as a binary version of this score (ff=1) **[6]**
traj(loc)	location features plus trajectory of location features over windows of 5, 10, 20, 30 frames **[118]**
traj(loc+ff)	location and face frontal features, as well as trajectory of location and of face-frontal features over windows of 5, 10, 20, 30 frames **[172]**
traj(loc+attn)	location and manually labeled attention features, as well as trajectory of location and of attention over windows of 5, 10, 20, 30 frames **[133]**

Table 2. Feature sets for detecting engagement intention.

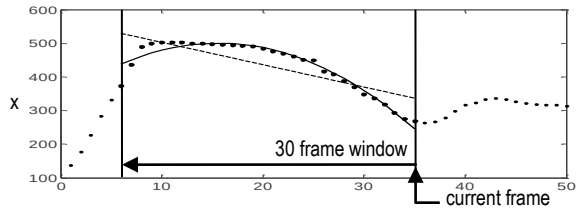

Figure 4. Trajectory features extracted by fitting linear and quadratic functions.

a manually annotated binary attention score, attn. The attention of each detected face was manually tagged throughout the entire dataset. This information is not available at runtime, and we use it only to identify an upper performance baseline.

The maximum entropy model is not temporally structured. The temporal structure of the spatial and attentional trajectory is captured via a set of additional features, derived as follows. Given an existing feature f, we compute a set of trajectory features traj.w(f) by accumulating aggregate statistics for the feature f over a past window of size w frames. We explored windows of size 5, 10, 20, 30. For continuous features, the trajectory statistics include the min, max, mean, and variance of the features in the specified window. In addition, we performed a linear and a quadratic fit of f in this window, and use the resulting coefficients (2 for the linear fit and 3 for the quadratic fit) as features (see the example in Figure 4). For the binary features, the trajectory statistics include the number and proportion of times the feature had a value of 1 in the given window, and the number of frames since the feature last had a value of 1.

7 Experimental Results

We trained and evaluated (using a 10-fold cross-validation process) a set of models for each of the two system orientations shown in Figure 3 and for each of the 5 feature subsets shown in Table 2. The results on the per-frame classification task, including the ROC curves for the different models are presented and discussed in more detail in Appendix A.

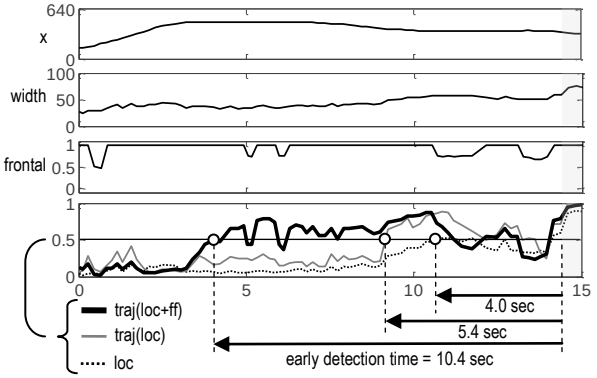

Figure 5. Example predictions for three different models.

At run time, the system uses these frame-based models to predict across time the likelihood that a given agent intends to engage (see Figure 5). In this context, an evaluation that counts the errors per person, rather than errors per frame is more informative. Furthermore, since early detection is important for supporting a natural engagement process, an informative evaluation should also capture how soon a model can detect a positive engagement intention (see Figure 5).

Making decisions about an agent's engagement intentions typically involves comparing the probability of engagement against a preset threshold. Given a threshold, we can compute for each model the number of false-positives at the trace level: if the prediction exceeds the threshold at any point in the trace, we consider that a positive detection. We note that, if we aim to detect people who will actually engage, there are no false negatives at the trace level. The system can use the machine learned models in conjunction with the previous heuristic (a user is detected standing in front of the system), to eventually detect when people engage. Also, given a threshold, we can identify how early a model can correctly detect the intention to engage (compared to the existing F-formation heuristic that defined the moment of engagement in the training data). These durations are illustrated for a threshold of 0.5 in Figure 5, and are referred to in the sequel as *early detection time*. By varying the threshold between 0 and 1, we can obtain a profile that links the false-positive rate at the trace level to how early the system can detect engagement, *i.e.* to the mean early detection time.

