Natural Communication about Uncertainties in Situated Interaction

Tomislav Pejsa*
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI, 53706
tpejsa@cs.wisc.edu

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

Michael F. Cohen
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
mcohen@microsoft.com

Chit W. Saw
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
nick.saw@microsoft.com

James Mahoney
Black Lobster Digital Arts
30 Bucket Ln
Yarmouth, ME 04096
james@BlackLobsterDigitalArts.com

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

ABSTRACT

Physically situated, multimodal interactive systems must often grapple with uncertainties about properties of the world, people, and their intentions and actions. We present methods for estimating and communicating about different uncertainties in situated interaction, leveraging the affordances of an embodied conversational agent. The approach harnesses a representation that captures both the magnitude and the sources of uncertainty, and a set of policies that select and coordinate the production of nonverbal and verbal behaviors to communicate the system's uncertainties to conversational participants. The methods are designed to enlist participants' help in a natural manner to resolve uncertainties arising during interactions. We report on a preliminary implementation of the proposed methods in a deployed system and illustrate the functionality with a trace from a sample interaction.

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*

General Terms

Algorithms; Human Factors

ICMI '14, November 12–16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2663249

* Research conducted during an internship at Microsoft Research

Keywords

Embodied conversational agents; situated interaction; multiparty; uncertainty; grounding

1. INTRODUCTION

Physically situated, multimodal interactive systems often rely on probabilistic inferences drawn from multiple streams of noisy perceptual data. Multiple uncertainties come to the fore during interactions. For example, dialog systems reason about uncertainties that arise in recognizing natural language via measures of recognition confidence and may employ clarification strategies such as explicit or implicit confirmations to keep a conversation on track. Robots or virtual agents that interact with users in physically situated settings grapple with uncertainties extending well beyond speech. These systems use probabilistic models to continually make state inferences based on streaming evidence from multiple sensors such as cameras, microphones, and proximity sensors. They must reason about people engaging with the system as well as those in the periphery, considering their physical position, orientation, and motion, focus of attention, interaction roles, intentions, and relationships. Beyond reasoning about the presence and contents of speech, these systems must monitor engagement (understanding who is involved in the interaction, when participants are joining or leaving, *etc.*) and turn-taking (understanding who is talking to whom, who has the conversational floor, to whom they are releasing the floor, *etc.*) Uncertainties arise in each of these processes and managing communication becomes particularly challenging.

Physically situated conversational agents may have at their disposal the affordances of a virtual or physical embodiment, with the ability to communicate via facial expressions, posture, and gaze. Just as in human-human interaction, these nonverbal affordances can be leveraged and coordinated with verbal grounding acts, so as to communicate about and help to resolve uncertainties.

We present a methodology for reasoning and communicating about the multiple uncertainties that arise in multiparty situated dialog. We use entropy as a measure of uncertainty about inferred beliefs

and consider the magnitude and the sources of each uncertainty. We introduce a set of policies to schedule and coordinate the production of verbal utterances and nonverbal behaviors that reflect the system's uncertainties. The affordances of embodiment (facial expressions, posture, and eye gaze) are leveraged to signal the inferred underlying causes of confusion and to direct signals to the users who are best positioned to assist with resolving the uncertainties. We have implemented the methodology in the context of a deployed, physically situated automated assistant.

We begin with a review of related work, followed by a description of the various types of inferences and associated uncertainties considered by a physically situated system. In Section 4, we describe the proposed approach and each of its components. In Section 5, we illustrate the approach with an analysis of a trace from a multiparty interaction with the implemented system. Finally, in Section 6 we summarize and present directions for future work.

2. RELATED WORK

In previous research in psycholinguistics, models of grounding [1] have been proposed to explain how participants establish and maintain mutual understanding over the course of the conversation. Participants coordinate their actions as they present utterances to each other, and they produce evidence of understanding in return. This evidence can come in multiple forms, *e.g.*, continued attention, a relevant next contribution, or an acknowledgement, such as a head nod or short verbal utterance (*"Uh-huh", "Yeah", etc.*). The latter acts are also referred to as feedback or backchannel responses [2]. Grounding is achieved if the provided evidence satisfies the grounding criterion [3], defined as follows: "The contributor [of the utterance] and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for the current purpose." If the grounding criterion is not satisfied, the contributing participant can initiate a repair, for example, by repeating or restating their previous utterance.

Computational models of grounding have also been proposed and have found practical implementation in human-computer interaction [4]. Such models include Traum's *grounding acts* model [5], which represents the grounding process as a series of communicative actions that serve specific functions, such as providing acknowledgment of understanding or requesting repair. An adaptation of the model has been proposed [6] within a framework for multiparty interaction with embodied agents, which considers conversation as a multilayer process requiring inferences about engagement, turn-taking, and utterance contents of multiple actors. Paek and Horvitz [7,8] explicitly model spoken dialog as decision-making under uncertainty. Their system uses Bayesian models to represent and infer uncertainty about the contents and targets of spoken utterances, requesting clarifications when uncertainty is high or when the expected value of seeking clarification is positive. Nakano et al. [9] propose an approach to multimodal, embodied grounding, which takes into account nonverbal grounding acts produced by users. Their system monitors users' gaze and head nods for evidence of understanding.

