Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions

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 report on an empirical study of a multiparty turn-taking model for physically situated spoken dialog systems. We present subjective and objective performance measures that show how the model, supported with a basic set of sensory competencies and turn-taking policies, can enable interactions with multiple participants in a collaborative task setting. The analysis brings to the fore several phenomena and frames challenges for managing multiparty turn taking in physically situated interaction.

1. Introduction

Effective dialog relies on the coordination of contributions by participants in a conversation via turn taking. The complexity of understanding and managing turns grows significantly in moving from dyadic to multiparty settings, including situations where groups of people converse as they collaborate on shared goals. We are exploring computational methods that can endow dialog systems with the ability to participate in a natural, fluid manner in conversations involving several people.

In Bohus and Horvitz (2010a), we presented a computational model for managing multiparty turn taking. The model harnesses multisensory perception and reasoning and includes a set of components and representations. These include methods for tracking multiparty conversational dynamics, for making turn-taking decisions, and for rendering decisions about turns into an appropriate set of low-level, coordinated gaze, gesture and speech behaviors. We implemented the model and have been testing it in several domains. The investigations have been aimed at characterizing the system's performance in complex multiparty settings.

In Bohus and Horvitz (2010b), we examine data collected during a user study to evaluate the ability of the system to shape the flow of multiparty conversational dynamics. In this paper, we focus our attention on the performance of the inference and decision-making models. We analyze the accuracy of current turn-taking inferences, the influence of inference errors on decisions, and the overall effectiveness of the system's decision making. We report on subjective and objective measures of the system's turn-taking performance. We find that the turn-taking methodology enables our system to successfully participate in multiparty interactions, even when relying on relatively coarse models for inference and decision making. The analysis highlights several general phenomena including standing bottlenecks and difficulties, and opportunities for enhancing multiparty turn taking in dialog systems. Based on the results, we discuss challenges and directions for research on turn taking in physically situated dialog.

2. Related Work

We begin by placing this work within the larger context of research on multiparty interaction and turn taking. In a seminal paper on turn taking in natural conversations, Sacks, Schegloff and Jefferson (1974) proposed a basic model for the organi-

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 98–109,
Portland, Oregon, June 17-18, 2011.

zation of turns in conversation. The model is centered on the notion of *turn-constructional-units*, separated by *transition relevance places* that provide opportunities for speaker changes. In later work, Schegloff (2000) elaborates on several aspects of this model, including interruptions and overlap resolution devices. Other researchers in conversational analysis and psycho-linguistics have highlighted the important role played by gaze, gesture, and other non-verbal communication channels in regulating turn taking. For instance, Duncan (1972) discusses the role of non-verbal signals, and proposes that turn taking is mediated via a set of verbal and non-verbal cues. Wiemann and Knapp (1975) survey prior investigations on turn-taking cues in several conversational settings, in an effort to elucidate differences. Goodwin (1980) discusses various aspects of the relationship between turn taking and attention. More recently, Hjalmarsson (2011) investigates the additive effect turn-taking cues have on listeners in both human and synthetic voices.

Within the dialog systems community, efforts have been made on designing and implementing computational models for managing turn taking (e.g., Traum, 1994; Thorrissön, 2002; Raux and Eskenazi, 2009; Selfridge and Heeman, 2010). Moving beyond the dyadic setting, Traum and Rickel (2002) describe a turn management component for supporting dialog between a trainee and multiple virtual humans. Kronlid (2006) describes a Harel state-chart implementation of the original SSJ model. Researchers studying human-robot interaction have developed prototype robots that can interact with multiple human participants (e.g. Matsusaka et al., 2001; Bennewitz et al., 2005). In our previous work Bohus and Horvitz (2009; 2010a; 2010b), we describe a platform that leverages multimodal perception and reasoning to support multiparty dialog in open-world settings.

3. Multiparty Turn-Taking Model

We engaged in a set of experiments to probe the inference and decision making competencies of a computational model for multiparty turn taking (Bohus and Horvitz 2010a; 2010b). To set the stage for the analysis to follow, we briefly review the proposed approach.

