Managing Human-Robot Engagement with Forecasts and… *um…* Hesitations

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+1-425-706-5880
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+1-425-706-2127
horvitz@microsoft.com

ABSTRACT

We explore methods for managing conversational engagement in open-world, physically situated dialog systems. We investigate a self-supervised methodology for constructing forecasting models that aim to anticipate when participants are about to terminate their interactions with a situated system. We study how these models can be leveraged to guide a disengagement policy that uses linguistic hesitation actions, such as filled and non-filled pauses, when uncertainty about the continuation of engagement arises. The hesitations allow for additional time for sensing and inference, and convey the system's uncertainty. We report results from a study of the proposed approach with a directions-giving robot deployed 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**]: Multimedia Information Systems – *Audio input/outputs;* User Interfaces – *Natural Language*; I.4.8 **[Scene Analysis]**: Tracking, Sensor Fusion

General Terms

Algorithms, Human Factors

Keywords

Engagement; hesitation actions; filled pauses; forecasting models; human-robot interaction; multimodal interaction; self-supervision.

1. INTRODUCTION

Situated, multimodal interactive systems typically grapple with multiple uncertainties as they perceive, infer, and act. In addition to challenges with speech recognition, such systems face uncertainties in the visual space, and in the multimodal inferences they make such as tracking the location of users, their focus of attention, gestures, engagement, floor control actions, and intentions. Beyond real-time inferences about the state of world, the systems may leverage the ability to forecast future states, and use such forecasts for planning and decision making. For instance, in making a decision about taking a turn in a multiparty conversation, a robot or virtual agent may need to consider who is likely to start talking next and when? Similarly, the ability to anticipate that people are about to enter or leave a conversation, or that their focus of attention may shift, can help support the planning of verbal and gestural outputs in a manner that leads to more fluid interactions.

We explore the construction and use of forecasting models that enable interactive systems to make inferences about the near-term future. In particular, we investigate a self-supervised approach to constructing these models that does not require manual annotations. Furthermore, we investigate the use of linguistic hesitation actions that can signal the system's state of confusion and generate additional time for collecting evidence and resolving uncertainties. We explore the use of forecasts and hesitations both separately and together, showing how forecasts can guide hesitations.

While the proposed forecasting and hesitation mechanisms are applicable to a variety of interaction problems, we explore their use in the context of managing conversational disengagement in a physically situated interactive system. Starting with a conservative, heuristic model for disengagement, we present methods for learning a forecasting model that is guided by this heuristic and that aims to anticipate whether participants will terminate their interactions with the system. We show how this self-supervised model can guide disengagement policies that use hesitations when situations of high uncertainty about the future arise. Specifically, if the model indicates that participants might disengage, the system may insert a filled pause, *"So ..."*, and, as it collects more evidence, the next contribution can be generated accordingly, for instance: *"Anything else I can do for you?"* if the participants remain engaged, or *"Well, I'll catch you later then."* if they disengage.

We begin with a review of related work. Next, in Section 3, we motivate and articulate the open-world disengagement challenge. We outline the proposed approach in Section 4, present empirical results in Section 5, and discuss limitations, lessons learned and future opportunities in Section 6. Section 7 summarizes the paper.

2. RELATED WORK

The ability to establish and maintain an open communication channel is a foundational competency for spoken language

ICMI '14, November 12 - 16 2014, Istanbul, Turkey.
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-2885-2/14/11…$15.00
http://dx.doi.org/10.1145/2663204.2663241

interaction. In physically situated settings, participants committed to an interaction typically enter into and maintain a certain spatial orientation, or *f-formation* [1], and actively signal and manage their conversational engagement by a variety of verbal and non-verbal cues, including gaze, gesture and body language.

Several strands of research have investigated various aspects of the engagement process in situated interactive systems. Sidner and colleagues [2] showed that people direct their attention to a robot more often if the robot performed engagement behaviors. A model of engagement for human-robot interaction based on connection events is proposed and implemented in [3]. In [4] a model of engagement based on proxemics was proposed and evaluated with a receptionist robot. In our own previous work, we have proposed a computational model for managing engagement with a situated agent in multiparty, open-world settings [5]. Furthermore, in [6], we have showed how, by starting with a conservative heuristic for detecting when people are initiating their engagement with a conversational agent, we can learn to predict this event from visual features. In this work, we focus instead on the problem of *disengagement*, and show how a similar, self-supervised forecasting approach can be used to anticipate when participants are about to finish their interactions. In addition, we conduct an online evaluation of the learned model, and investigate how it can guide disengagement policies that leverage hesitation actions to mitigate situations of high uncertainty.

Hesitations and disfluencies abound in conversational speech, and have previously been the object of linguistic investigation. For instance, [7] discusses the use of filled pauses like *uh* and *um* and, based on a corpus analysis, conclude that they are used to signal short (*uh*) or long (*um*) production pauses. In [8], a survey of previous analyses of fillers and hesitations is presented, indicating that fillers occur in situations where speakers are uncertain about choices they need to make, and that they facilitate understanding and allow listeners to assess the speaker's level of confidence.

In human-computer interaction, methods have been proposed for detecting filled pauses during speech recognition [9, 10], and for modeling and producing them during synthesis [11]. In efforts more closely related to this work, filled pauses have also been used in incremental natural language generation systems. For instance, results from a Wizard of Oz study [12] indicate that an incremental natural language generation system leveraging filled pauses, as well as overt and covert self-corrections, can achieve shorter response times and is perceived as more efficient than a non-incremental generator, even though it produces longer utterances. [13] presents an analysis of types of user reactions occurring while pausing by using filled pauses, gaze, and syntactic completeness. Reinforcement learning methods for guiding incremental generation of natural language were shown to avoid long waiting times and minimize the use of fillers and self-corrections [14].