In addition to making inferences about the users' grounding acts, embodied agents need to produce coordinated verbal and nonverbal acts in response. Poggi and Pelachaud [10] propose a representation of communicative acts that decouples propositional content from performative aspects, mapping them to verbal utterances and nonverbal behaviors, respectively. Marsi et al. [11] model the agent's head motions and facial expressions to convey uncertainty about the information the agent is conveying through speech. Other work has focused on providing incremental embodied feedback to ongoing user utterances [12, 13], using verbal confirmations and clarification requests, as well as nonverbal cues such as head nods and facial expressions, to indicate the system's degree of understanding. The DeepListener system [8] uses verbal clarification requests paired with smooth shifting of the color of a glowing lens graphic to provide incremental feedback about the system's confidence and readiness to accept utterances.

Our work is related to the aforementioned efforts and aims to construct an integrative solution that enables embodied communication under uncertainty in situated, multiparty contexts. Like Traum and Rickel [6] and Paek and Horvitz [7, 8], we model conversation as action that happens at multiple levels, each supported by a different set of communicative competencies. The latter work also focuses on representations and handling of uncertainty. The main difference is that we focus on a broader range of communicative competencies, and construct an implementation that operates in the more complex context of a deployed multiparty, situated interactive system.

3. UNCERTAINTIES AND INTERACTION

Developing systems that are capable of engaging in interaction in physically situated settings leveraging multiple streams of sensory evidence hinges on a minimal set of communicative competencies, such as the ability to manage engagement and turn-taking and to recognize and understand speech. These competencies rely in turn on lower-level perceptual processes like vision and auditory scene analyses, and often involve probabilistic inferences. Inaccuracies in inferences can lead to failures at one or more levels of analysis, and ultimately to the breakdown of interaction.

Table 1 enumerates layered competencies for situated dialog and highlight key uncertainties encountered at each level. For concreteness, we shall focus on examples from the *Assistant* project. The Assistant is an embodied conversational agent that make continuous inferences about the presence and availability of its owner, and interacts with visitors on the handling of various administrative tasks, such as setting up meetings or relaying messages. The system is deployed outside its owner's office, and displays an animated, expressive head. A typical interaction might start with the Assistant looking up, making eye contact and saying, *"Hi, are you here to see [owner]?"*

Perception. Communicative competencies are anchored in a *perceptual layer*, which enables the system to sense its surrounding environment through the use of cameras, microphones, and other sensors. The visual channel often carries large amounts of information, and vision algorithms can be used to detect and track people in the environment, as well as their orientation, body pose, and gaze direction or visual focus of attention. Audio data, captured via microphone arrays, can be processed to infer the content and directional source of speech. Beyond sensing the physical environment, we also include in this category the processes that enable the system to access domain-specific information relevant for the task at hand. For instance, the Assistant has access to its owner's calendar and computing activities, as well as to inferences from predictive models about the expected cost of interruption of the owner, the probability distribution over time until the system's owner will return to his office or read email, and the likelihood of the owner attending particular meeting. Across all of these different channels, perception involves the use of probabilistic models, which produce beliefs over hidden world state variables, such as the location of a participant in the scene, their focus of attention, and the directions sounds are coming from. As these inferences underpin many of the higher-level communicative abilities, uncertainty and inaccuracies often propagate to the higher levels, where they can lead to interaction breakdowns.

Table 1: Uncertainties at various levels of processing modeled in the current implementation

Level	Uncertainty type	State variable	Potential uncertainty sources	Instances
Intention Recognition	Concept	Domain specific relevant intentions and information	UserIdentity, SpeechInputPresent, SpeechInputTarget, SpeechInputContents, SpeechInputSource	Per-concept
	UserIdentity	The identity (e.g. name) of a user	-	Per-user
Speech Understanding	SpeechInputTarget	Target of a spoken utterance	-	Per-input
	SpeechInputSource	Source of a spoken utterance	-	Per-input
	SpeechInputContents	Contents of a spoken utterance	-	Per-input
	SpeechInputPresent	Whether or not a spoken utterance was produced since the last system floor release	FloorReleaseDidAnyUserTake	1
Turn taking	FloorReleaseTarget	Target of the last user floor release	-	1
	FloorReleaseDidAnyUserTake	Indicates whether any user took floor since the last system floor release	FloorReleaseDidUserTake	1
	FloorReleaseDidUserTake	Indicates whether a given user took the floor since the last system floor release	-	Per-user
Engagement	Engagement	Whether a given user intends to be engaged with the system	FaceTracking	Per-user
Perception	FaceTracking	Whether a user is present at the location specified by the face tracker	-	Per-user

Engagement. A base-level challenge with interaction that must be resolved is that of initiating, maintaining, and terminating a connection with one or more participants in the conversation. We refer to this process as *engagement* [14]. In human-human interaction, engagement is a mixed-initiative process that involves verbal and nonverbal signals, including facial expressions, proximity, and body posture. In multimodal interaction, inferences about engagement intentions and actions are made with models that leverage lower-level perceptual and sensory evidence [15], such as a participant's location or focus of attention. Uncertainties may arise about whether a participant is still maintaining or terminating an interaction, or whether someone is initiating a new engagement. Inaccurate inferences may lead to incorrect initiations of engagement with people passing by, or to abrupt, early terminations of engagement while participants expect the interaction to continue.