Figure 6 shows the false-positive rate as a function of the mean early detection time for models trained using each of the five feature subsets shown in Table 2, in the *side* orientation. The model that uses only location information (including the size of the face and proximity) performs worst. Adding automatically extracted information about whether the face is frontal or leads to only a marginal improvement. However, adding information about the trajectory of location and of attention, leads to larger cumulative gains. Adding the more accurate (ma-

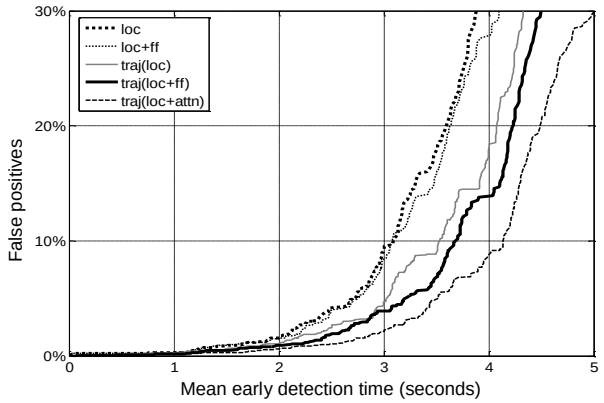

Figure 6. False-positives vs. early detection time (side).

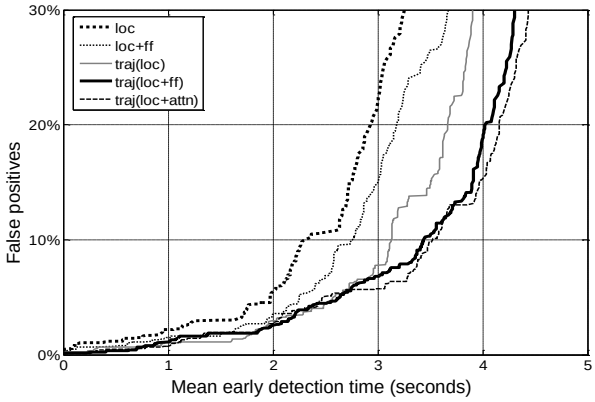

Figure 7. False-positives vs. early detection time (front).

	False positive rate					
Model	EDT=1	EDT=2	EDT=2.5	EDT=3	EDT=3.5	EDT=4
loc	0.31%	1.6%	4.3%	9.4%	18.4%	32.6%
loc+ff	0.31%	1.5%	4.1%	8.7%	18.3%	28.6%
traj(loc)	0.31%	1.1%	2.6%	4.8%	9.3%	18.6%
traj(loc+ff)	0.15%	0.9%	2.0%	4.0%	7.1%	14.3%
traj(loc+attn)	0.26%	0.6%	1.1%	2.2%	5.1%	8.9%

Table 3. *False-positive rate at different EDT (side)

	False positive rate					
Model	EDT=1	EDT=2	EDT=2.5	EDT=3	EDT=3.5	EDT=4
loc	2.3%	5.8%	11.3%	23.0%	35.2%	44.5%
loc+ff	1.6%	3.7%	7.3%	15.8%	28.5%	41.7%
traj(loc)	1.1%	3.1%	4.7%	8.2%	15.6%	36.8%
traj(loc+ff)	1.2%	2.7%	4.7%	7.2%	10.9%	19.8%
traj(loc+attn)	0.8%	2.9%	5.4%	5.4%	10.3%	16.1%

Table 5. *False-positive rate at different EDT (front)

	Early detection time			
Model	FP=2.5%	FP=5%	FP=10%	FP=20%
loc	2.18	2.72	3.09	3.59
loc+ff	2.25	2.74	3.08	3.63
traj(loc)	2.51	3.03	3.53	4.07
traj(loc+ff)	2.68	3.20	3.68	4.22
traj(loc+attn)	3.08	3.52	4.13	4.49

Table 4.*Early detection times at different FP rates (side).

	Early detection time			
Model	FP=2.5%	FP=5%	FP=10%	FP=20%
loc	1.14	1.97	2.29	2.92
loc+ff	1.70	2.25	2.74	3.18
traj(loc)	1.93	2.57	3.13	3.66
traj(loc+ff)	1.99	2.64	3.44	4.02
traj(loc+attn)	1.97	2.47	3.52	4.15

Table 6 * Early detection times at different FP rates (front).

***s**haded cells in Tables 3-6 show statistically significant improvements in performance ($p<0.05$) over the corresponding model that uses the immediately previous feature set (*e.g.* the cell right above). The traj(loc), traj(loc+ff), traj(loc+attn) always statistically significantly ($p<0.05$) improve upon the loc models

nually tagged) information about focus of attention yields the best model. The relative performance of these models (which can be observed at the frame-level in Appendix A) confirms our expectations and the importance of trajectory features (both spatial and attentional) in detecting engagement intentions. The results also indicate that the differences, and hence the importance of these features, are larger when trying to detect engagement early on, *i.e.* at larger early detection times. Tables 3 and 4 further highlight these differences. For instance, when detecting engagement intentions at a mean early detection above 3 seconds, the model that uses trajectory information, traj(loc+ff), decreases the false positive rate by a factor of 3 compared to the location-only model.

