Computational Models for Multiparty Turn-Taking

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

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

Abstract

We describe a computational framework for modeling and managing turn-taking in open-world spoken dialog systems. We present a representation and methodology for tracking the conversational dynamics in multiparty interactions, making floor control decisions, and rendering these decisions into appropriate behaviors. We show how the approach enables an embodied conversational agent to participate in multiparty interactions, and to handle a diversity of natural turn-taking phenomena, including multiparty floor management, barge-ins, restarts, and continuations. Finally, we discuss results and lessons learned from experiments.

1 Introduction

Dialog among people is a highly coordinated mixed-initiative process, regulated by a stream of verbal and non-verbal cues. Participants in a conversation continuously (and often unconsciously) produce and monitor a large variety of cues, including verbal affirmations, head and hand gestures, and patterns of establishing and breaking eye contact. People in a conversation appear to understand with efficiency and ease the state and dynamics of who is signaling what to whom. They leverage this understanding to make contributions while seeking to minimize overlaps and gracefully resolve channel conflicts.

We focus our attention in this paper on the challenges of endowing a situated spoken dialog system with the ability to do appropriate and effective turn-taking in multiparty settings. The work comes in the context of a larger research effort aimed at developing a set of core competencies that would allow computer systems to interact naturally in open, dynamic, relatively unconstrained environments, and embed the interaction deeply into the flow of everyday tasks, activities and collaborations (Bohus and Horvitz, 2009a).

Most spoken dialog systems developed to date make a simplifying "you speak then I speak" assumption, which links input processing with turn-taking, *i.e.* the dialog manager produces an output after each received input. Various ad hoc solutions have been used to address common interaction phenomena which challenge this assumption, such as timeouts or user barge-ins. These methods are often insufficient and lead to turn-overtaking problems, even in dyadic conversations. The inadequacy of simple heuristics for guiding turn-taking become even more salient in multiparty settings, where several participants vie for the floor, where contributions that participants address to the system are interleaved with contributions they address to each other, and where events external to the conversation can impinge on the urgency of a participants' need to make a contribution.

We describe efforts to develop a principled framework that can endow a situated spoken dialog system with effective turn-taking competencies, leveraging audiovisual and contextual evidence. The approach centers on developing machinery enabling dialog systems to explicitly represent, reason about, and make real-time decisions about turn-taking in multiparty interactions. The methods rely on components for tracking the conversational dynamics (*e.g.*, who is talking to whom) and the floor control actions that regulate these dynamics, for making turn-taking decisions (*e.g.*, when is it my time to speak, or to stop), and for rendering these decisions into a set of corresponding behaviors (*e.g.*, establishing and breaking eye contact).

2 Related work

In a seminal paper on turn-taking in natural conversations, Sacks, Schegloff and Jefferson (1974) proposed a basic model for the organization of turns in conversation. The authors highlight the *locally managed* and *interactive* nature of the turn-taking process, and propose that discourse contributions are constructed from successive *turn-constructional-units*, separated by *transition-relevant-places* which provide opportunities for speaker changes. Turn allocation is performed either by the last speaker, or via self-selection, and is governed by a set of rules which apply at transition-relevant-places. In later work, Schegloff further elaborates a number of aspects of this model, including interruptions (Schegloff, 2000a) and overlap resolution devices (Schegloff, 2000b).

Subsequent works in the conversational analysis community have highlighted the important role played by gaze, gesture, and other non-verbal communication channels in regulating turn-taking in interaction. For instance, Duncan (1972) brings to fore the role of non-verbal signals in turn-taking, and proposes that turn-taking is mediated via a set of verbal and non-verbal cues. Wiemann and Knapp (1975) survey a number of previous investigations on turn-taking cues in a variety of conversational settings, and perform a quantitative analysis of such cues in dyadic conversations in an effort to elucidate some of the reported differences. Goodwin (1980) also discusses various aspects of the relationship between turn-taking and attention.

The first computational implementation of a comprehensive turn-taking model was done by Thorissön (2002). Thorissön discusses the need for a layered architecture with several update loops operating at different speeds in order to support locally regulated and dynamically produced turn-taking behaviors. Sensing is performed by binary unimodal perceptors fused via multimodal integrators described by heuristic rules. Turn-taking decisions are also implemented by rules based on Boolean combinations of observed features.

More recently, Raux and Eskenazi (2009) describe a turn-taking model for dyadic interactions based on a 6-state non-deterministic finite-state-machine, and show how it applies to the problem of end-of-turn detection.

Moving beyond dyadic interactions, Traum and Rickel (2002) describe a turn-taking management component, as part of a layered architecture for supporting multiparty dialogue between a trainee and multiple virtual humans in immersive environments. The proposed approach centers around five turn-taking acts: take-turn, request-turn, release-turn, hold-turn, and assign-turn.

More recently, several efforts have been directed at modeling specific aspects of the turn-taking process. For instance, Bell et al. (2001) describe a system that uses a semantic parser to classify incoming utterances as *closing* or *non-closing*. Others have used machine learning in conjunction with prosodic, syntactic, semantic, or dialog features to make predictions about closing (Sato et al., 2002; Ferrer et al., 2003; Takeuchi, 2004; Schlangen, 2006), and to optimize end-pointing thresholds (Raux, 2008) in spoken dialog systems.

The new work described here draws inspiration from prior research and extends it in several ways. We present details of a comprehensive computational framework for managing turn-taking in multiparty, open-world spoken dialog systems. We discuss key abstractions, models and inferences, and demonstrate how we can enable a system to handle a broad spectrum of naturally occurring turn-taking phenomena (*e.g.* multiparty floor management, barge-ins, restarts, continuations, etc.) We test the approach and report results collected during an evaluation of the turn-taking framework in a multiparty setting.