We model turn taking as an interactive, collaborative process by which participants in a conversation monitor one another and take coordinated actions to ensure that (generally) only one person speaks at a given time. The participant ratified to speak via this process is said to have the *floor*. Each participant engaged in the interaction continuously produces (*i.e.* at every time tick) one of four *floor management actions*: a *hold* action indicates that a participant is maintaining the floor; a *release* action indicates that the participant is yielding the floor to a set of other participants (which could be void, allowing for self-selection next turn allocation); a *take* action indicates that the participant is trying to acquire the floor; finally, a *null* action indicates that a participant is not making any floor claims. The floor shifts from one participant to another as the result of the joint, cooperative floor management actions taken by the participants. Specifically, a *release* action must be met with a *take* action for a floor shift to occur; in all other cases the floor stays with the participant that currently holds it.

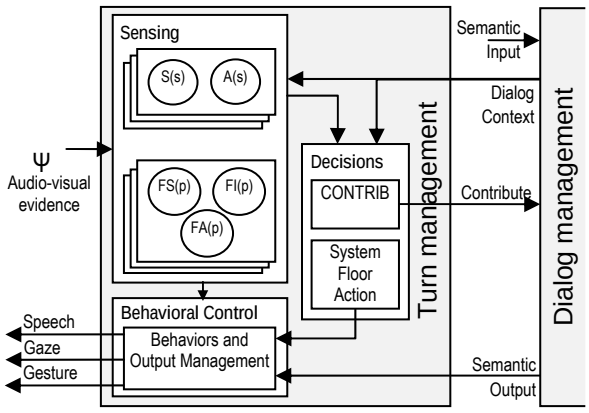


Figure 1. Components of turn-taking model.

Figure 1 illustrates the main components and key abstractions in the model. The sensing subcomponent tracks the conversational dynamics, and includes models for detecting spoken signals *s*, inferring the source *S(s)* and the set of addressees *A(s)* for each signal, as well as the floor state *FS(p)*, actions *FA(p)* and intentions *FI(p)* of each participant *p* engaged in a conversation. This information is used in conjunction with higher-level dialog context to decide when the system should generate new contributions and which floor action should be produced at each point in time. Finally, floor actions are rendered by a behavioral component into a set of coordinated gaze, gesture and speech behaviors. By harnessing these different components, the proposed model can enable an

embodied conversational agent to handle a broad spectrum of turn-taking phenomena.

4. User Study

We implemented an initial set of turn-taking inference and decision making models in the context of a multiparty dialog system, and we conducted a large-scale multiparty interaction user study with this system. The study, described in more detail below, was designed to fulfill two goals: (1) to ascertain an initial performance baseline and identify current bottlenecks and challenges to be addressed moving forward, and (2) to collect a large set of multiparty human-computer dialog data that can be used to study and improve multiparty turn taking in dialog systems.

4.1. System

The platform used in these experiments, described in detail in Bohus and Horvitz (2009), takes the form of a multimodal interactive kiosk that displays an avatar head which plays a questions game with multiple participants. The system leverages audiovisual information and employs components for visually tracking multiple people in the scene, sound source localization, speech recognition, conversational scene analysis, behavioral control and dialog management. Figure 2 shows a screen generated by the system, with the rendered avatar and a sample challenge question. Users can collaborate on selecting an answer, and, after a confirmation, the system provides an explanation if the answer is incorrect, before moving on to the next question. Sample interactions are found in Appendix C and videos are available online (Situated Interaction, 2011).

4.2. Turn-Taking Inference and Decisions

In the current system, a voice activity detector is used to identify and segment spoken utterances. The source of each utterance is assumed to be the participant who is closest in the horizontal plane to the sound direction identified by the microphone array. The set of addressees is identified by fusing information probabilistically about the focus of attention of the source, as obtained through face detection and head pose tracking, while the utterance is being detected. In addition, the system assumes that non-understandings are addressed to other engaged participants, since initial tests indicated that in this domain about 80% of utterances that led to non-understandings were in fact addressed to others. Similarly, the system assumes that utterances longer than three seconds are addressed to others (responses addressed to the system tend to be short in this domain)