Figure 7 and Tables 5 and 6 show the results for the *front* orientation. The relative trends are similar to those observed in the *side* orientation, highlighting again the importance of trajectory features. At the same time, the models are performing slightly worse in absolute terms, which is consistent with the increased difficulty of the task. Several contributing factors can be identified in Figure 3: people may simply pass by in closer proximity to the system; people who come from the corridor are generally frontally oriented towards the system, making frontal face cues less informative; and finally, people who will engage need to deviate less from the regular trajectory of people who are just passing by.

Next, we review how well the models trained generalize across the two different setups, by evaluating the trajectory models traj(loc+ff) across the two datasets. The results indicate that the models are attuned to the dataset they are trained on (see Figure 7). As we discussed earlier, we expect this result given the different geometry of the relative trajectories of engagement in the two orientations. These results highlight the importance of learning in situ, and show that the proposed approach can be used to learn the specific patterns of engagement in a given environment automatically, without explicit developer supervision.

Finally, we performed an error analysis. We focused on the *side* orientation and visually inspected the 79 (4%) false-positive errors committed by the traj(loc+ff) model when using a threshold corresponding to a mean

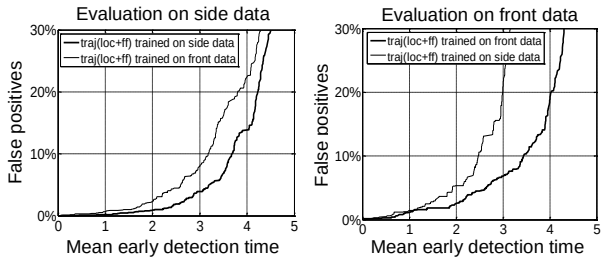

Figure 7. Model evaluation across orientations.

early detection time of 3 seconds. This analysis indicates that in 22 out 79 of these errors (28%) the person did actually exhibit behaviors consistent with an intention to engage the system, such as stopping by or turning around after passing the system, and approaching and maintaining sustained attention for a significant amount of time. These cases represent false-negatives committed by our conservative F-formation heuristic with respect to engagement intention; the user did not approach close enough for the system to trigger engagement. The actual false-positive rate of the trained model is therefore 2.9% rather than 4%. The system was able to correctly identify these cases because the behavioral patterns are similar to the ones exhibited by people who did approach close enough for the heuristic detector to fire. We plan to assess the false-negative rate of the current heuristic more closely and explore how many false negatives are actually recovered by the trained model. This analysis will require that multiple judges assess engagement intentions on all 3274 traces.

8 Summary and Future Work

We described an approach to learning engagement intentions in a situated conversational system. The proposed models fit into a larger framework for supporting multiparty, situated engagement and open-world dialog (Bohus and Horvitz, 2009a; 2009b). Experimental results indicate that a system using the proposed approach can learn to detect engagement intentions at low false positive rates up to 3-4 seconds prior to the actual moment of engagement. The models leverage features that capture spatiotemporal and attentional cues that are tuned to the specifics of the physical environment in which the system operates. Furthermore, the models can be trained in previously unseen environments, without any explicit developer supervision.

We believe the methods and results described represent a first step towards supporting fluid, natural engagement in open-world interaction. Numerous challenges remain. While we confirmed the importance of spatiotemporal and attentional features in detecting engagement intentions, we believe that leveraging additional and more accurate sensory information (*e.g.* body pose, eye gaze, more accurate depth information, agent identity coupled with longer term memory features) may improve performance. Secondly, while the current models where trained in a batch fashion, the proposed method naturally lends itself to an online approach, where the system starts with a prior model for detecting engagement intentions, and refines this model online. More importantly, rather than just learning to detect engagement intentions, we plan to focus on the more general problem of controlling the engagement process: how should the system time its actions (*i.e.* gaze and sustained attention, smiles, greeting, etc.) to create natural, fluid engagements in the open world. Also, introducing mobility to dialog systems brings another interesting dimension to this problem: how can a mobile system, such as a robot, detect engagement intentions and respond to support a natural engagement process? We believe that there is great opportunity to address these interaction challenges by learning predictive models from data.