We use filled pauses as hesitation devices to endow a system with the ability to generate additional time for perception and decision making, as well as to display the uncertainty the system has. In distinction to prior efforts, we operate in the more complex setting of a multimodal, physically situated interactive system. Our focus is on uncertainties and decisions related to managing engagement. The generation of filled pauses is driven by a learned model that uses multimodal information to continuously forecast whether participants are about to terminate their interactions with the system. Finally, we conduct and report results from an end-to-end, in-the-wild study of the proposed approach with a deployed human-robot interactive system.

3. PROBLEM

We focus on the problem of managing conversational engagement, defined in [2] as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". As an experimental platform, we use Directions Robot [15], a system that couples a platform of physically situated interaction [16] with a Nao humanoid robot [17]. The system uses language synchronized with gestures to provide directions to people's offices, conference rooms, and other public areas inside our building; a video is available at http://sdrv.ms/15Yay8V. The robot is deployed in an open space, in front of the elevators on the 3rd floor, in a standing setup shown in Figure 1. The usual traffic in this area includes people with offices on the floor, as well as visitors, who come to see people or attend meetings in the building, and who are often unfamiliar with the surroundings.

Managing engagement in physically situated settings is a challenging multimodal problem. Engagement is a mixed-initiative, jointly coordinated process, involving multiple signals across different channels, including proxemics and body orientation, eye gaze, head and hand gestures, verbal and non-verbal behaviors. The open-world deployment of the Directions Robot brings additional challenges: interactions often involve groups of people, and are sometimes driven by actual needs and often by curiosity about the robot. People come and go at will, and their focus of attention is often shared between the robot and other people in a group, bystanders, or people passing by.

The Directions Robot manages engagement via a computational model previously described in [5]. For every person in the scene, the robot reasons about the engagement *state*, *actions,* and *intentions*, and makes engagement decisions based on these inferences. The *engagement state* captures whether or not a person is engaged in an interaction with the system. The *engagement intention* reflects whether or not the person wants to be engaged with the system. Finally, *engagement actions* model whether a person is performing a maintaining or disengaging action (if engaged), or an initiating action versus no action (if not engaged).

In a previous set of experiments [15], we explored the use of heuristic and machine-learned models for inferring engagement actions performed by people in the scene. In addition to bringing to the fore some of the challenges with managing engagement that we have already outlined, the study also highlighted a tradeoff between making early, incorrect disengagement decisions, versus being more conservative, and making these decisions too late. We found that, when using a model that we had learned from data, the robot would sometimes disengage abruptly if a participant would turn their attention towards another person in their group, or towards the direction that the robot was pointing to while it provided directions. Such head turns and motions resemble the beginning of disengagements. Committing false-positive errors on detecting disengagement is very costly as the robot disengages and bids the user a goodbye inappropriately, while the user expects the conversation to continue. A more conservative approach for inferring disengagement (such as a heuristic rule used in the study) may largely avoid such false positives, but can in turn lead to late disengagement decisions. As a result, the robot sometimes tries to continue the dialog while the participants are leaving.

4. APPROACH

Motivated in part by lessons learned from this initial user study, we investigate methods for managing disengagement decisions that address this tradeoff. We present a methodology where the system