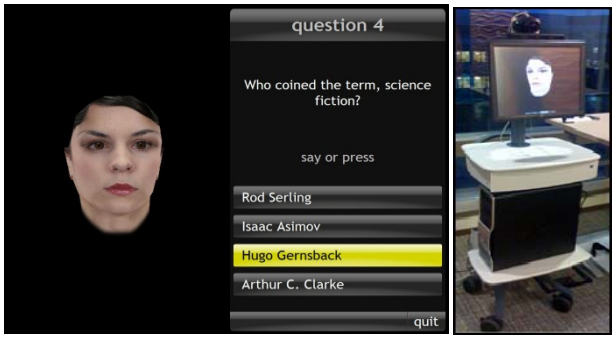


Figure 2. Questions game: screen and kiosk.

Floor management actions are inferred as follows. If a participant has the floor, we assume they are performing a hold action if speaking and a release action otherwise. The release is assumed to be towards the addressees of the last spoken utterance. Although the latter assumption on releases may not hold in the most general case, it is a reasonable one for the questions game domain. If a participant does not have the floor, the system assumes they perform a take action if speaking or a null action otherwise. The system also assumes that the floor intentions are fully reflected by the floor actions, i.e., a participant intends to have the floor if and only if she performs a hold or take action. Floor states are updated based on the joint, coordinated floor actions of all participants, as described earlier.

Turn-taking decisions are based on a simple heuristic policy. The system takes the floor if (1) the floor is being released to it or (2) a participant releases the floor to someone else, but no one claims the floor for a preset duration. In most cases, this duration is set to 3.5 seconds. However, if the floor is released to someone else after the system is interrupted during a *question* dialog act, the system will try to quickly reacquire the floor should no one else be speaking, so as to finish or restate its question. The waiting duration is set in the latter case to 500 milliseconds. If after 500ms, when the system tries to take the floor another conflict occurs (followed by a floor release to someone else), the waiting duration is increased again to 3.5 seconds. Finally, if a third consecutive conflict oc-

curs when the system tries to acquire the floor, the waiting duration is set to a longer, 20 seconds.

The system releases the floor at the end of its own outputs. In addition, it has to decide whether it should release the floor when a user performs a take action (i.e. barges in) while the system is speaking. The heuristic policy currently implemented by the system releases the floor only for barge-ins occurring during question dialog acts.

Finally, the behavioral models employ policies informed by the existing literature on the role of gaze in regulating turn taking. In particular, the system's gaze is directed towards the speaking participant, or, if the system is speaking, towards the addressees of the system's utterance. During silences, the system's gaze is directed towards the participants that the floor is being released to.

The models and policies described above represent a starting point for inference and action, constructed to enable data collection and an initial evaluation in this domain. We are working to update the turn-taking architecture with more sophisticated evidential reasoning and utility-theoretic decision making. Nevertheless, when harnessed as an ensemble within the turn-taking approach that we have described, the current procedures provide for an array of complex, multiparty turn-taking behaviors. For instance, the system can address each participant individually or all participants as a group via controlling the orientation of its head pose. When participants talk amongst themselves, the system can monitor their exchanges and wait until the floor is being released back to it. If an answer is heard during such a side conversation (e.g., one participant suggests an answer to another), the system highlights it on the screen (see Figure 2). If a significant pause is detected during this side conversation, the avatar takes the floor and the initiative, e.g., *"So, what do you think is the correct answer?"* Once a participant provides an answer, the system seeks confirmation from another participant before moving on. In some cases, the avatar passes back the floor and seeks confirmation non-verbally, by simply turning towards another participant and raising its eyebrows. The system can try to require the floor immediately after being interrupted, but can also back off, giving the participants a chance to finish a side conversation, if successive floor conflicts occur. Sample interactions can be viewed in Appendix C and online (Situated Interaction, 2011).

4.3. Study Design

The user study was conducted in a usability lab and involved a total of 60 participants recruited as pairs of people from the general population who previously knew one another (30 male and 30 female, with ages between 18 and 61). The study was structured in 15 one-hour sessions, with each session involving four participants, i.e., two pairs of two previously acquainted participants. In each session, we formed all possible subgroups of size two (6 subgroups) and of size three (4 subgroups) with the four participants. Each subgroup played one game with the system. This setup allowed us to collect a large set of multiparty interactions under diverse conditions (e.g., all-male, all-female, mixed-gender groups; groups where people were previously acquainted *vs.* not, etc.). At the end of each session, participants filled in a subjective assessment survey.