3 Models

Once engaged, participants in a conversation coordinate with each other on the presentation and recognition of various verbal and non-verbal signals. Spoken language generation and processing fundamentally occurs via a serial verbal channel. Thus, conversations are largely constrained to a volley of contributions generated and received by actors, and depend critically on coordinative signals about the role and timing of speaking and listening. Cues are used in such inferences as the likely end of verbal contributions, intentions to contribute, successful transmissions, and the current and next targets of utterances.

We model turn-taking as a *collaborative, interactive process* by which participants in a conversation monitor each other and take coordinated actions in order to ensure that (generally) only one participant speaks at a given time. The conversation participant that is ratified to speak via this collaborative process is said to have the *conversational floor* (henceforth *floor*). We shall assume in our modeling that only one of the participants involved in a given conversation has the floor at any point in time. An open-world interactive system working with multiple parties can however keep track of several conversations, and hence multiple floors, one associated with each conversation.

The floor shifts from one participant to another based on coordinated *floor management actions* that are continuously produced by the participants. Specifically, the model allows four floor management actions that each participant may perform at any point in time. The *Hold* action indicates that a participant who has the floor is engaged in a process of holding the floor. The *Release* action indicates that a participant who has the floor is in the process of yielding it to another participant. The *Take* action indicates that a participant who does not have the floor is trying to acquire it. Finally, the *Null* action indicates that a participant who does not have the floor is simply observing the conversation, without issuing any floor claims.

Floor shifts happen in the proposed model as the result of the joint, cooperative floor management actions taken by the participants. Specifically, a *Release* action by one participant has to be met with a *Take* action by another in order for a floor shift to occur; in all other cases, the floor remains with the participant who initially had it. A number of concrete floor management examples are discussed in Section 4, and in Appendix A.

Below, we discuss mechanisms for representing, reasoning, and decision making about turn-taking. The proposed framework includes components for: (1) sensing conversational dynamics, (2) real-time turn-taking decisions, and (3) rendering decisions into appropriate behaviors. A high-level view of these components and their relationships is displayed in Figure 1. In the following three subsections, we describe each component in more detail, including key abstractions, inferences and models. We also describe the current implementations and review opportunities for machine learning.

3.1 Sensing Conversational Dynamics

The sensing component is responsible for real-time tracking of the conversational dynamics, and includes models for detecting spoken signals, inferring the source and the target of each detected signal, and the floor state, actions and intentions of each participant engaged in a conversation.

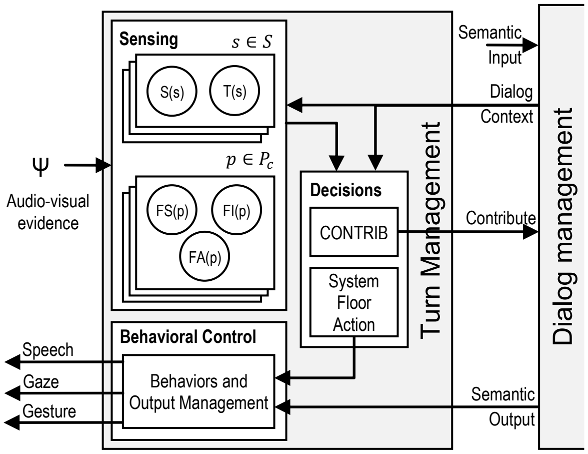


Figure 1. Components of a turn-taking architecture

Detecting spoken signals. Let $S = \{s\}$ be the set of all spoken signals being performed at a given point in time. We use an energy-based voice-activity detector to identify and segment these signals in the audio stream. This solution has a number of limitations, especially in multiparty setting when multiple users might speak simultaneously, or after one another. We plan to investigate audio-visual speaker diarization using sound-source localization provided by a microphone array.

Sensing signal source and targets. For each detected signal s, we represent the source of that signal as a multinomial variable $S(s)$ ranging over $A \cup \{\eta\}$, where A is the set of all observed actors and η denotes an unknown source or background.

Creating an appropriate representation for the signal target raises additional challenges. Clark and Carlson (1982) identified several roles that participants can have with respect to a given utterance: *addressees, side participants, overhearers,* and *eavesdroppers*, described in Table 1 and illustrated in Figure 2. To capture these roles, we represent the signal target via a couple $T(s) = \langle c, ADR \rangle$, where c denotes a conversation and AD denotes a subset of the participants in c ($ADR \subset P_c$) which are the addressees of signal s. Since we are assuming that the set of participants P_c is tracked by the engagement components in the system (Bohus and Horvitz, 2009b), this representation automatically determines the addressees (ADR), the side participants ($P_c \setminus ADR$), and overhearers ($A \setminus P_c$).

In the current implementation, the source is assigned to be the participant whose location is clos-

Role	Description
Addressee	participant that utterance is addressed to
Side participant	participant that utterance is not addressed to
Overhearer	others known to the speaker who are not participants in conversation but will hear the utterance
Eavesdropper	others not known to the speaker who are not participants in the conversation but will hear the utterance

Table 1. Addressee roles in multiparty interaction.

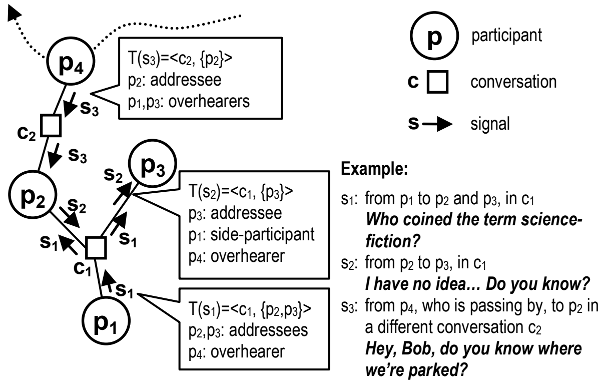


Figure 2. Sample multiparty interaction with illustrated addressee roles on three different signals.

est in the horizontal plane to the sound source direction identified by the microphone array. The target conversation is assumed to be the active system conversation. The addressee set is identified by a model that integrates information about the attention direction for the source participant (obtained through a face detection and head pose tracking algorithm) while the utterance is produced. In addition, non-understandings are assumed to be addressed to the set of other engaged participants, rather than to the system, since initial tests with the system indicated that about 80% of utterances that lead to non-understandings are in fact addressed to others.