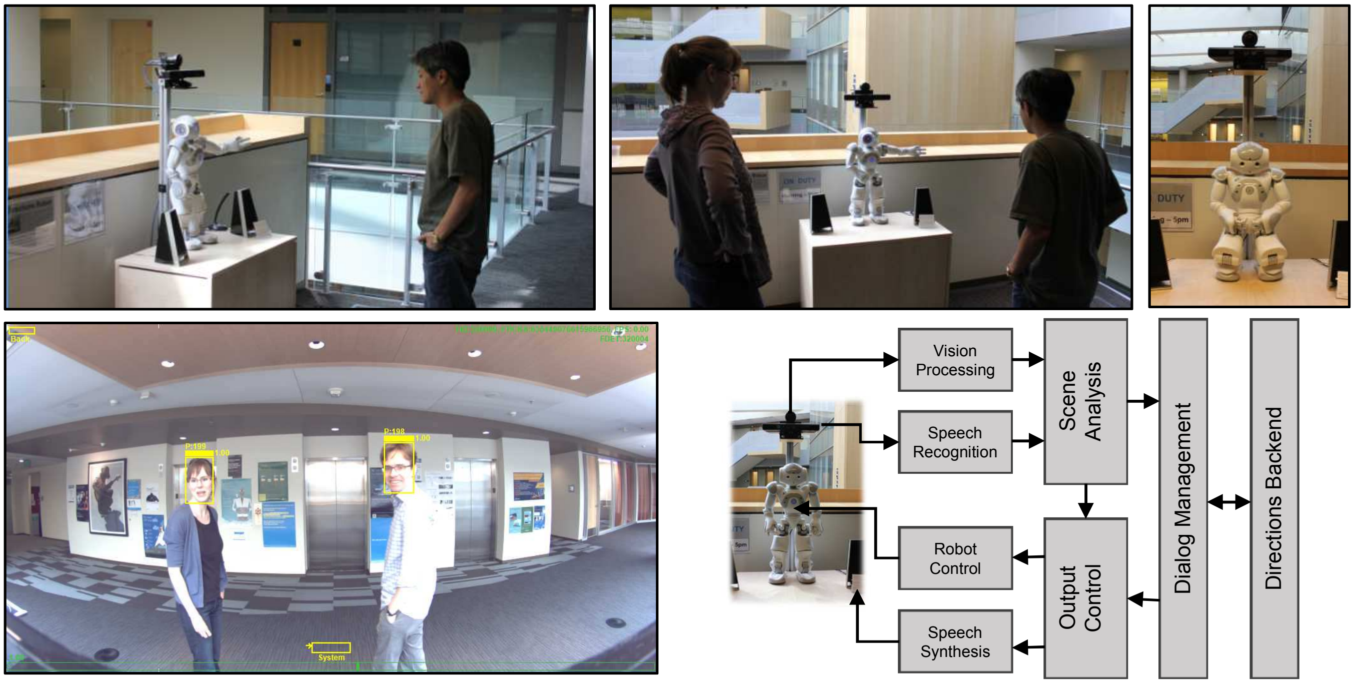

Figure 1. Directions Robot, from top-left to bottom-right: robot interacting with one participant; robot interacting with two participants; robot in offline position; sample view from system camera; system diagram.

automatically learns, in a self-supervised manner, to *forecast* whether participants will shortly disengage. This prediction is accomplished by starting with a conservative heuristic for detecting disengagement, and learning to predict when this heuristic will signal disengagement *ahead of time*, from available real-time multimodal features. In addition to this forecasting model, we explore the use of *hesitation actions* such as filled and non-filled pauses, in situations of high uncertainty about the future. Producing filled pauses can reflect the system's state, allow for more time for gathering evidence, and mitigate disengagement decisions that may have a high cost. We describe the details of the approach below.

4.1 Forecasting Disengagement

The forecasting model for disengagement aims to assess the likelihood that a participant will disengage with the system within some small time interval or *lookahead* (*e.g.*, 5 seconds) into the future. We expect that the ability to forecast disengagement can inform better decision-making, and help avoid late contributions.

The self-supervised methodology for training this predictive model relies on initially running the system with a heuristic rule for assessing disengagement that is conservative, or high-precision: the rule aims to commit no false positives, *i.e.*, it only indicates disengagement when it has indeed happened, at the cost of potentially being late. We construct this heuristic by leveraging features that capture how close the participant is, whether a participant is stationary or moving, and whether or not a participant is attending to the robot.

As an example, Figure 2 shows a trace of key variables for a participant, over time. The top three plots show the location of the face in the x dimension (A), the width of the face (B), and the probability that the participant's attention is on the system computed via a probabilistic model (C). These raw streams are used to compute measures of proximity, stability, and attention, which are in turn used via a heuristic rule (see more details in the next section) to estimate the probability of disengagement – shown in Figure 2.D. Based on interaction data collected by the system, we can automatically define a target label for the forecasting model at every frame: the label (also shown in Figure 2.D) is 1 if the conservative heuristic signals disengagement with probability > 0.8 within the lookahead window (in this case 5 seconds), and 0 otherwise. The disengagement forecasting model is trained to predict this label at every frame, from existing features. The probability of future disengagement computed by running a trained forecasting model is also shown in Figure 2.D.

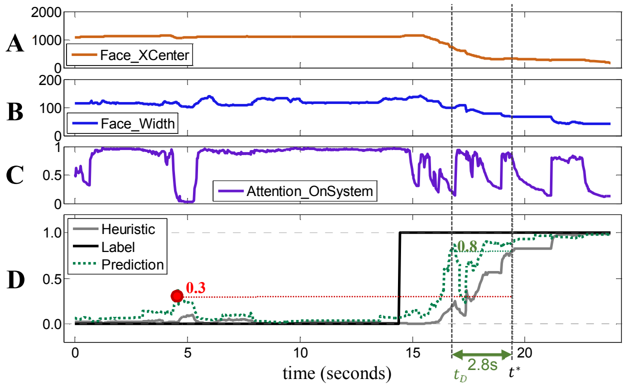

Figure 2. User trace. A: location of face, B: width of face, C: inferences about attention of user on system, and D: heuristic inference, labels, and forecasting model

4.2 Disengagement Policy with Hesitations

The second focus of this work is the study of a disengagement policy that relies on hesitation actions such as filled and non-filled pauses to mitigate high-cost disengagement decisions. The central idea is that, in situations of high uncertainty about the future engagement state of the participants, instead of issuing the next dialog contribution, the system can instead produce a hesitation, *e.g.,* slowly saying *"So ..."* as a non-committal linguistic action that leaves the options open for either continuing or finishing the engagement.

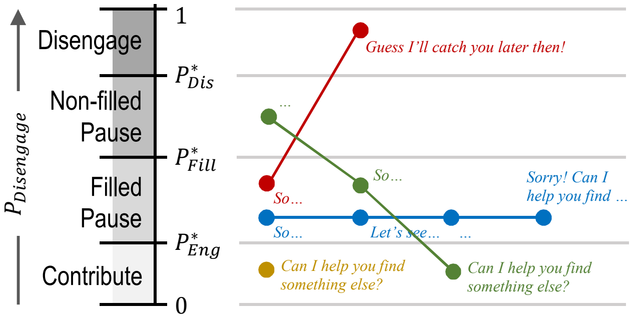

Figure 3. Disengagement policy with hesitation actions.