4.4. Corpus, Annotations, and Cost Assessment

In total, 150 multiparty interactions were collected: 90 with two participants and the system, and 60 with three participants and the system. A professional annotator transcribed the utterances detected by the system at runtime, and labeled them with *source* and *addressee* information.

The system was noted to commit several types of turn-taking errors. To expand the error analysis beyond occurrence statistics and to characterize the impact of various types of errors, we conducted a follow-up study. In this second study, a set of additional participants were recruited to review videos of interactions from the first study and asked to (1) identify the turn-taking errors committed by the system and (2) to assess the costliness of the error on a five-point scale.

A total of 9 interactions (5 with two participants and system; 4 with three participants and system) were randomly sampled from the collected corpus, while ensuring that each turn-taking outcome of interest (discussed in Section 5 and summarized in Table 1) was sufficiently represented. Nine participants were recruited via an email request to employees at our organization. Each participant reviewed three interactions, and each interaction was reviewed by three different participants. Prior to the experiment, each of the annotators received a brief review of the turn-taking process in human-human interaction. Next, they used a multimodal

annotation tool that we created to review the interaction videos. As each video played, the annotator pushed a button at each point they believed that the system had committed a turn-taking error. In a second pass, each annotator was asked to review the errors that they had previously identified and to assess the relative cost of the error, on a scale from 0 ("no error") to 5 ("worst error"). In a final step, the authors manually aligned each identified turn-taking error with a turn-taking decision made by the system and its corresponding outcome.

5. Evaluation

We now focus on the various types of turn-taking errors, the outcomes that these errors lead to, and the costs assessed for the outcomes. We begin by focusing on diarization challenges described in Section 5.1. In Sections 5.2 and 5.3, we review the accuracy of the system's turn-taking inferences and decisions, and their corresponding outcomes. Finally, in Section 5.4, we turn our attention to the subjective assessment results obtained via the post-experiment user survey.

Before diving into the details, we note that we eliminated 7 out of the total 150 interactions from the analysis due to significant problems with acoustic echo cancellation. In the remaining 143 interactions, we also identified and eliminated 24 utterances in the transitional engagement stages, e.g., when the users were not ready or properly setup in front of the system. The analysis below is based on the remaining 4379 utterances.

5.1. Diarization

The system uses a voice activity detector which leverages energy, acoustics and grammar to detect spoken utterances. Our experiments indicate that this type of black-box solution can make diarization errors, especially in multiparty settings where people may speak simultaneously, at a fast pace, and address each other with language outside the system's grammar. Results show that only 72% of the detected segments contain speech from a single participant. Another 2% contain background noises incorrectly identified as speech. Most often these are instances where the system heard itself due to acoustic echo-cancellation problems; the ratio grows to about 6% among all utterances detected while the system is speaking. The remaining 26% contain overlapping or successive utterances from multiple speakers. Inspection of the data reveals that some utterances spoken softly by participants were not detected and that segmentation boundary errors are also sometimes present. While such errors may be mitigated by inferences at higher levels in the turn-taking model, they can significantly influence the system's ability to track the conversational dynamics and make appropriate turn-taking decisions. We plan to pursue more robust audiovisual diarization methods that integrate sound localization as detected by a microphone array, along with higher-level interaction context.

5.2. *Take* versus *Null*

We now turn our attention to the system's floor control decisions. The analysis below is based on the utterances and segmentation *detected by the system at runtime*. We note that a more precise analysis could be conducted with a ground truth segmentation of utterances. Utterances detected by the system can be classified into three categories, based on their relationship to system outputs, as shown in Figure 3: *overlaps*, which start and end during a system's output, *continuers*, which begin during but finish after a system output has ended, and *responses*, which do not overlap anywhere.