Turn taking. Once engaged in an interaction, the system must coordinate with other participants on the production of verbal outputs throughout the conversation in a process known as *turn taking*. Like engagement, turn taking involves nonverbal signals and cues [16, 17, 18]. Managing multiparty turn taking is challenging and hinges on accurate tracking of several variables, including who is currently speaking or expected to speak, the sources and targets of utterances, and the floor control actions performed by the participants. The Assistant uses a turn-taking model [19] that relies on tracking which participant has the conversational floor, and the floor actions produced by each engaged participant, *e.g.* holding, taking, or releasing the floor. In this context, uncertainties can arise about which participant took the floor and to whom a floor release action is targeted towards. Making good turn-taking decisions, such as whether now is a good time for the system to start an utterance, depends on the probabilities inferred about who is likely to next start talking and when they will start [20] and inaccurate inferences about these variables may lead to interaction breakdowns. For instance, if the system misrecognizes a release of the floor by one participant to another as a release to the system, it may start talking at an inopportune moment, leading to a potential floor battle. On the other hand, if the system is not able to correctly recognize that a participant has taken the floor and has provided a response, it may pause for an unusually long time while waiting for the participant's turn, creating an awkward gap in the conversation.

Language understanding. Typically, automatic speech recognition and natural language understanding systems analyze the audio signal and produce an *n*-best list of hypotheses, which captures the belief over the contents of the utterance. In multiparty settings, systems must additionally identify the source and target of each utterance. Such an understanding of intentions and roles in a conversation is important in understanding contributions and in guiding plans for interaction. Left unchecked, uncertainty in speech recognition often leads to misunderstandings and can throw an interaction off track.

Intention understanding. At the next level, the contents of the decoded utterances must be placed in context and used in conjunction with the discourse history and other perceptual evidence to continuously track beliefs over the high-level user goals and intentions, as well as other world state information that are relevant to the task at hand. For instance, if the identity of a user is important for a given task, a physically situated system may update its belief over the identity by integrating probabilistic evidence from multiple decoded utterances across a clarification dialog about identity, *e.g.*, "*Sorry, did you say you're John?*", with evidence that is streaming continuously from a face recognizer. Inaccurate inferences at lower levels may lead to incorrect understanding at the intention level, *e.g.*, the system might incorrectly infer the user's identity on the basis of inaccurate facial recognition and speech recognition results.

In summary, uncertainties arise at many levels of the system, ranging from perceptual uncertainty (visual and auditory), to uncertainty about the state of the conversation (engagement and turn taking), to the higher-level uncertainty about intentions. We next discuss how we incorporate communications about important uncertainties into the interactions through nonverbal expressions and verbal utterances.

4. METHODOLOGY

We now describe methods that can provide interactive systems with capabilities to communicate about key uncertainties arising during an interactive session. Our goal is to give systems the ability to enhance interactive sessions by sharing uncertainties and enlisting the help of participants to resolve them—all in a natural manner. Key components of such capabilities include: (1) *uncertainty state estimation*—assessing and diagnosing relevant uncertainties in the system, (2) *uncertainty communication policy*—selecting and coordinating verbal and nonverbal behaviors that communicate one

or more uncertainties, with a goal of enhancing interaction by sharing or working to reduce uncertainty, and (3) *uncertainty behavior execution*—rendering these behaviors.

Complex interactive systems often use layered inferences per considerations of modularity and tractability. As such, key variables typically depend on inferences made about other variables, as well as attributes observed in streams of perceptual information. For instance, inferences for engagement intention may leverage inferences about lower-level variables, such as tracking confidence, estimated visual-focus of attention, proximity, and motion. In light of such dependencies, multiple steps can be taken to resolve the degree of uncertainty about one or more variables deemed to be important in the success of an interaction. In the general case, plans for sharing the uncertainty for a target variable can be guided by computing the *expected value of information* over all influencing variables and seeking to identify observations or states that would best resolve the uncertainty. Beyond information value, plans for sharing and coordinating about uncertainties must also consider the constraints of available gestures and perceived naturalness of interaction. The general problem of computing policies for identifying and communicating about key uncertainties is a complex decision-theoretic challenge. Below, we describe key representation elements and an initial framework and implementation that we have developed. The implementation uses handcrafted heuristics for identifying and communicating about key sources of uncertainty arising during interactions.

4.1 Uncertainty State Estimation

We assume as a starting point an existing system that tracks the world state ($\boldsymbol{S}$) via probabilistic inference models, *i.e.*, for each *world state variable* $X_i \in \boldsymbol{S}$ (or *variable* in short), an inference model computes the *belief* over X_i, *i.e.*, $p(X_i)$. Variables may represent any relevant properties of an individual user (*e.g.*, engagement intention, floor action, goals and domain specific intentions), of an utterance (*e.g.*, its source, target, or contents), of a floor action, etc. The belief captures the system's probabilistic estimate for the variable's possible values.