The disengagement policy chooses among four possible actions by comparing the probability of disengagement against three predefined thresholds, as shown in Figure 3. As with the standard policy, the probability of disengagement is compared at every frame against a disengagement threshold, and, if it exceeds the threshold ($P_D > P^*_{Dis}$), the system disengages. In addition, every time the system is about to take a turn, a decision is made about whether a filled or non-filled pause should be triggered instead. If the probability of disengagement is low, *i.e.* smaller than a corresponding engagement threshold ($P_D < P^*_{Eng}$), the system takes the turn and generates its next contribution. If the probability of disengagement falls the middle range ($P^*_{Eng} < P_D < P^*_{Dis}$), the new policy triggers instead a hesitation action, in the form of a filled or a non-filled pause. Filled-pauses, *i.e., "So…," or "Let's see…,"* accompanied by an extended arm gesture toward the participant are produced when the probability of disengagement is smaller than a preset filled-pause threshold ($P^*_{Eng} < P_D < P^*_{Fill}$); non-filled pauses are produced otherwise ($P^*_{Fill} < P_D < P^*_{Dis}$).

The filled and non-filled pauses are hesitation actions that allow the system to buy more time, at a relatively low cost. As the hesitation action is being produced, the system can gather additional evidence over time that will hopefully lead to resolving the uncertainty: the system will eventually figure out whether or not the participants are disengaging and can perform the correct action. The expected cost of these hesitations is lower compared to a potentially incorrect engaging or disengaging action. In addition, the production of a filled pause also conveys to the user the system's difficulty with making a choice [8] at this time.

If the uncertainty persists, the policy re-runs, and may produce a succession of hesitations. If the probability of disengagement remains in the filled-pause region and the system should take a turn, the system produces up to two successive filled pauses, *i.e., "So…," "Let's see…,"* followed by a non-filled pause, and eventually a contribution if the uncertainty still persists (blue trace in Figure 3) If the probability of disengagement drops back into the engaging region after the system produced a hesitation, and more than 2 seconds have elapsed since, the system inserts an overt self-repair—*"Sorry!"* before issuing its next contribution, to share a reflection with the user about the hesitation, and prior uncertainty and its resolution. We believe the apology can help further mitigate the costs of the filled pause.

5. EXPERIMENTS AND RESULTS

We now describe the experiments conducted with the proposed approach. The test-bed for these experiments is the *Directions Robot,* a robotic system deployed in the wild that can interact with one or multiple participants via natural language, and provide directions to people and offices in our building. The system combines a Nao humanoid robot with off-board sensors and computation. A high-resolution wide-angle camera and a Kinect sensor are placed above the robot, and a multi-core desktop computer runs the software infrastructure and controls the robot. The software subsumes components for making inferences from audio-visual signals (*e.g.*, face tracking, speech recognition, *etc.*) and combines them with interaction planning, decision making, and output generation. In addition, the robot leverages directory information and a building map framework to construct spoken directions. A video is available at http://sdrv.ms/15Yay8V.

5.1 Baseline Inference and Policy

The baseline system was equipped with a conservative inference model for assessing disengagement. This model was manually crafted, taken into account lessons learned from a previous user study [15]. It constructed an estimate for the probability that a participant is disengaging by combining three continuous scores (in the 0-1 interval), capturing three signals of engagement: *proximity* (P), *stability* (S), and *attention persistence* (A), according to the equation below:

$$P_D = 1 - P \cdot (1 - (1 - S) \cdot (1 - A))$$

This heuristic requires that proximity is high and at least one of stability or attention is high in order for the user to be considered engaged. If the proximity score is low, or alternatively if both stability and attention are low, the probability of disengagement moves towards one. The scores for proximity, stability and attention persistence were computed by applying a sigmoid transform to a base feature. For proximity, the base feature was the size of the tracked face. For stability, the base feature measured the ratio between the max horizontal excursion of the face throughout the last one second and the size of the face. Finally, for attention persistence the base feature was computed by averaging over the past two seconds the probability that the user is attending to the system. The latter probability is computed by a machine learned model that uses face tracking and head pose information. The parameters of each sigmoid normalization function were manually tuned to ensure a conservative inference model, which aims to minimize the false-positive rate.

The default policy for making engagement control decisions relies on the computed probability of disengagement: when this exceeds a preset threshold, the robot disengages with the participant. The interaction is terminated when the last participant is being disengaged. The final disengagement action performed by the robot depends on the current dialog state. If disengagement occurs near the beginning of the dialog, the robot terminates the interaction without any speech or gesture: this covers mostly cases when the engagement was incorrectly initiated with a person going by. If the disengagement occurs in the middle of the dialog, the robot says "*Well, guess I'll catch you later then!*" communicating its surprise at the early disengagement. Finally, if the engagement terminates after directions were given, a simple salutation such as *"Bye bye!"* is performed, accompanied by a hand-waving gesture. In addition, when interrupting its own speech to terminate an interaction, the robot produces a synthetic disfluency (if the break did not occur on a word boundary) to further indicate its surprise, followed by a final, *"Oops! Guess I'll catch you later then."*

Given challenges with tracking multiple participants in open-world settings, the vision system may sometimes lose track of an engaged participant. In this case, to avoid an immediate termination of engagement, the participant is persisted in an *invisible* mode for a short period, giving the tracker a chance to recover. While the participant is in this invisible mode, the heuristic rule increases the probability of disengagement from the last known value towards one, ensuring that the disengagement threshold is reached in five seconds since the face was lost. If the track is still not recovered at that point, the system disengages with the invisible participant. The baseline policy is adjusted to avoid producing any verbal contributions if all engaged participants are in this invisible state. Since lost faces often occur when people turn and leave abruptly, this policy reduces the chances of the robot issuing a late verbal contribution, while people are trying to disengage.

5.2 Forecasting Disengagement

5.2.1 Data

The data for constructing the forecasting model for disengagement was collected by running the baseline system described above for a period of 5 days. Throughout this time, the robot initiated 133 interactions. The resulting dataset for training the forecasting model contained 158 user traces (due to the multiparty interaction setting, there are more user traces than interactions), and ~126K frames or data-points.