These initial models were created to enable multiparty interaction data collection and to provide a baseline for evaluating the proposed approach. We believe that models learned from data that perform joint inferences about all participants and leverage audiovisual information (*e.g.* prosody, head and body pose, etc.) and high-level interaction context (*e.g.* who spoke last, where is the system looking at, etc.) can perform significantly more accurately than our initial implementation. We are investigating such solutions, based on the collected data.

Sensing floor state, actions and intentions. For each engaged participant $p \in P_c$, we represent whether or not the participant has the floor with a binary variable $FS(p)$.

Floor management actions for each participant are represented by a couple, $FA(p) = \langle fma, FRS \rangle$. fma denotes the floor action type, and is a multinomial variable ranging over the four actions previously described: *Take, Release, Hold,* and *Null*. The floor release set FRS denotes the subset of conversational participants ($FRS \subset P_c$) that the floor is being released to during *Release* actions; for all other actions, $FRS = \emptyset$. Note that this representation allows us to model both last-speaker based turn allocation and self-selection based turn allocation, as per (Sacks et al., 1974): $FRS \neq \emptyset$ in the first case, and $FRS = \emptyset$ in the latter.

For each participant, we also model the *floor intention*. Like the floor state, this is represented by means of a binary variable $FI(p)$ which indicates whether or not p intends or desires to have the floor. Although floor intentions may generally be captured by the observed floor actions, we model them separately as representing and reasoning about intentions may be valuable in predicting actions and in guiding turn taking.

We currently use a simple model for inferring floor management actions: if a participant has the floor, we assume they are performing a *Hold* action if speaking and a *Release* action otherwise; the floor release set is assigned to the set of addressees for the last spoken utterance. When a participant does not have the floor, we assume they are performing a *Take* action if speaking or a *Null* action otherwise. For now, we assume the floor intentions are reflected in the floor actions, *i.e.* a participant intends to have a floor if and only if he or she performs a *Hold* or *Take* action. Finally, the floor states are updated based on the joint floor actions of all participants, as described earlier.

The floor management action inference models described above use limited evidential reasoning. We are investigating the development and use of more accurate, data-driven models that leverage a rich set of audiovisual and contextual features for inferring floor actions and intentions.

3.2 Turn-Taking Decisions

The proposed framework eliminates the traditional "you speak then I speak" turn-taking assumption, and decouples input processing from response generation and turn-taking decisions. All inputs are processed (*i.e.,* the dialog manager performs discourse understanding and state updates) as soon as they are detected. However, the decisions (1) to generate a new contribution and (2) the selection of the floor management action to be performed by the system are made separately.

The decision to generate a new contribution ($CONTRIB$) is currently based on a set of rules that take into account the turn-taking context (*i.e.*, floor state, actions and intentions for each participant in the scene), and capture basic turn-taking norms. Specifically, the signal for a new system contribution is triggered when the floor is being released to the system, or when the floor is being released (by the system or another participant) to someone else, but is not taken by anyone for more than a threshold amount of time.

When the need for a new contribution is signaled, the dialog manager generates a new semantic output. Semantic outputs are eventually rendered by the behavioral component, which coordinates with the system's floor management actions (*e.g.* a spoken output is not generated until the system has acquired the floor, etc.) Each semantic output can include parameters that configure the system's floor management policy about that output. Examples include whether the output should be rendered immediately or pending on a floor release to the system, whether the output can be interrupted by the user, whether to release the floor at the end of that output, etc.

The system employs the same four floor management actions described above. The action to be performed (SFA) is selected by a set of rules that leverage the real-time turn-taking context, as well as high level dialog information, such as the set of planned system outputs. For instance, if the system has an output that is pending on a floor release, and a user is releasing the floor to the system, then the system will execute the *Take* action.

3.3 Rendering Behaviors

The system's floor management actions must be rendered in a set of accurately timed behaviors, contextualized to the particular embodiment of the system and to the state of the conversational scene. For instance, in an embodied conversational agent, apart from triggering the beginning or the end of the system's utterances, actions such as *Take* and *Release* minimally involve coordinated gaze and gesture behaviors. The specific structure of these behaviors can further depend on the state of the scene (*e.g.* the way the system behaves during a *Hold* action is sensitive to whether another participant is trying to take the floor.)

Action	Behavioral implementation
Hold	When speaking to one addressee, gaze is directed to that addressee; when speaking to multiple addressees, gaze is directed successively (based on a stochastic model) to each addressee; when the system is not speaking (but still holding the floor) it avoids eye contact by looking away from all participants.
Release	The system gazes at one of the participants in the floor release set (FRS). If FRS includes multiple participants and the participant being gazed at does not take the floor for a period of time, or directs their attention away from the system, the system switches gaze to another participant in FRS.
Take	The system gazes at the participant that holds the floor.
Null	The system gazes to the speaking participant, or, when no participant is speaking, to one of the participants that the floor is being released to.

Table 2. Behavioral implementation for floor actions.

The current behavioral models are informed by the existing literature on the role of gaze in regulating turn-taking, but are still relatively coarse. They are described in Table 2 and several examples are presented in Subsection 4.2 and Appendix A. We are investigating further refinements including modulating gaze and prosody, adding new gestures, producing backchannels, and leveraging additional conversational and scene context.