With the current policy, the system chooses whether it should take the floor following each detected *continuer* and *response*. The dataset contains a total of 3265 such instances. The system's decision at each of these points hinges on the results of its inferences about the participants' floor actions, and thus of inferences about the addressees of each utterance. Table 1 displays a tabulation of the release actions performed by the participants versus the actions identified by the system. The release actions are determined from labels assigned manually by the professional annotator. Recall that we make an assumption that the release is towards the set of addressees of an utterance. For segments that were labeled as containing multiple utterances, the release is made to the addressee of the last utterance. The last row in Table 1 corresponds to background noises and system speech incorrectly

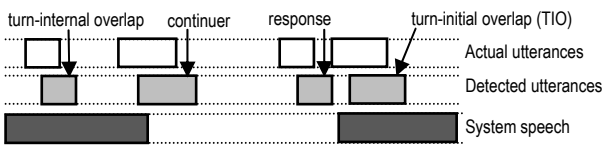


Figure 3. Schematic of different classes of overlap.

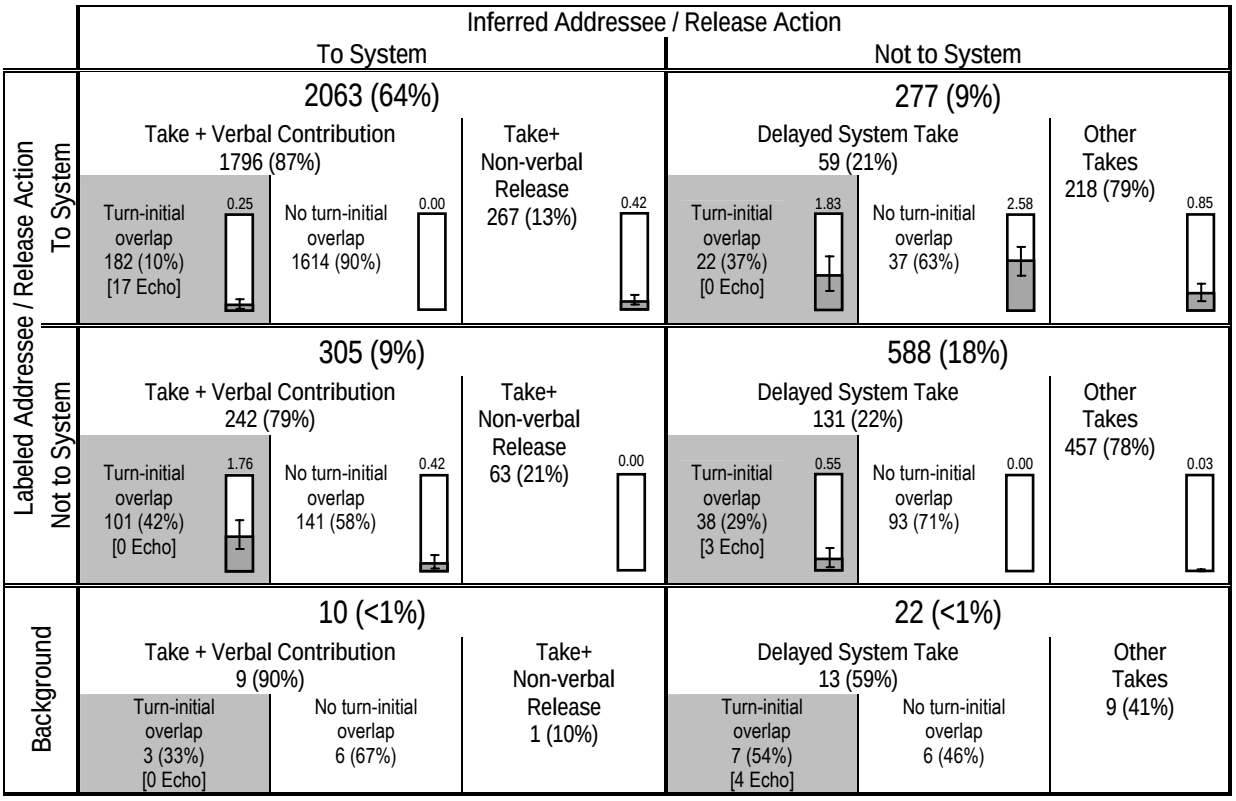


Table 1. Decisions to *take* floor (vs. *null*), outcomes, and estimated costs (bar graph with confidence intervals). *Echo* denotes cases where the turn initial overlap is created by utterances where the system hears itself because of errors with echo cancellation.

identified as utterances.