We measure the uncertainty in the inferred belief of a variable X_i by computing the entropy $H(X_i)$, defined as:

$$H(X_i) = \sum_j -p\left(X_i = x_j^i\right) \log_b \left(p\left(X_i = x_j^i\right)\right)$$

We consider influences among X_i and other variables used in the estimation of X_i in considering plans for communicating about and resolving potential failures in understanding. We use a recursive *uncertainty state representation* that captures not just the magnitude of the uncertainty (entropy), but also contributing influences or sources of error.

Specifically, we define an *uncertainty state* u_i as a tuple consisting of the *world state variable* X_i, its *uncertainty score* $H(X_i)$, and a set of *uncertainty sources* U_i, which are other uncertainty states:

$$u_i = (X_i, H(X_i), U_i)$$

In the current implementation of the framework, we identify uncertainty sources based on heuristic rules, informed by the structure of inferences in the system. For instance, at the highest level of the communicative stack (Intention level in Table 1), our system reasons about *concepts*, state variables which capture domain-specific information such as user goals and intentions, *e.g.*, *IsPersonOnCalendar*, *IsLookingForOwner*, etc. The system's belief over concepts is typically updated based on speech recognition results. As a consequence, a low-confidence recognition result can give rise to uncertainty over the concept. In such a case, the lower level *SpeechInputContents* uncertainty would be assigned as a source of the *Concept* uncertainty. However, speech recognition confidence is not the only potential source leading to concept uncertainties. In Section 5 we present in more detail an example where concept uncertainty arises based on uncertainty about which user spoke the utterance (*SpeechInputSource* uncertainty). The set of uncertainty states implemented so far, together with their corresponding state variables and their potential sources are shown in Table 1.

4.2 Policies for Communicating Uncertainties

At any given time, based on its assessment of the uncertainties present, the Assistant must decide which uncertainties to communicate to the users and how to communicate them. The actions are guided by the *uncertainty communication policy*. Given the state of the art with perception and recognition, uncertainties generally abound in situated systems, to a much larger degree than in human-human interaction. For instance, the Assistant will be confused much more often than a human would be about where a person is, whether they are still engaged, and who is speaking to whom. Thus, it may not be useful and could even appear unnatural if all uncertainties that arose were communicated. The uncertainty communication policy judiciously selects among *multiple* uncertainty states that may exist at any given time, and coordinates the production of corresponding communicative behaviors *across time*. This must be done in a manner that focuses the effort on the uncertainties that are most critical to the task at hand, while working to keep the interaction natural and useful. In the most general sense, the policy must take into account not only the instantaneous joint uncertainty state over all variables, but the history and dynamics of the uncertainty state.

In an effort to alleviate challenges with coordination between verbal and nonverbal behaviors for communicating uncertainty, we structured the uncertainty communication policy into two sub-policies: (1) a *speaking policy* selects and triggers verbal dialog acts, whose production may be accompanied by synchronized nonverbal gestures, and (2) a *listening policy* selects and triggers nonverbal gestures that communicate uncertainty when the agent is not speaking. We now describe each of these components.

4.2.1 Speaking Policy

The task of a *speaking policy* is to trigger verbal dialog acts that alert users about uncertainty and attempt to resolve grounding failures. Given the current world state and joint uncertainty state, every time the system is about to take a turn, the speaking policy may decide to produce a dialog act, *i.e.* a set of verbal utterances, coupled with nonverbal behaviors, that communicates and attempts to resolve a given uncertainty. This is a two-step process. In the first step, a dialog act is constructed, specifying the *semantic content* that is being communicated (a semantic representation of the prompt, including the domain-specific information communicated or requested), a set of *addressees* (users at whom the act will be addressed), and an *uncertainty communication action* that captures the uncertainty states being communicated and the set of *resolving users*, *i.e.*, users expected to be involved in the resolution of these uncertainties. In a second step, the dialog act is mapped into a lexical form (prompt) accompanied by nonverbal behaviors, which are then rendered via the embodied agent.

In the most general case, speaking policies produce dialog acts that aim to clarify uncertainties over concepts, so the uncertainty communication action will typically reflect a *Concept* uncertainty state. As previously described, the *Concept* uncertainty state may in turn point to other uncertainties as its sources. This representation of uncertainty sources enables the speaking policy to

Figure 1. Nonverbal expressions of uncertainty.

construct a more refined rendering of the dialog act that communicates not just general uncertainty about a concept, but also indicates the suspected sources of this uncertainty. For instance, if the source of the *Concept* uncertainty is low recognition confidence, *i.e.*, *SpeechInputContents* uncertainty, the policy may trigger a dialog act requesting a confirmation or clarification. Another example, discussed in Section 5, has *SpeechInputSource* uncertainty attached as source of the *Concept* uncertainty. In this case, the speaking policy can trigger a dialog act that points to the problem ("*Sorry, I can't tell who is speaking…"*) and asks for a clarification ("*Which one of you said they were John?*")

The set of resolving users specified as part of the uncertainty communication action is usually the same as the set of addressees; the system communicates the uncertainty to people who are expected to assist with its resolution. There are counter-examples, however. Sometimes concepts are not elicited from the users, but may be updated from external knowledge sources, *e.g.*, information about the owner's whereabouts and expected return time. When such information is uncertain, the uncertainty communication action does not specify any resolving users. As such, the produced verbal utterance and nonverbal behavior communicate the uncertainty in a way that does not explicitly solicit external help, for instance by producing a confused expression accompanied by a gaze avert (see example from Section 5.)