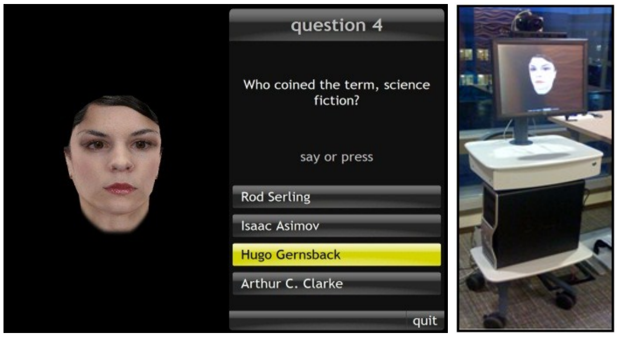
4 Experiments

We implemented the proposed multiparty turn-taking models in the context of a larger project on open-world spoken dialog systems (Bohus and Horvitz, 2009a), and conducted multiparty interaction experiments with one such system.

4.1 System and Application

The system used in the experiments described below takes the form of a multimodal interactive kiosk that displays an avatar head with controllable pose and limited facial gestures. We explored turn-taking competency by running a trivia questions game on the system. The game employs natural language, and can involve one or multiple participants.

The system uses a wide-angle camera and a microphone array, and includes components for detecting and tracking multiple participants in the scene, sound source localization, speech recognition, conversational scene analysis (*e.g.* running inferences about focus-of-attention, engagement, turn-taking, long-term goals and activities of actors in the scene, etc.), behavioral control and dialog



question 4

Who coined the term, science fiction?

say or press

Figure 3. System running the trivia game application.

management. Details on the system and its components are available in (Bohus and Horvitz, 2009a).

Each session of the trivia questions game proceeds as follows: after an initial opening phase, the system begins to challenge users with a sequence of trivia questions, displaying the set of possible answers for each question on the screen (see Figure 3). The users pick an answer, and, after a confirmation, the avatar provides a quick explanation if the answer was incorrect, and then moves to the next question. When playing with multiple people, the avatar can address and orient its head pose towards each participant individually. If one participant provides an answer, the avatar seeks confirmation from another participant before moving on. In some cases, the avatar will seek confirmation nonverbally, by simply turning towards a different participant and raising its eyebrows. If the participants talk amongst themselves, the avatar monitors their exchanges and waits until the floor is being released back to it. During this period, if the system hears one of the answers in this side-conversation (*e.g.* one participant suggests it to the other participant), the answer will be highlighted on the screen, as shown in Figure 3. If a significant pause is detected in during this side conversation, the avatar can take the initiative, *e.g. "So, what do you think is the correct answer?"* These, as well as a number of other multiparty turn-taking behaviors are illustrated and discussed in more detail in the next subsection, and in Appendix A.

We conducted a user study with the system described above, consisting of 15 sessions with a total of 60 participants. In each session, 4 participants were organized into 6 pairs and 4 triplets, and each group played one game with the system (the group order was randomized.) Each session thus resulted in 10 multi-participant interactions with the system (a total of 150 interactions).

4.2 Sample Interaction

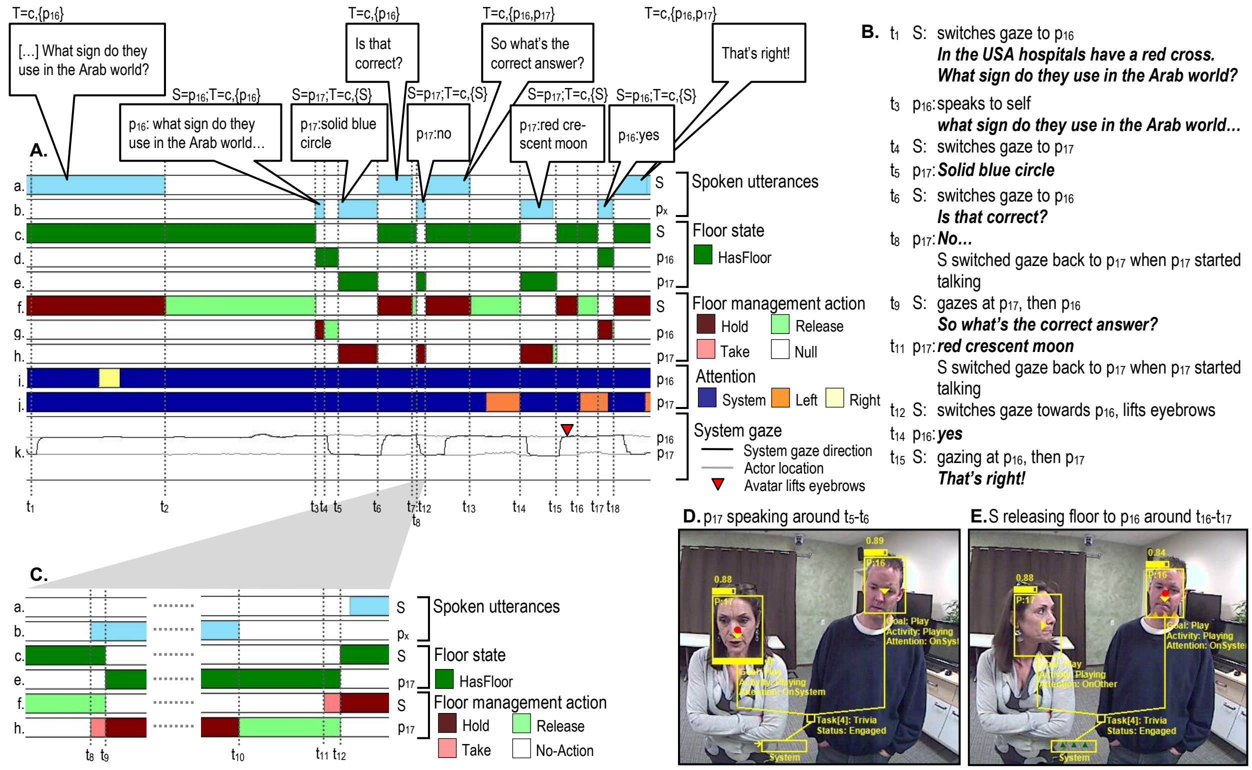
Figure 4 illustrates an interaction segment from the study, showing how the proposed models enable the system to handle several multiparty turn-taking phenomena; the corresponding video sequence is available online (Situated Interaction, 2010). An additional example is discussed in Appendix A.