On the task of detecting addressees, and thus floor release actions, the results show an error rate of 18%, including 305 false-positives (erroneous detections) and 277 false-negatives (missed detections) of floor releases to the system. These errors influence the quality of turn taking in a variety of ways and underscore the need for more robust inferences about speech source and target, and floor release actions. We believe that more sophisticated models learned from audiovisual information (e.g., prosody, head and body pose, etc.) and attributes of the interaction context (e.g., who spoke last, where is the system looking, etc.) can reduce errors significantly.

Table 1 indicates that in 305 (9%) of the cases the system incorrectly inferred that the floor was being released to it. In 79% of these cases, the system took the floor and produced a verbal contribution. Since the floor was not released to the system, such errors can lead to significant turn-taking problems, which often manifest as floor conflicts marked by *turn-initial overlaps*, where a participant and the system start speaking around the same time (see Figure 3). Operationally, we define *turn-initial overlaps* as all detected overlaps with an actual onset of less than 300 milliseconds from the beginning of the system's utterance (see discussion in Appendix A); the other overlaps are dubbed *turn-internal*. We note that the time at which an overlap is *detected* by the system lags behind the actual onset of the utterance by an average of about 700 milliseconds, due to core latencies in our audio and speech processing pipeline. Accounting for these computational lags, and others arising at different places in processing pipelines, raise challenges for turn taking in spoken dialog systems.

42% of the verbal takes performed incorrectly by the system led to turn-initial overlaps. This is not surprising, as the system starts speaking when the floor was not released to it. In some of these cases the same participant continues (e.g., diarization errors incorrectly segmented the utterance), or someone else starts speaking. The cost assessment experiment confirmed the impact of these errors – the average estimated cost was 1.76. If no turn-initial overlap occurred after the system incorrectly took the floor, the average cost was 0.42. Clearly

floor conflicts come with a cost. The specific cost assessments we obtained are perhaps influenced to a degree by the role of *game mediator* played by the system. With this role, taking the floor in cases when the system was not addressed is perhaps not as costly as it might be in other domains.

Note that 182 turn-initial overlaps also occur when the system takes the floor after correctly identifying that the floor was released to it (upper-left quadrant in Table 1). 17 of them are created by the system hearing itself as it starts speaking, due to errors in acoustic echo cancellation; these instances are marked *Echo* in Table 1. While the relative percentage of turn-initial overlaps is smaller after a floor release to the system (~10%), the majority of all turn-initial overlaps (shaded cells in Table 1) occur in this context, because of the larger incidence of the situation. Often, these utterances contain an immediate answer or a short confirmation from another participant. The cost of these turn-initial overlaps is also much lower: 0.25 versus 1.76 (again, the cost structure is probably sensitive to details of the domain).

We believe the turn-initial overlaps that occur when the floor is released to the system can be explained in part by the interpretation of the system's short delay in responding (per processing) as a signal that the system is not taking the floor, leading other participants to take initiative. As another factor, turn taking is a mixed-initiative process, and other participants might vie for the floor and issue their own contributions immediately after an answer directed to the system. These observations bring to the fore two questions: (1) how can we minimize the number of turn-initial overlaps, and (2) how can the system gracefully handle such overlaps once they occur?

One approach to minimizing turn-initial overlaps is to reduce the system's response delays via faster processing or via the use of predictive models to anticipate the end of turns (e.g. Ferrer et al., 2003; Schlangen, 2006; Raux and Eskenazi, 2008; Skantze and Schlangen, 2009). Multiparty settings require methods for forecasting not only when a current speaker will finish, but also whether any participant will try to take (or release) the floor within a small window of time in the future, i.e., accurately modeling all floor intentions. Our turn-taking framework includes components for representing and modeling floor intentions, but these are not used in the current system. We believe there is promise in learning models to predict floor intentions and the timing of ends of utterances from interaction data. The availability of such predictions can fuel additional turn-taking strategies and also pave the way to more graceful handling of turn-initial overlaps after they occur. For instance, if the system can anticipate that someone else might start speaking, it might still decide to take the floor but it might start with a filler, e.g., *"So* [pause] *What do you think?"* constructing a natural opportunity for resolving a potential conflict after *"So"* We plan to investigate the use of decision-theoretic methods to anticipate and resolve such conflicts by introducing and modulating an array of strategies, including the use of fillers, restarts, and acknowledgment gestures.