5.2.2 Labels and Features

For each frame in a user trace, labels indicating whether disengagement would occur (according to the baseline heuristic) within a future time window were automatically constructed, as described in Section 4.1. We explored the use of models with different lookaheads, ranging from 3 to 6 seconds. We also investigated the use of different classes of features, spanning multiple knowledge sources and modalities:

- *FaceLocation* (51 features): the horizontal location and the size of the tracked face, and derived feature streams capturing statistics of these signals such as average, slope, standard deviation, excursion over past time windows ranging from 250ms to 8 seconds. In addition, we added the base stability feature that was used in the baseline heuristic.
- *TrackingConfidence* (18 features): the tracking confidence score, and derived feature streams capturing the average and slope of this signal over past time windows; in addition we used information about whether the face was currently invisible, and for long this situation persisted.
- *Focus-of-Attention* (19 features): the probability that the user is attending to the system produced by the attention inference model, together with derived streams capturing un-weighted and time-based exponentially weighted averages, as well as the slope of this signal across past time windows.
- *Interaction/Dialog* (116 features): features that encode the current dialog state as well as derived features that describe and how long the system has been in that state; the dialog state generally corresponds to the system's semantic output, *e.g., DoYouNeedHelp*, *LocationConfirm*, *WhatAreYouLooking-For*. We also include here turn-taking information such as whether the system or a user is currently speaking, and related timing information. Finally, we include the total number of people detected in the scene and the number of users engaged in the interaction.
- *BaselineHeuristic* (1 feature): the baseline heuristic described above was also considered as a feature.

5.2.3 Model Training and Selection

We trained logistic regression and boosted decision tree models, with various parameterizations: for logistic regression we explored the use of different L1 and L2 regularization weights, while for boosted trees we changed the number of leaves (2, 4, 8), the number of trees constructed (500, 1000, 2000), and the learning rate. We identified the best logistic regression and boosted tree model across these different parameterizations for each specified lookahead.

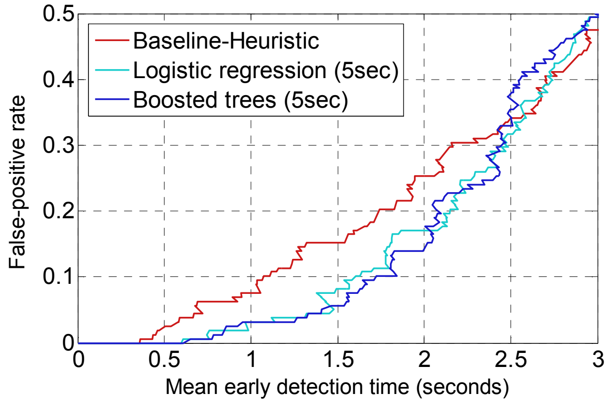

Figure 4. False-positive rate versus mean early-detection-time at different thresholds.

Each of these models implements a particular tradeoff between the false-positive rate and how early the future disengagement may be detected. To illustrate this, consider again the trace representing the predictions from a given model shown in Figure 2.D. By applying a disengagement threshold, such as 0.8, a correct prediction of disengagement is made 2.8 seconds ahead of the moment that the heuristic signals disengagement. We refer to this time as the *early-detection-time*. If we use a lower disengagement threshold, the prediction might be made even earlier. At the same time, lowering the threshold too much, for instance to 0.3, can lead to false-positives, as marked by the red dot in Figure 2.D.

For each of the pre-selected models, we conducted an analysis where we varied the disengagement threshold and plotted the mean early-detection-time versus the false-positive rate. In this analysis, false positives were defined as any detection earlier than 6 seconds before the heuristic detection; if a user engaged and disengaged multiple times, we considered only the first disengagement event on the user trace. Figure 4 shows the plots for the heuristic baseline, and the best performing logistic regression and boosted trees models. Based on this analysis, we selected the 5-second lookahead logistic regression model. In an effort to keep the false-positives to a minimum, we chose a high disengagement threshold, based on a methodology that aimed to minimize false-positives: we incrementally lowered the threshold from 1.0, until a first false-positive error at a large early-detection-time was found.

Finally, for the chosen model setup, we also investigated how performance varies when using different subsets of features. The frame-based classification error and mean squared error attained by using different subsets of the features are shown in Table 1.

Table 1. Performance with different feature sets

Model	**Classification error**	**Mean squared error**
Majority Baseline	**24.9%**	**0.1880**
Heuristic based model (H)	13.9%	0.1134
Focus-of-Attention (A)	17.0%	0.1249
FaceLocation (L)	16.3%	0.1265
TrackingConfidence (C)	18.8%	0.1468
Interaction/Dialog (D)	21.3%	0.1526
A+L	13.5%	0.1057
A+L+C	11.7%	0.0955
A+L+C+D	11.5%	0.0924
Full model (A+L+C+D+H)	**10.9%**	**0.0863**

Training a model using the manually engineered baseline heuristic as a feature produces a model (H) that outperforms the models trained on any of the other feature classes in isolation: focus-of-attention (A), face-location (L), tracking-confidence (C) and interaction/dialog (D). Of these, the focus-of-attention and location features seem to be most informative. Adding them together leads to a model (A+L) that already exceeds the model based on the heuristic (H). Adding the other two feature classes (tracking-confidence and interaction/dialog) further improves performance—see A+L+C+D model in Table 1. Finally, the results indicate that the manually constructed heuristic contains information orthogonal to the models constructed with the other feature classes, as adding it as a feature yields further performance gains for the Full model.