In the example from Figure 4 the system asks a question at time t_1. While producing this utterance, (track *a*, t_1 to t_2) the system has the floor (track *c*), and is performing a *Hold* action (track *f*). Since the utterance is addressed to p_{16}, the *Hold* behavior shifts the avatar gaze at the beginning (time t_1) towards p_{16} (track *k*). At t_2 the system finishes the utterance and switches to a *Release* action (track *f*). The floor release target is p_{16}; at the behavioral level the system continues to gaze at p_{16}.

At time t_3, the system detects that p_{16} starts talking: he is echoing the system's question. This utterance leads to a non-understanding, and the system infers (according to current models) that the utterance is addressed to p_{17}, and is followed by a floor release by p_{16} to p_{17} starting at time t_4 (track *g*). These inferences are in fact inaccurate: in reality, the utterance was self-addressed and p_{16} was performing a *Hold* action, even after he finished talking. While the representations described earlier allow us to capture self-addressed utterances, and floor *Hold* actions performed when a user is not speaking, the current model implementation does not detect these phenomena. At the same time, the costs of this particular error are not very high.

At time t_4, since p_{17} becomes the floor release target, the behavior associated with the system's *Null* action switches the gaze from p_{16} to p_{17}. Shortly thereafter, at t_5, p_{17} answers by proving a response. Next, the system attempts to check whether p_{16} agrees this is the correct answer – notice again the floor shifting to the system at time t_6 (tracks *e*,*c*), and the gaze shifting to p_{16} (track *k*). As soon as the system releases the floor to p_{16} though, p_{17} intervenes and corrects herself: "No!"

The fine details of the corresponding floor shift from the system to p_{17} and then back (time t_8-t_{12}) are shown in Figure 4.C. As soon as speech is detected (t_8, track *b*), the source is identified as p_{16}. The floor activity inference model therefore recognizes a *Take* action from p_{17} (t_8, track *h*). Since the system is performing a *Release* action at this point (track *f*), the floor shifts from the system to p_{17} at



B. t1 S: switches gaze to p16
In the USA hospitals have a red cross. What sign do they use in the Arab world?
t3 p16: speaks to self
what sign do they use in the Arab world…
t4 S: switches gaze to p17
t5 p17: ***Solid blue circle***
t6 S: switches gaze to p16
Is that correct?
t8 p17: ***No…***
S switched gaze back to p17 when p17 started talking
t9 S: gazes at p17, then p16
So what's the correct answer?
t11 p17: ***red crescent moon***
S switched gaze back to p17 when p17 started talking
t12 S: switches gaze towards p16, lifts eyebrows
t14 p16: ***yes***
t15 S: gazing at p16, then p17
That's right!

Figure 4. Logs and images from a sample interaction collected in the multiparty turn-taking user study.

the next tick (t_9, tracks c and e). At time t_{10} the system detects the end of the user's utterance. The addressee is the system (inferred based on the fact that p_{17}'s focus-of-attention stays on the system throughout this utterance – track i). The floor inference models detect that p_{17} is performing a *Release* action to the system at time t_{10}. This in turn triggers a new contribution signal. The dialog manager plans the new output, this time addressed to both p_{16} and p_{17}: *"So what's the correct answer?"* To render the output, the system performs a *Take* action at time t_{11}, which at the next time tick, t_{12}, shifts the floor to the system (tracks *f, h, c, e*). The system then begins speaking and the *Hold* behavior keeps the avatar's gaze initially towards p_{17}, and then turns it to p_{16} (track k) since this utterance is addressed to both participants.

At time t_{14}, p_{17} provides an answer. At time t_{15}, the system takes the floor again, but this time silently redirects its gaze towards p_{16}, and lifts its eyebrows in a questioning look (tracks *c, k*). Immediately, p_{16} responds *"Yes!"* This is an example of a floor release where the system induces a contribution from one of the participants in an entirely non-verbal fashion. Next, at t_{18} the system informs the participants that the answer was correct.

4.3 Empirical evaluation

Each participant filled out a post-experiment survey (Situated Interaction, 2010) containing questions about the system's turn-taking capabilities, as well as open-ended questions (*e.g.* what did you like best / worst). Figure 5 shows the resulting mean of responses (on a 7-point Likert scale) and the corresponding 95% confidence intervals. Sample responses to the open-ended questions are shown and discussed more in Appendix B.

Participants generally rated the system's turn-taking abilities favorably. An in-depth analysis of the data and performance of various component models is beyond the scope of this paper and is described elsewhere (Bohus and Horvitz, 2010). Anecdotally, initial analyses revealed that a large proportion of errors stem from shortcomings of the voice activity detector and of the current models for identifying the signal source and addressees.