4.2.2 Listening Policy

The listening policy enables the system to communicate about uncertainty nonverbally, during listening periods, without committing to verbal dialog acts. This is in accordance with the principle of least joint effort in grounding [1], which states that partners in a conversation expend the minimal effort required to achieve mutual ground.

In general, the listening policy maps the various uncertainty states to certain nonverbal behaviors. For example, partial hypotheses generated by the speech recognizer during a user utterance may already indicate uncertainty about speech contents, as the utterance is in progress. Based on this, the listening policy can trigger a corresponding nonverbal expression of straining to hear, in an attempt to convey the uncertainty and also potentially shape future user behavior – in this case, speak more clearly. Similarly, a *FaceTracking* uncertainty is mapped to a straining to see expression, and so on. The set of expressions we have currently implemented is described in the next section.

When multiple uncertainties occur simultaneously, the listening policy must select which one to attend to. The prioritization is done based on two criteria. First, the system considers it more important to communicate uncertainty about things said or done by users who are currently involved in the interaction. The listening policy therefore prioritizes uncertainties pertaining to engaged users over ones pertaining to bystanders. Second, it is more critical to resolve uncertainties about communicative processes that are more fundamental to supporting the interaction. The listening policy therefore prioritizes uncertainties on the lower levels in the communicative stack (Table 1).

The listening policy also considers history when deciding whether it communicates about a particular uncertainty state. Specifically, we prevent excessive nonverbal feedback about the same communicative problem by using a simple principle: only one non-verbal communicative behavior is produced per listening stage for each uncertainty state.

4.3 Behaviors

Policies coordinate the production of verbal utterances and nonverbal behaviors to signal uncertainty. Nonverbal behaviors are constructed by coordinating facial expression changes, postural shifts, and eye gaze. In this section we provide a brief overview of the behaviors supported in our current implementation and explain the coordination mechanisms.

4.3.1 Facial Expressions and Posture

To support the production of nonverbal behaviors, we have developed a set of six predefined facial *expressions: StrainToHear, StrainToSee, StrainToPerceive, Bewildered, Confused, Understood.* These expressions were authored by a professional 3D artist, who was guided by his intuitions about human motion and prior literature on nonverbal expressions of uncertainty [21, 22].

Each expression includes both facial movements and posture changes. The expressions are parameterized by a continuous *intensity* score in the 0-1 interval. At the lower level, the animation procedure for each expression constructs a set of animation curves that smoothly animate the agent's facial expression and posture at the specified intensity. The construction of these curves involves a certain degree of randomization, so facial actions and posture shift directions are not always the same, even when the expression is triggered with the same intensity. Figure 1 shows the peak points of the six predefined expressions, applied at intensity 0.7.

The performance of nonverbal behaviors is carried out via a sequence of expression intensity updates. For listening policies, these updates are set to happen at predefined times. For example, when the listening policy executes the behavior for communicating *Engagement* uncertainty towards an engaged participant, the behavior first applies the *Bewildered* expression at intensity 0.15, then after 2.0 seconds it increases the intensity of the expression to 0.3. For speaking policies, expression changes are synchronized with verbal utterances via tags embedded into the prompts that can specify starting points, duration and intensities for expressions.

4.3.2 Eye Gaze

Apart from facial expressions and posture changes, the nonverbal uncertainty communication behaviors also control eye gaze. This is a particularly important affordance in multiparty interaction, as it enables the agent to unambiguously indicate addressees of communicative acts [23]. In the implementation, the agent relies on gaze to communicate uncertainty towards or to solicit help from specific participants.

Addressees for communicative acts about uncertainty are generally chosen on the basis of utility; the system conveys uncertainty to participants who are most likely to assist with resolution. In practice, this can mean gazing towards participants who are

believed to be "responsible" for uncertainty in the first place. For example, when there is *SpeechInputContents* uncertainty about an utterance, the non-verbal feedback produced by the listening policy is accompanied by gaze towards the user who is the source of the utterance. In other cases, uncertainty signals are addressed at multiple parties; an example is the dialog act about *SpeechInputSource* uncertainty presented in Section 5. On the other hand, informing acts that communicate a *Concept* uncertainty not arising from the communicative process are accompanied by averted gaze (see example in Section 5), as the agent does not expect users' help in its resolution.

5. EXAMPLE

We implemented the methodology described above within a previously developed infrastructure for physically situated, spoken language interaction used to construct the Assistant [24]. We demonstrate the operation of the implementation with a sample interaction with the Assistant. In the scenario, two people arrive for a meeting with Zack, the Assistant's owner. The Assistant knows about the meeting from Zack's calendar, but only one person is expected to arrive for that meeting. In addition, Zack himself is not there at the moment. Both of these facts coupled with ambiguity in verbal interactions lead to various forms of uncertainty within this short interaction. This scenario was constructed to illustrate the various functional aspects of the proposed methodology. A trace of the interaction with relevant key frames is shown in Figure 2 and a video is available at http://1drv.ms/1uwhfrm.