5.3 In-the-Wild Study

The performance assessment presented above indicates that the forecasting model can indeed anticipate disengagement, as signaled by the conservative heuristic, while maintaining a low false-positive rate. In order to understand the effect of this model and hesitation-based policy on the end-to-end system, we deployed the selected model in the live system, and ran an observational user study. To tease apart the effects of the forecasting model and of the hesitation policy, we implemented and contrasted four different conditions for managing disengagement:

- *Baseline (B)*: the system uses the baseline, conservative heuristic for inferring disengagement and the baseline policy.
- *Baseline+Hesitations (B+H)*: the system uses the baseline heuristic for inferring disengagement in conjunction with the hesitation-action policy.
- *Forecast (F)*: the system uses the forecasting model for disengagement, and the baseline policy.
- *Forecast+Hesitations (F+H)*: the system uses the forecasting model for disengagement, and the hesitation-action policy.

The study ran for 19 days, and each time an interaction was initiated, the robot randomly chose one of the four conditions above to use for disengagement. Throughout this time, the robot initiated 589 interactions. 80 of them were eliminated from the analysis: 19 due to various crashing bugs, and 61 due to false initiations of engagement. The remaining corpus contains 509 interactions: 133 (26.1%) in the *Baseline* condition, 123 (24.2%) in the *Baseline+Hesitations* condition, 125 (24.6%) in the *Forecast* condition, and 128 (25.1%) in the *Forecast+Hesitations* condition.

During the user study, we found a subtle bug that postponed the termination of the interaction by a single frame for the F and F+H conditions, in situations where the probability computed by the forecasting model exceeded the disengagement threshold. We believe this very small delay does not affect the results.

We focused the assessment on the system's behavior during the final moments of the interaction, when the last engaged participant was disengaging. The goal was to identify situations where the system continued the dialog inappropriately while participants were leaving, but also situations where the system terminated the interaction too early, prior to the moment participants wished to disengage. An inherent tradeoff exists between these categories: when using a conservative disengagement model and policy, the robot can avoid incorrect, early disengagements, but this comes at the expense of producing more late disengagements, *i.e.*, the system will tend to continue the interaction *after* participants have already terminated their engagement, as they are on their way out. Also, the system often interrupts itself on these contributions, as the disengagement model finally indicates that users are leaving. We expect that, with use of the forecasting model and hesitations, the number of late disengagements would be reduced, without a significant increase in the number of early disengagements.

To identify late disengagements, we developed a tagging scheme that focused on the last system contribution that was not a departure salutation, and identified whether this contribution was produced late with respect to the engagement state of the participants. A contribution was denoted late if it was started after the last engaged participant broke the f-formation and terminated their engagement – *LateStart*, or if it was started correctly, while at least one participant was engaged, but continued after the last participant had left the engagement – *LateContinue*. Otherwise, if at least one participant maintained engagement throughout the system's contribution, the contribution was denoted *OnTime*.

The annotations were performed by a professional tagger, who had access to the video and audio of the interactions from the system's viewpoint, as shown in Figure 1. If the last contribution was a *LateStart*, the annotator also inspected the previous contributions, in order to identify the earliest one that might have been started or continued late. In addition, the annotator also identified early disengagements, *i.e.*, situations when the system stops the conversation early, before the participants actually disengaged; these were assessed at the moment the system started its departure salutation or finished the interaction non-verbally.

To gain a better understanding of the effect of the hesitation actions, we also instructed the annotator to inspect each hesitation action produced (including the non-filled pauses), and mark whether it was acceptable or costly in terms of perceived awkwardness and overall influence on the interaction, when viewed in context. We asked the annotator to write down their observations about how costly hesitation actions influenced the interactions. This analysis aimed to shed more light on classes of problems that occur with hesitation actions, and to identify other aspects that should be taken into account when constructing hesitation policies.

With respect to the perceived naturalness and cost to the interaction, a complex interplay exists between the verbal outputs and hesitations produced by the robot, the timing of their production with respect to the engagement status of the participants, and the broader dialog and situational context. From a disengagement perspective, we considered interactions that contain a late contribution that tries to advance the dialog while participants are disengaging to be unnatural or costly. There are two types of dialog acts however that we believe are less costly if produced even while participants are disengaging. One of them is the *NoProblem* dialog act. Oftentimes, after receiving directions, people thank the robot, and leave. If the thank you is confidently understood, the robot responds with a *no-problem* dialog act, rendered by saying *"You're welcome!"*, or *"No problem!"* This contribution is acceptable even when participants are on their way out as it does not attempt to move the dialog forward and can be seen as a closure. Such closures at a distance occur naturally among people after a disengagement. We denote these contributions *Late(NoProblem)*.

At the onset, we also started with the assumption that the filled pause dialog act – *Late(Hesitation)*, is less costly than a regular dialog act when produced during moments of disengagement. For instance, *"So…,"* followed often by a departure salutation like *"Well, guess I'll catch you later!"* generally has a lower cost than issuing a contribution that tries to advance the dialog. Subsequent data inspection and the annotator analysis revealed however that filled pauses produced late were sometimes judged as costly / awkward; often this had to do with the filled pause causing the disengaging participants to turn their attention back to the system, and sometimes even stop in their tracks. We therefore further