References

M. Argyle and M. Cook, 1976, *Gaze and Mutual Gaze*, Cambridge University Press, New York

D. Bohus and E. Horvitz, 2009a, *Open-World Dialog: Challenges, Directions and Prototype*, to appear in KRPD'09, Pasadena, CA

D. Bohus and E. Horvitz, 2009b, *Computational Models for Multiparty Engagement in Open-World Dialog*, submitted to SIGdial'09, London, UK.

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

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

A. Kendon, 1990a, *A description of some human greetings*, Conducting Interaction: Patterns of behavior in focused encounters, Studies in International Sociolinguistics, Cambridge University Press

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

M.P. Michalowski, S. Sabanovic, and R. Simmons, *A spatial model of engagement for a social robot*, in 9th IEEE Workshop on Advanced Motion Control, pp. 762-767

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

C. Peters, 2005b, *Direction of Attention Perception for Conversation Initiation in Virtual Environments*, in Intelligent Virtual Agents, 2005, pp. 215-228.

C.L. Sidner, C.D. Kidd, C. Lee, and N. Lesh, 2004, *Where to Look: A Study of Human-Robot Engagement*, IUI'2004, pp. 78-84, Madeira, Portugal

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

R. Vertegaal, R. Slagter, G.C.v.d.Veer, and A. Nijholt, 2001, *Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes*, CHI'01

Appendix A. Per-frame evaluation of maximum entropy models for detecting engagement intentions

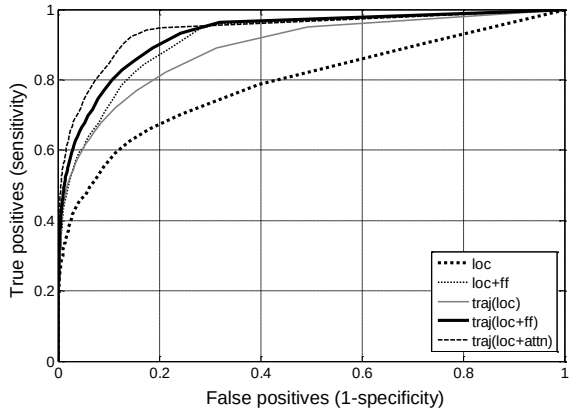

Figure 1. Per-frame ROC for side orientation models

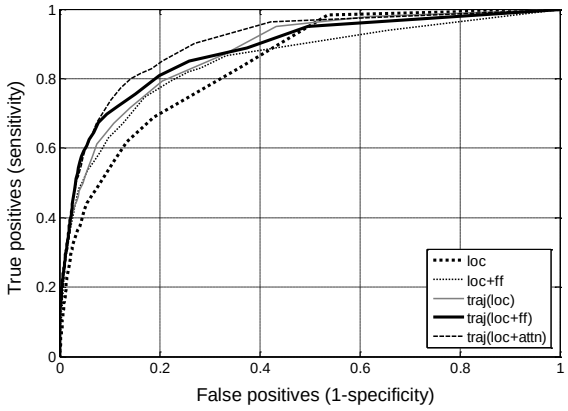

Figure 2. Per-frame ROC for front orientation models

Model	Avg. log-likelihood			Hard error		
	Base	Train	CV	Base	Train	CV
loc	-0.1651	-0.1222	-0.1259	3.91%	3.22%	3.25%
loc+ff	-0.1651	-0.0962	-0.0984	3.91%	3.01%	3.07%
traj(loc)	-0.1651	-0.0947	-0.1073	3.91%	2.88%	3.06%
traj(loc+ff)	-0.1651	-0.0836	-0.0904	3.91%	2.69%	2.85%
traj(loc+attn)	-0.1651	-0.0765	-0.0810	3.91%	2.47%	2.56%

Table 1. Baseline, training-set and cross-validation performance (data average log-likelihood and classification error) for side orientation models

Model	Avg. log-likelihood			Hard error		
	Base	Train	CV	Base	Train	CV
loc	-0.1875	-0.1451	-0.1498	4.63%	4.58%	4.72%
loc+ff	-0.1875	-0.1326	-0.1392	4.63%	4.22%	4.39%
traj(loc)	-0.1875	-0.1262	-0.1338	4.63%	3.99%	4.24%
traj(loc+ff)	-0.1875	-0.1159	-0.1298	4.63%	3.91%	4.38%
traj(loc+attn)	-0.1875	-0.1150	-0.1267	4.63%	4.04%	4.47%

Table 2. Baseline, training-set and cross-validation performance (data average log-likelihood and classification error) for front orientation models