5 Conclusions and Future Work

The performance of the multiparty turn-taking models described here suggests that the approach can serve as a base-level platform for research on

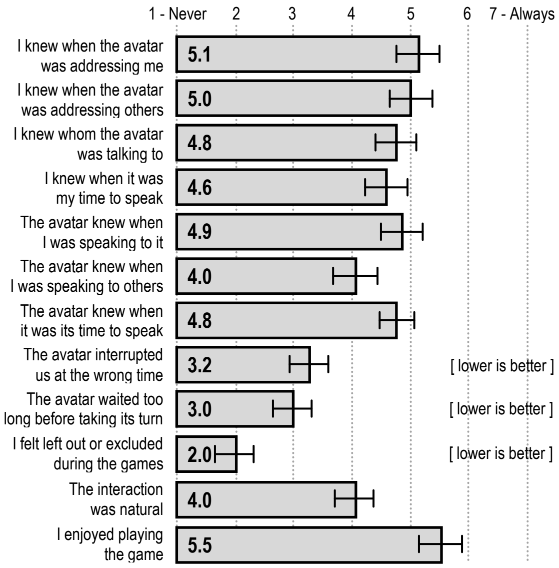


Figure 5. Responses to post-experiment survey.

problems of sensing and decision making for multiparty turn-taking. The methods can be enhanced with use of predictive models learned from case libraries of interactions, including inferences about utterance end points, source and target of utterances, and about actor intentions more generally. Inferential reasoning will likely benefit from the use of methods that can perform continuing analyses that can trade off timely actions with the greater accuracies promised by delays to collect additional audiovisual evidence. Utility-theoretic methods can be employed to perform context-sensitive analyses of the expected value of these and other turn-taking actions. There is also much to do with the appropriate generation and consideration of subtle back-channel cues for enhanced signaling and naturalness of conversation. Further research also includes developing competencies that imbue spoken dialog systems with insights about social norms, including appropriate behaviors for handling situations of overhearing and eavesdropping. We are excited about tackling these and other challenges on the path to fielding fluent conversational systems in the open world.

Acknowledgments

We thank Anne Loomis, Qin Cai, Cha Zhang, and Zicheng Liu for their contributions to this project. We also thank our colleagues who participated in pilot experiments for the user study.

References

Bell, L., Boye, J., and Gustafson, J., 2001. *Real-time handling of fragmented utterances,* in Proc. NAACL-2001 workshop on Adaptation in Dialog Systems.

Bohus, D., and Horvitz, E., 2009a. *Dialog in the Open-World: Platform and Applications*, in Proc ICMI'09, Boston, US.

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

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

Clark, H., and Carlson, T., 1982. *Hearers and speech acts,* Language, 58, 332-373.

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

Ferrer, L., Shriberg, E., and Stolcke, A. 2003. *A Prosody-Based Approach to End-Of-Utterance Detection That Does Not Require Speech Recognition,* in Proc. ICASSP-2003, 608-611, Hong Kong.

Goodwin, C. 1980. *Restarts, pauses and the achievement of mutual gaze at turn-beginning,* Sociological Inquiry, 50(3-4), 272-302.

Raux, A. and Eskenazi, M., 2008. *Optimizing end-pointing thresholds using dialogue features in a spoken dialogue system,* in Proc. SIGdial-2008, Columbus, OH.

Raux, A. and Eskenazi, M., 2008. *A Finite-State Turn-Taking Model for Spoken Dialog Systems,* in Proc. HLT-2009, Boulder, CO.

Sacks, H., Schegloff. E., and Jefferson, G. 1974. *A simplest systematics for the organization of turn-taking in conversation*, Language, 50, 696-735.

Sato, R., and Higashinaka, R., 2002. *Learning decision trees to determine turn-taking by spoken dialogue systems,* in Proc. ICSLP-2002 861-864, Denver, CO.

Schegloff, E. 2000a. *Accounts of Conduct in Interaction: Interruption, Overlap and Turn-Taking,* The handbook of sociological theory, 287-321, New York: Plenum.

Schegloff, E. 2000b. *Overlapping talk and the organization of turn-taking in conversation,* Language in Society, 29, 1-63.

Schlangen, D., 2006. *From reaction to prediction: Experiments with computational models of turn-taking,* in Proc. Interspeech 2006, Pittsburgh, PA.

Situated Interaction, 2009. *Project page:* http://research.microsoft.com/~dbohus/research_situated_interaction.html