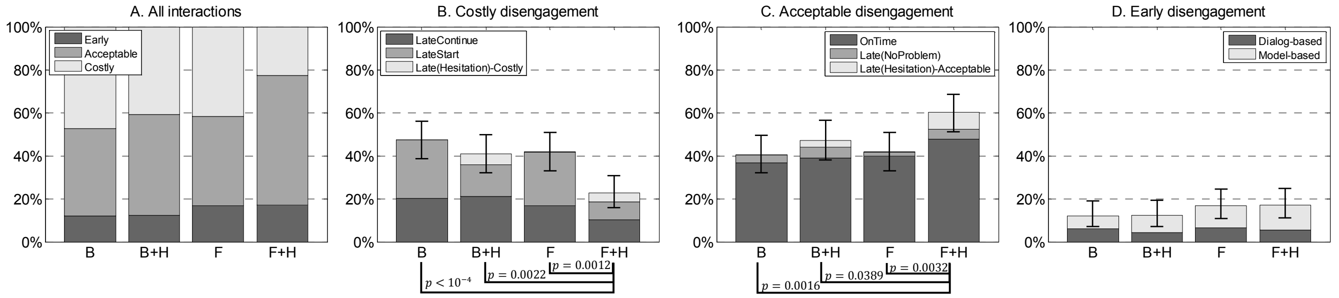

Figure 5. Breakdown of disengagement types.

decomposed these *Late(Hesitation)* acts into *Late(Hesitation)-Costly*, and *Late(Hesitation)-Acceptable*.

Overall, we believe there are various cost differences between interactions that end with a *LateContinue*, *LateStart*, *Late(No-Problem), Late(Hesitation)-Acceptable, Late(Hesitation)-Costly,* or with an *OnTime* output. As such, for completeness, we present the corresponding statistics for each of these classes across conditions in Table 2. As a simplifying first order approximation, we considered as *Costly* disengagements the interactions with disengagements ending with a *LateContinue*, *LateStart* and *Late(Hesitation)-Costly* contributions. We also grouped together as *Acceptable* disengagements the ones with *OnTime, Late(No-Problem)* and *Late(Hesitation)-Acceptable* contributions. Finally, the third, *Early* disengagement class captures the interactions where the system terminated the engagement prematurely. With this grouping, the proportions of interactions in each of these categories across the four conditions are shown in Table 2 and Figure 5, together with the 95% confidence bounds. Statistically significant differences at a level below 0.05, based on a chi-squared test are shown with the corresponding *p*-values in Figure 5.

When comparing the baseline and forecasting model conditions, no significant differences are observed in the percentage of *Acceptable* disengagements. The percentage of *Costly* disengagements decreases slightly when using the forecasting model, but the difference appears at the expense of a corresponding increase in *Early* disengagements.

Using the hesitations-based policy leads however to larger reductions in the percentages of *Costly* disengagements: 47.4% (B) > 40.7% (B+H), and 41.6% (F) > 22.7% (F+H). The smallest percentage of *Costly* disengagements is attained when using the forecasting disengagement models in conjunction with the hesitation policy (F+H). The results indicate that a large proportion of these gains stem from reducing the number of *LateStart* contributions. As Table 5 indicates, the reductions in *Costly* disengagements when using the hesitation policy happen largely in conjunction with corresponding increasing in the *Acceptable* disengagements, rather than the *Early* ones. The decomposition of *Acceptable* disengagements in Table 5 indicates that in part, the gains observed when using the hesitation policy, *i.e.* when moving from B to B+H, or F to F+H, correspond to interactions where hesitations are inserted late in a manner that does not appear costly to the interaction – the *Late(Hesitation)-Acceptable* line.

Finally, with respect to *Early* disengagements, these can be triggered not only by false-detections of disengagement by the inference model, but also by speech recognition errors: for instance, the system asks *"Is there anything else?"* the participant responds *"Yes"*, and this is misunderstood as *"No"*; the system then terminates the interaction and bids the user goodbye. We report these incorrect *Dialog-based* terminations separately. The results indicate that with the forecasting model (F and F+H conditions) leads to a slightly increased percentage of *Model-based* early (compared to the B and B+H conditions); the observed differences do not reach statistical significance.

Disengagement type	B	B+H	F	F+H
Costly	**47.4%**	**40.7%**	**41.6%**	**22.7%**
LateContinue	20.3%	21.1%	16.8%	10.2%
LateStart	27.1%	14.6%	24.8%	8.6%
Late(Hesitation)-Costly	0.0%	4.9%	0.0%	3.9%
Acceptable	**40.6%**	**47.2%**	**41.6%**	**60.2%**
On-time	36.8%	39.0%	40.0%	47.7%
Late(NoProblem)	3.8%	4.9%	1.6%	4.7%
Late(Hesitation)-Acceptable	0.0%	3.3%	0.0%	7.8%
Early	**12.0%**	**12.2%**	**16.8%**	**17.2%**
Dialog-based	6.0%	4.1%	6.4%	5.5%
Model-based	6.0%	8.1%	10.4%	11.7%

Table 2. Breakdown of disengagement types

This analysis indicates that overall, the hesitation-based policies based on the forecasting model led to better behaviors on disengagement. However, these policies may also sometimes inadvertently trigger filled and non-filled pauses during the dialog, at times when participants are not disengaging. We therefore also took a closer look at the hesitations triggered by the system. Recall that the annotator tagged each hesitation as costly or not. We eliminated from the analysis below a set of 15 hesitation actions that were triggered in succession within a single interaction, where the robot incorrectly continued an engagement with a person that was close to the robot but involved in a conversation with someone else. Over the remaining hesitation actions, 81% were judged acceptable. This proportion was larger when using the forecasting model than when using the baseline heuristic: 86% (F+H) > 63% (B+H). A significantly larger number of hesitations were triggered when using the forecasting model: 130 (F+H) > 38 (B+H).