At the beginning of the interaction, two participants enter in the space in front of the Assistant, as they approach Zack's office. They do not directly engage with the Assistant, rather they peer into the office (located to the left). During this time (segment A) the engagement inference model indicates that the probability that the participants are trying to engage is in the middle range, which leads to an increase in the uncertainty scores about engagement for both users – see *Engagement* scores, Figure 2.A. As the Assistant has not yet engaged in a conversation, the listening policies, based on the high uncertainty state, activate and trigger a *Bewildered* expression (Figure 2.A.1), which persists for a few seconds (Figure 2.A.2). Several seconds later, as participants turn their attention and orient towards the system, uncertainty about engagement intention dissipates, and the system begins the interaction (Figure 2.A.3).

Later in the dialog (segment B), the Assistant tries to determine if one of the participants is the person expected for the 2 o'clock meeting (captured by the concept *IsPersonOnCalendar*), by asking: "*Is one of you John?*" The left participant responds "*Yes*". As the utterance is being produced, information from the sound source localizer is used to infer which participant is speaking. As the participants are close together, there is high uncertainty about the utterance source (Figure 2.B, *SpeechInputSource* score, pink graph). Throughout the production of the utterance, the system is listening; the listening policy activates (Figure 2.B.1) and responds to the *SpeechInputSource* uncertainty by triggering the *Bewildered* expression at a low intensity. Once the utterance finishes, while its contents are correctly understood, because the utterance source is uncertain a *Concept* uncertainty state for *IsPersonOnCalendar* is constructed that points to the lower level *SpeechInputSource* uncertainty state as its source. The speaking policy constructs an appropriate grounding act that both points to the source of the problem "*Sorry, I can't tell who is speaking when you stand so close together*", and requests a clarification "*Which one of you said they're John?*" The production of this verbal utterance is coupled with a nonverbal *Bewildered* expression which temporally extends and increases the intensity of the *Bewildered* expression already started by the listening policy (Figure 2.B.2). Furthermore, since this grounding act is directed towards both participants, the Assistant's gaze shifts between them throughout the production of these utterances (see System Gaze in Figure 2.B).

Informed by the agent about the speech source inference issue, participants move apart and the left participant responds "*I did.*" (Figure 2.B.3) The utterance is recognized with fairly low confidence, but sufficient for current purposes. This time there is little *SpeechInputSource* uncertainty, and the Assistant considers the *IsPersonOnCalendar* concept grounded. The speaking policy first triggers a verbal grounding act that presents evidence of understanding, and then continues the interaction. The understanding act is rendered by saying "*Right!*" coordinated with a nonverbal *Understood* behavior, rendered as a head-nod and confident smile (Figure 2.B.4).

Next, the Assistant tries to determine whether the participant on the right will be joining the meeting (Segment C). At this point the participants move slightly back and start speaking softly to each other. As time elapses and the Assistant does not receive an utterance, the floor inference model leads to increased uncertainty in the system about whether any of the participants took the floor (see *FloorDidAnyUserTake* score, Figure 2.C.1). The listening policy responds by beginning a *Bewildered* expression at low intensity (Figure 2.C.1). Simultaneously, uncertainty about *Engagement* also increases. With both uncertainties present, the engagement issue takes precedence–as it is on a lower level of the communicative stack–and the corresponding policy continues the existing *Bewildered* expression at low intensity (Figure 2.C.2).

Eventually, after enough time has elapsed with no contribution detected from the users, the system decides to take the turn. The speaking policy checks the *Engagement* uncertainty for each user and, seeing that it is high for both, plans a dialog act that communicates the system's uncertainty: the Assistant produces a filled pause "*So…*" (Figure 2.C.3.) Note that, at the moment that the Assistant takes the turn, the *FloorDidAnyUserTake* uncertainty collapses, as now the system has the floor. As the participants start turning back towards the system, the *Engagement* uncertainty for the right participant is reduced (*Engagement* score, dark red graph), which triggers the next system contribution. The *UserJoiningMeeting* is still not grounded, and in this case the speaking policy produces a clarification request for that concept, accompanied again by an extension of the *Bewildered* expression.

After the participant on the right responds, the Assistant informs the participants that the owner might be running late (segment D). The dialog act (Figure 2.D.1) presents the concept *OwnerShortReturn*, which is ungrounded and has a high *Concept* uncertainty score (Figure 2.D, *Concept* score, green plot). The verbal utterance reflects the agent's uncertainty (*"I think Zack is running a little late."*) and is coordinated with the production of a nonverbal *Confused* expression (Figure 2.D.2). The dialog act is targeted at both participants, but the system does not solicit help from either of them to resolve the *Concept* uncertainty. For that reason the *Confused* expression is accompanied with a coordinated gaze aversion gesture, where the Assistant's gaze is directed upwards, away from the participants (indicated by the gray region on System Gaze track in Figure 2.D)

6. CONCLUSION

We discussed key uncertainties that can arise in physically situated interactive systems, explored methods for endowing a system with the ability to reflect about and share its uncertainties with people, and characterized the general challenge of identifying and resolving key uncertainties. We described a specific implementation based on