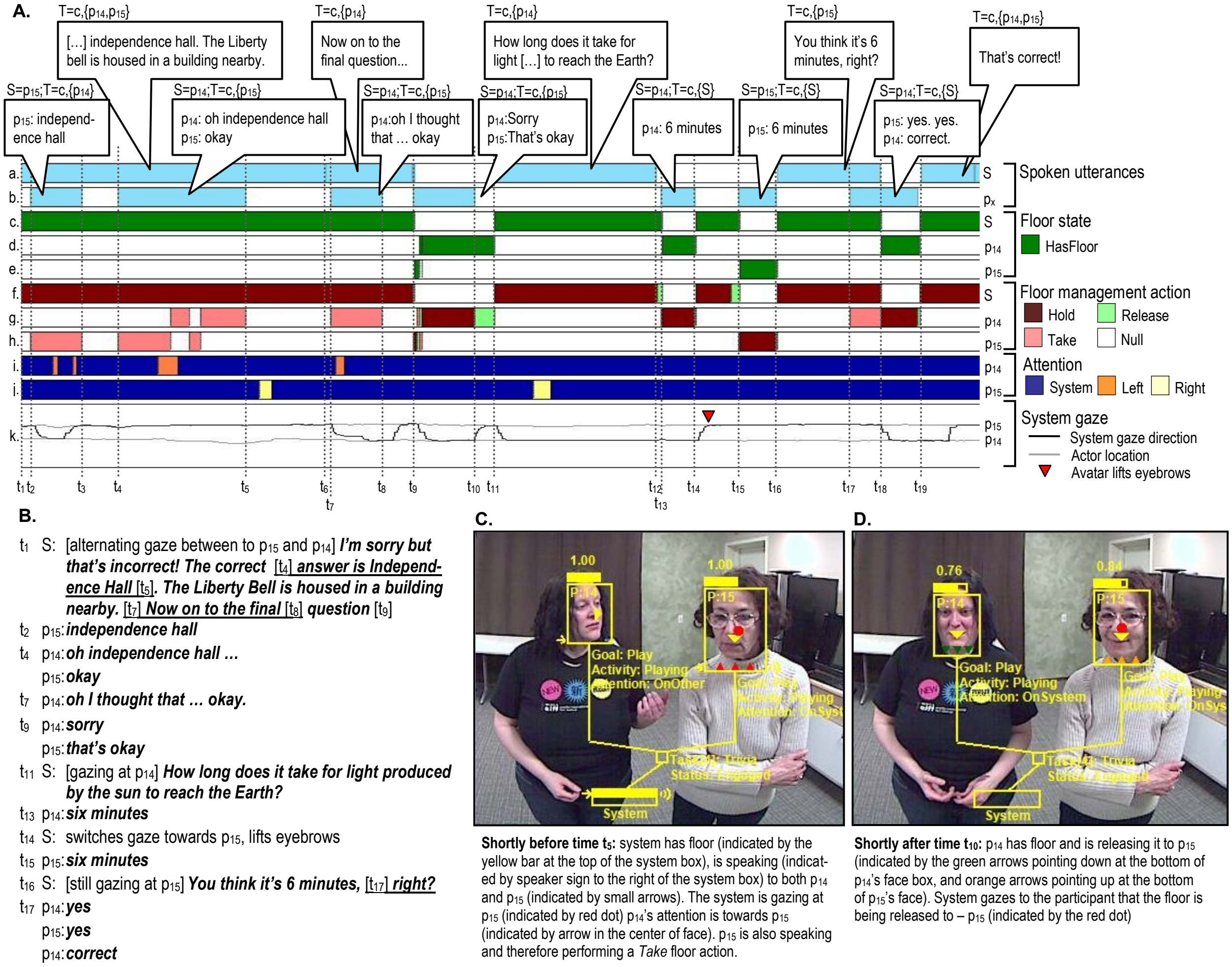
Takeuchi, M., Kitaoka, N., Nakagawa, S. 2004. *Timing detection for realtime dialog systems using prosodic and linguistic information*, in Proc. International Conference: Speech Prosody 2004, 529-532.

Thorisson, K.R. 2002. *Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perceptions to Action*, Multimodality in Language and Speech Systems, Kluwer Academic Publishers, 173-207.

Traum, D., and Rickel, J., 2002. *Embodied Agents for Multi-party Dialogue in Immersive Virtual World*, in Proc. AAMAS-2002, 766-773.

Wiemann, J., and Knapp, M., 1975. *Turn-taking in conversation*, Journal of Communication, 25, 75-92.

Appendix A. Sample multi-participant interaction segment



B.

t_1 S: [alternating gaze between to p_{15} and p_{14}] ***I'm sorry but that's incorrect! The correct*** [t_4] ***answer is Independence Hall*** [t_5]***. The Liberty Bell is housed in a building nearby.*** [t_7] ***Now on to the final*** [t_8] ***question*** [t_9]
t_2 p_{15}: ***independence hall***
t_4 p_{14}: ***oh independence hall …***
p_{15}: ***okay***
t_7 p_{14}: ***oh I thought that … okay.***
t_9 p_{14}: ***sorry***
p_{15}: ***that's okay***
t_{11} S: [gazing at p_{14}] ***How long does it take for light produced by the sun to reach the Earth?***
t_{13} p_{14}: ***six minutes***
t_{14} S: switches gaze towards p_{15}, lifts eyebrows
t_{15} p_{15}: ***six minutes***
t_{16} S: [still gazing at p_{15}] ***You think it's 6 minutes,*** [t_{17}] ***right?***
t_{17} p_{14}: ***yes***
p_{15}: ***yes***
p_{14}: ***correct***
t_{19} S: ***That's correct!***

Shortly before time t_5: system has floor (indicated by the yellow bar at the top of the system box), is speaking (indicated by speaker sign to the right of the system box) to both p_{14} and p_{15} (indicated by small arrows). The system is gazing at p_{15} (indicated by red dot) p_{14}'s attention is towards p_{15} (indicated by arrow in the center of face). p_{15} is also speaking and therefore performing a *Take* floor action.

Shortly after time t_{10}: p_{14} has floor and is releasing it to p_{15} (indicated by the green arrows pointing down at the bottom of p_{14}'s face box, and orange arrows pointing up at the bottom of p_{15}'s face). System gazes to the participant that the floor is being released to – p_{15} (indicated by the red dot)

Figure 6. Logs and images from a sample interaction collected in the multiparty turn-taking user study.

At the start of this segment the system is providing an explanation about the previous incorrect answer (t_1 to t_7). While the system is speaking, participants p_{14} and p_{15} are speaking with each other. The system detects three utterances (track *b*, time t_2 to t_3, t_4 to t_5, and t_7 to t_8). However, because its floor management policy is configured to disallow barge-ins during explanations, the system continues to perform a *Hold* action throughout this time (track *f*). The floor therefore remains with the system (track *c*, and Figure 6.C). Note that, even though the system rejected the barge-ins, the corresponding semantic inputs are still received and processed by the dialog manager. In this case, all of these inputs resulted in non-understandings, and the dialog state was not affected. Also note that, throughout the period t_4-t_5 the *Take* action alternates between p_{14} and p_{15}. This occurs because the models for inferring the speech source alternate between the two participants while the utterance is being produced; in fact, while the voiced activity detector identified a single utterance from t_4 to t_5, each participant produced an utterance in this time span (p_{14}:*"oh independence hall"* and p_{15}:*"okay"*). The example highlights the need for more accurate signal segmentation.