In 21% of the 305 incorrectly detected floor releases to the system, our system immediately performed a non-verbal floor release to another participant by turning the avatar's face towards them and raising its eyebrows (Take + Non-verbal Release in Table 1). These situations are not costly, as the system's action does not interrupt the flow of the conversation. Indeed they were never penalized in the cost assessment experiment that we conducted. However, the same action, performed when the floor is actually released to the system (13% of 2063 cases), has the potential to create problems if not properly recognized by the targeted participant as a floor release by the system; the average cost assessed in this case was 0.42.

The right-hand column in Table 1 shows cases where the system detected that the floor was not released to it. In these cases, the system waits (performs null) for a specified duration. The cost assessment indicates that waiting in this situation is overall costly, and the cost depends on the ultimate outcome. If no one else takes the floor, the system will eventually do so (Delayed System Take cases in Table 1). In some of these cases, turn-initial overlaps also occur. The 277 cases in which the system fails to detect that the floor was in fact released to it lead to no immediate response from the system. In these cases the system can be perceived as unresponsive and the participants eventually repeat themselves. We believe that performance can be improved with the use of an ongoing decision-theoretic analysis that continuously reassesses the situation while the system waits. Such an analysis would consider the delay, floor holder's previous actions, inferences about participants' floor inten-

tions, and cost-benefit tradeoffs of different floor actions.

5.3. *Release* versus *Hold*

We now turn our attention to the system's decisions to release the floor. Recall that, according to the current policy, the system performs a floor hold while it is speaking and a floor release at the end of its outputs. In addition, if an overlap (i.e., barge-in) was detected during question dialog acts, the system performed a floor release immediately, interrupting its own output and allowing for the user barge-in.

Since such barge-ins were allowed only during the question dialog acts, as Table 2 shows, the current policy leads to an abundance of cases in which the system performs hold when an overlap is detected. Some of these cases are continuers: the overlap only happens at the very end of the system's output. These cases do not create significant turn-taking problems, as the floor still transitions to the participant relatively quickly (the system releases at the end of its output). However, in a significant number of cases the system appears to ignore the participants (shaded cells in Table 2). About three quarters of these overlaps occur while the system is providing an explanation after an incorrect answer. Observations of the data indicate that in these cases participants may discuss or give their opinion on the answer or some aspect of the system's explanation, while ignoring the system as it blindly continues the explanation.

We have separated in Table 2 turn-initial from turn-internal overlaps. The two types of overlaps reflect different phenomena. As we have discussed, turn-initial overlaps mark floor conflicts, and various strategies could be used to negotiate such conflicts (e.g., Yang and Heeman, 2010). In contrast, turn-internal overlaps may reflect efforts by other participants to take the floor, or might simply be backchannels, laughter, exclamations or other lexical or non-lexical events that do not mark a claim for the floor. Making appropriate floor control decisions in this case will require models for reliably distinguishing between the two, i.e., between the take or null floor actions of the participants. This is an especially challenging inference problem as decisions need to be made as early as possible after the onset of an utterance.

We note the relatively large incidence of failures in echo cancellation in our microphone array. On the utterances marked *Echo* in Table 2, the system heard itself and thought a user was speaking. We believe these failures could be significantly reduced with better acoustic echo cancellation.

		Action performed by system when overlap detected		
		HOLD		RELEASE
Overlap Type	Turn Initial	315 (23%)		43 (3%) [7 Echo]
		Overlap 285 (90%) [14 Echo]	Continuer 30 (10%) [3 Echo]	
	Turn Internal	968 (69%)		73 (5%) [13 Echo]
		Overlap 828 (86%) [44 Echo]	Continuer 140 (14%) [7 Echo]	

Table 2. Decisions to *release* floor (vs. *hold*).