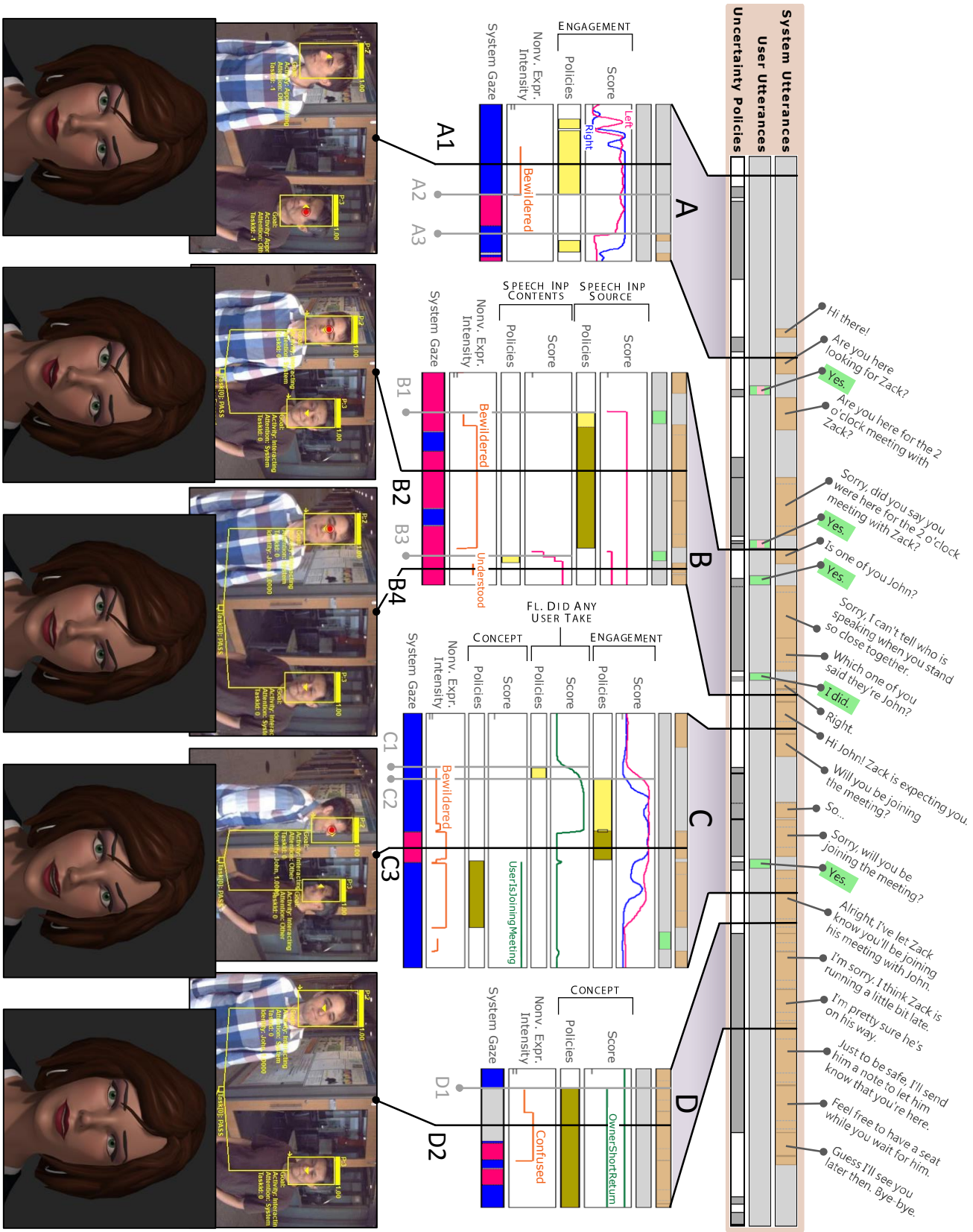

Figure 2. Trace of a demo interaction. (Top) System and user utterances, uncertainty policy activities. (Middle) Interaction segments, highlighting uncertainty scores (color-coded by the object of uncertainty; *purple* – right user; *blue* – left user; *green* – both users or concept), policy execution (*light yellow* – listening; *dark yellow* – speaking), nonverbal expression intensity, and system gaze (*purple* – right user; *blue* – left user; *grey* – averted gaze). (Bottom) System view and agent. In system view images, red dot is the agent's gaze target, yellow arrows indicate the participants' attention direction, while yellow lines show engagement.

a representation that captures both the magnitude and sources of uncertainty. We presented policies to coordinate the production of verbal utterances and nonverbal behaviors to communicate the system's uncertainties to the conversational participants and to enlist their help in resolving them. We illustrated the functionality of the proposed approach with a trace from an example interaction.

The described implementation, while still preliminary, provides a foundation for future developments and experimentation. Extensions to the approach include introducing uncertainty states for other inferences in the system, developing more sophisticated blame-assignment models to identify and communicate uncertainty sources, building robust, utility-driven policies that take into account the dynamics of uncertainty over time, and explicitly computing expected costs of persisting or resolving uncertainties when deciding whether to communicate about them. Finally, we have found that even subtle changes in the rendering of nonverbal expressions can have significant influence on the meaning seen in the signals by people, so we wish to further explore the rich lexicon of nonverbal uncertainty behaviors and their link to utterances.