The annotator's observations revealed several classes of problems. Perhaps not unexpected, the production of hesitations actions at inopportune moments keeps the participants waiting too long for an answer that they expect to get immediately, or appears awkward, *e.g.*, when produced after a departure salutation from the user. At other times, the production of a filled pause appears to interrupt the users. Finally, sometimes, filled pauses produced while participants are actually disengaging can cause them to turn their attention back to the system, and sometimes stop in their tracks – the already discussed *Late(Hesitation)-Costly* case. An inspection of the *Late(Hesitation)* actions produced by the robot while participants were disengaging showed that 60% (6 out of 10 total) were costly when using the baseline heuristic (B+H), whereas on 33% (5 out of 15 total) were costly when using the forecasting model (F+H).

6. DISCUSSION

Below, we discuss limitations of this study, and future directions that remain to be explored.

Although the results indicate improvements in disengagement behaviors, recall that the baseline heuristic was designed to be conservative in order to avoid false disengagements and provide reliable labels. Future work should assess how the proposed approach fares in comparison to models learned based from labeled ground truth, or models manually tuned by other experts. Forecasting the actual moment of disengagement, as opposed to the estimate a conservative heuristic can provide automatically, may also be a more challenging learning problem. Finally, future work should assess the performance of the conservative heuristic and forecasting model against manually labeled ground truth.

One advantage of the approach we have described is that it does not require manual supervision. The system learns by itself, guided by a conservative heuristic, and tunes its behaviors to the specifics of its environment. Future work should investigate whether the type of conservative heuristic that we have constructed generalizes well to other settings, or whether it needs to be adapted. Even if some tuning is required for new environments, adapting the heuristic might still be less costly than performing manual annotations.

Another area of future exploration is the setting of thresholds: while the disengagement threshold P_{Dis}^{*} was based on a method that aimed to minimize false positives, the other thresholds for disengagement in the conservative heuristic, or for producing filled and non-filled pauses were chosen in a fairly arbitrary manner—by observing the behavior of the forecasting model on a few interactions, and were kept the same for both the baseline (B+H) and forecasting conditions (F+H). Better results may be attained with tuning, for both the baseline heuristic and forecasting model.

Perhaps even more importantly, the analysis of the form, function, and impact / cost of hesitation actions is a task that clearly deserves additional attention. The impact of a hesitation in dialog depends both on its form and on many contextual factors. Our study was constrained by the speech synthesis engine, but future work should investigate the use of other fillers, like *um* or *uh*, which might be more appropriate during disengagement as they don't carry the interrogative and floor release function that a prolonged "*So…*" sometimes does. Hesitations implemented as non-filled pauses, perhaps accompanied by appropriate non-verbal behaviors may generally constitute a lower cost option than filled pauses. Future work should also investigate how the choice of hesitations may be contextualized based on other aspects of the situation at hand.

7. CONCLUSION

We have investigated a self-supervised methodology for constructing forecasting models for disengagement, and proposed the use of disengagement policies that leverage linguistic hesitation actions to mitigate uncertainty. An initial study of the proposed approach with a directions-giving robot deployed in the wild shows that the joint use of forecasts and hesitations can help manage disengagement, and points to several interesting directions for future work. We believe that the study of policies for introducing hesitations is ripe for future exploration and that mastery of such policies will help systems to engage in more seamless, natural interactions with people in open-world settings.

ACKNOWLEDGMENTS

We thank Rebecca Hanson for assistance with data annotation, Anne Loomis Thompson and Nick Saw for project contributions, and the anonymous reviewers for their feedback.

8. REFERENCES

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

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

[3] Rich, C., Ponsler, B., Holroyd, A., and Sidner, C.L., 2010. Recognizing engagement in human-robot interaction, in *Proc. of HRI'2010*, Osaka, Japan.

[4] Michalowski, M.P., Sabanovic, S., and Simmons, R., 2006. A spatial model of engagement for a social robot, in *9th IEEE Workshop on Advanced Motion Control*, pp. 762-767.

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

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

[7] Clark, H.H., and Fox Tree, J.E., 2002. Using uh and um in spontaneous speaking, *Cognition*, 84(1):73-111, May, 2002.

[8] Corley, M., and Stewart, O.W., 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*, *Language and Linguistics Compass*, 4, 589-602.

[9] Goto, M., Itou, K., and Hayamizu, S., 1999. A Real-time Filled Pause Detection System for Spontaneous Speech Recognition, in *Proc. of Eurospeech'99*, 227-230, Budapest, Hungary.

[10] An, G., Brizan, D.G., and Rosenberg, A., 2013. Detecting laughter and filled pauses using syllable-based features, in *Proc. of Interspeech'2013*, Lyon, France.

[11] Adell, J., Bonafonte, A., and Escudero, D., 2010. Synthesis of filled pauses based on a disfluent speech model, in *Proc. of ICASSP'2010*, Dallas, TX.

[12] Skantze, G., and Hjalmarsson, A., 2010. Towards incremental speech generation in dialogue systems, in *Proc. of SIGDial'2010*, Tokyo, Japan.

[13] Skantze, G., Hjalmarsson, A., and Oertel, C., 2013. Exploring the effects of gaze and pauses in situated human-robot interaction, in *Proc. of SIGDial'2013*, Metz, France.

[14] Dethlefs, N., Hastie, H., Reiser, V., and Lemon, O., 2012. Optimizing Natural Language Generation for Decision Making for Situated Dialogue, in *Proc. of INLG'2012*, 49-58, Utica, IL.

[15] Bohus, D., Saw, C.W., and Horvitz, E., 2014. Directions Robot: In-the-Wild Experiences and Lessons Learned, in *Proc. of AAMAS'2014*, Paris, France.

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

[17] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B. 2009. Mechatronic design of NAO humanoid. *In Proceedings of ICRA'09*, Kobe, Japan.