At time t_9, the system finishes the utterance *"Now on to the final question"* and begins to ask a new question. The system's floor management policy is configured so as to allow barge-ins when it is challenging users with questions. Thus, as soon as another user utterance and corresponding *Take* action is detected (track *b*, right after t_9), the system performs a *Release*, which stops the system's utterance and eventually shifts the floor to p_{14}. Immediately following t_9 the floor again alternates between p_{14} and p_{15} while the system is uncertain about the speech source (tracks *c*,*d*,*e*,*f*,*g*,*h*.) The participant's utterance leads to another non-understanding, and the system infers that it was spoken by p_{14} towards p_{15}; in fact this segment also contains two separate concatenated utterances (Figure 6.A). At t_{10} the system therefore infers that p_{14} is releasing the floor to p_{15} (see track *g*, and Figure 6.D), and, as soon as it detects this release, it switches gaze to p_{15} (track *k* and Figure 6.D). By default, in this situation (*i.e.* a participant releases the floor to another participant) the system waits for 3.5 seconds with no one taking the floor, before it would generate a new contribution signal, and therefore attempt to take the floor. However, as soon as it receives this input, as the current question was interrupted and still needs to be produced, the dialog manager configures the threshold duration in the floor management policy to only 0.5 seconds. This reflects the fact that the system would like to leverage any small break in the side conversation to insert its contribution (next question) and continue. At t_{11} therefore (0.5 seconds later), since no one took the floor, a new contribution signal is generated, the dialog manager re-issues the same output, and the behavioral layer successfully takes the floor and restarts the prompt. Note however that, if between t_{10} and t_{11} one of the participants had continued the side conversation, the system would have kept monitoring their conversation until either the floor was released to it, or 0.5 seconds had elapsed with the floor released to another participant but not taken by that participant. Alternatively, if a second interruption had occurred after the system started re-speaking its question at t_{11}, the dialog manager would have configured a longer threshold duration. As a result, if the system was interrupted a second time by an utterance from one participant to another, it would have waited for a longer silence before issuing a new contribution. If however the interruption was instead created by an utterance addressed to the system, such as an answer to the question (hence followed by a floor release to the system), the next system contribution would have been generated immediately. Finally, if the duration of the break in the system's prompt (*e.g.* from t_9 to t_{11} in this case) had been below 0.5 seconds, the behavioral layer would have continued the system's prompt from where it left off, producing a continuation instead of a restart.

At time t_{13}, p_{14} responds. The system switches gaze towards p_{15} and lifts eyebrows at t_{14}. p_{15} immediately also responds (this is another example of a gesture-based, non-verbal confirmation and floor release). Since the confidence score on this last utterance is low, the system confirms again the answer with p_{15} at time t_{16}. While the system is speaking the confirmation question, another utterance and corresponding *Take* action is detected (at t_{17}, from p_{14}). The system does not allow the barge-in, so it finishes its utterance at t_{18}, and only then releases the floor. The *Take* action by p_{14} therefore persists between t_{17} and t_{18} (track *g*), and the floor switches to p_{14} only at time t_{18}. Afterwards, the system informs the participants they selected the correct answer and moves on.

Appendix B. Open-ended comments in survey results

The table below contains a sampling of comments from users on two open-ended questions in the survey. We manually clustered the comments into several classes, as shown in the Table below. The large majority of comments in the "liked best" class fall into 2 broad categories: Turn-taking includes comments related to the system's multiparty interaction and turn-taking capabilities, and Questions Game includes comments related to the (educational) nature of the trivia game. For the class "would change" most participants focused on the audio and/or visual rendering of the avatar.

Category	#	Example comment
Please describe what you liked best about interacting with the system		
Turn-taking	17	- I think the avatar did a great job, looking at the person who was speaking. Sometimes she would just look at someone, raise an eyebrow to confirm an answer w/o speaking - The eyes followed the person talking so you really get the feeling that someone is talking with you instead of at you. - Being able to say my answer, think out loud with the others - The movement of the head when it addressed someone
Questions game	17	- I enjoyed trivia so I thought it was fun. I also liked the questions asked - I liked getting the information about the specific answer for the question. It did not just simply say the answer, it gave us more detailed info. I learned some new information today. - It's a great fun way to improve knowledge
Speech Recognition	5	- Voice recognition was fairly accurate, no need to repeat - The voice recognition was impressive. It understood much more than "yes" and "no". I could speak very naturally and it would understand
Other	17	- Greeting w/o having to push buttons. I liked that the experience just "began" w/o prompting - It was really simple to use - I enjoyed the avatar and how she interacted with us as participants
If there was one thing you could change about this system, what would it be?		
Avatar rendering	31	- The avatar should be a little less serious and more friendly. - Make the computer voice seem more real. - Facial expressions. I felt like there are a multitude of expressions that can be used to further the game such as frustrated look when you take too long
Turn-taking	7	- It was a bit sensitive in being able to pick up our "mumbling". We couldn't really talk to each other w/o the avatar trying to pick up an answer. - When in groups of 3, make it more clear about who is supposed to answer or confirm the question. - A few times it thought we had agreed on an answer and went ahead in the game before I was ready.
Speech Recognition	7	- Fine-tune the recognition so that it understands better. - What I would change is the error rate in recording the answer. Sometimes consensus was made in the group but the system recorded a different response. [...]
Other	11	- I would have the game ask harder questions - If I could change something, I would have it detect body and face expressions (nodding).

Table 3. Sample comments to open-ended questions in post-experiment survey