The motivation for the work reported here lies in the hypothesis that, using the affordances of embodiment and communicating about uncertainty not only explicitly, but via the broader nonverbal channel, can lead to increased naturalness and efficiency in interaction. We expect that, in communicating its internal states of uncertainty, an agent may better convey to users its limitations, and implicitly shape their behavior so as to optimize interactions over time. We expect that such capabilities may have significant influence on user empathy and on the overall perceptions of the system. User studies will be required to test these hypotheses.

Finally, we note that uncertainty is only one factor in the successful grounding during the volley of contributions in conversational dialog. Other aspects include signaling about expectation and surprise, as well as expressing confidence and eureka following the resolution of key uncertainties, e.g., the *Understood* expression used in the example. We plan to continue work to understand and integrate other signals that may be useful in the engagement and overall grounding of communication and collaboration. We believe that the capability to express such rich internal states verbally and nonverbally, in stream with the evolving situation, will come to serve a central role in human-computer interaction.

ACKNOWLEDGMENTS

We thank Anne Loomis Thompson for her contributions to the project.

7. REFERENCES

[1] Clark, H. H., and Brennan, S. A. 1991. Grounding in Communication, *Perspectives on Socially Shared Cognition*. 127-149, Washington, DC.

[2] Yngve, V. 1970. On Getting a Word in Edgewise, *Proc. CLS 6*, 567-577.

[3] Clark, H. H., and Schaefer, E. F. 1989. Contributing to Discourse, *Cognitive Science*, 13, 259-294.

[4] Traum, D. R. 1999. Computational Models of Grounding in Collaborative Systems, *Proc. AAAI Fall Symposium on Psychological Models of Communication*, 124-131.

[5] Traum, D. R. A Computational Theory of Grounding in Natural Language Conversation. Ph.D. Thesis, Computer Science Dept., U. Rochester, December 1994.

[6] Traum, D. R., and Rickel, J. 2002. Embodied Agents for Multi-Party Dialogue in Immersive Virtual Worlds, *Proc, AAMAS'02*, 766-773.

[7] Paek, T., and Horvitz, E. 2000. Conversation as Action under Uncertainty, *Proc. UAI'00*, 455-464.

[8] Paek, T., and Horvitz, E. 2000. DeepListener: Harnessing Expected Utility to Guide Clarification Dialog in Spoken Language Systems, *Proc. ICLSP'00*, Beijing.

[9] Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. 2003. Towards a Model of Face-to-Face Grounding, *Proc. ACL'03*, 1, 553-561.

[10] Poggi, I., and Pelachaud, C. 2000. Performative Facial Expressions in Animated Faces, *Embodied Conversational Agents*, 155-188, Cambridge, MA.

[11] Marsi, E., and Rooden, F. 2007. Expressing Uncertainty with a Talking Head in a Multimodal Question-Answering System, *Proc. MOG'07*, 105-116, Enschede, The Netherlands.

[12] Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., and Stocksmeier, T. 2008. Modeling Embodied Feedback with Virtual Humans, *Lecture Notes in Computer Science*, 4930, 18-37.

[13] Kopp, S., Stocksmeier, T., and Gibbon, D. 2007. Incremental Multimodal Feedback for Conversational Agents, *Lecture Notes in Computer Science*, 4722, 139-146.

[14] Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. 2004. Where to Look: A Study of Human-Robot Engagement, *Proc. IUI'04*, 78-84.

[15] Bohus, D., and Horvitz, E. 2009. Learning to Predict Engagement with a Spoken Dialog System in the Open-World, *Proc. SIGdial'09*, London, UK.

[16] Duncan, S. 1972. Some Signals and Rules for Taking Speaking Turns in Conversation, *Journal of Personality and Social Psychology*, 23, 283-292.

[17] Sacks, H., Schegloff. E., and Jefferson, G. 1974. A Simplest Systematics for the Organization of Turn-taking in Conversation, *Language*, 50, 696-735.

[18] Wiemann, J., and Knapp, M., 1975. Turn-taking in Conversation, *Journal of Communication*, 25, 75-92.

[19] Bohus, D., and Horvitz, E., 2010. Facilitating Multiparty Dialog with Gaze, Gesture and Speech, in *Proc. ICMI'10*, Beijing, China.

[20] Bohus, D., and Horvitz, E., 2011. Decisions about Turns in Multiparty Conversation: From Perception to Action, in *Proc. of ICMI'11*, Alicante, Spain.

[21] http://center-for-nonverbal-studies.org/uncert.htm

[22] Stone, M., and Oh, I. 2008. Modeling Facial Expression of Uncertainty in Conversational Animation, *Lecture Notes in Computer Science 4930*, 57-76.

[23] Vertegaal, R., Slagter, R., Veer, G., and Nijholt, A. 2001. Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes, *Proc. CHI'01*, 301-308, Seattle, WA.

[24] Bohus, D., and Horvitz, E. 2009. Dialog in the Open World: Platform and Applications, *Proc. ICMI'09*, Boston, MA.