5.4. Subjective Assessment

Finally, we present results from a subjective assessment of the system by participants, based on a post-experiment survey. The survey included several 7-point Likert scale questions related to turn taking, which are displayed in Figure 4, together with the mean user responses and the corresponding 95% confidence intervals. Generally, participants rated the system's turn-taking abilities favorably, with scores around 4.5-5. No statistically significant differences were detected in assessments across the participant's gender or previous familiarity with speech recognition systems. We also note that a parallel human—human interaction study would help us characterize better the system's performance relative to human dialog.

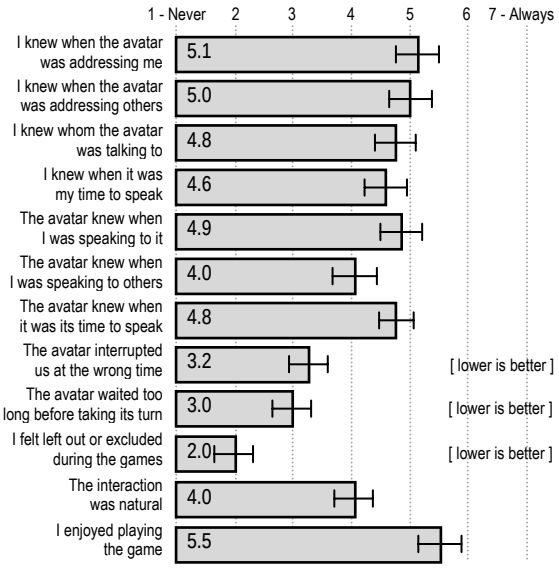


Figure 4. Results of subjective assessments.

In addition to the survey questions, participants were invited to describe in their own words what they liked best and the first thing they would change about the system. 21 of the 60 participants mentioned aspects of multiparty interaction in the "what I liked best" category, such as the system's ability to track the speaking participant and address people individually. Other frequent answers to this question called out the overall experience with the integrative intelligence of the system (15 answers), the fun/educational nature of the game (14), and aspects of speech recognition (11). On the "first thing you would change," the majority of answers (32) included references to shortcomings in rendering the avatar, while 13 answers included references to problematic aspects of the multiparty turn taking. Other answers included task domain suggestions (6) and comments about improving the speech recognition (5). A sampling of answers is presented in Appendix B.

6. Summary and Future Work

We reported on a user study of a multiparty turn-taking model. Objective measures of system performance and subjective assessments by participants indicate that the approach can enable successful multiparty turn taking in the questions game domain. When the correct turn-taking decisions are made, the multiparty interaction is seamless and resembles human-human collaboration. The conversations exhibit fluid exchanges among people and the system, including mixed-initiative, multiparty floor control, fluid back offs and restarts, natural use of non-verbal cues, such as participants' utterances being triggered by a turn of the avatar's head or a lift of the eyebrows. In contrast, turn-taking failures lead to a striking loss of fluidity and a qualitative jump out of an engaged process, where the system rapidly shifts from a collaborating *participant* into a distant and uncoordinated *appliance*.

The results we have discussed are based on an initial set of coarse perceptual and decision-making models and thus reflect an initial baseline; there is significant room for improvements. A careful dissection of the outcomes demonstrates the subtleties of multiparty turn taking and highlights several directions we plan to address in future work. First, our experiments have highlighted the importance of accurate diarization in multiparty dialog settings. Minimizing errors requires rich perceptual and inferential competencies, leveraging audiovisual evidence, general patterns of human discourse, and attributes of the task-specific goals and context. We plan to explore the use of machine learning procedures for constructing predictive models that harness richer streams of evidence to identify and segment utterances, and to make inferences about their sources and targets, and the floor state, actions and intentions of all participants. Better turn-taking decisions can also be supported by inferences about social norms, roles and dynamics, pace of interaction, and engagement.

Although handcrafted turn-taking policies went a long way in this domain, enabling more general multiparty turn taking will require continuous inference and decision making under uncertainty that considers subtleties of intention and timing, and that takes into consideration tradeoffs associated with different courses of actions. We foresee the value of extending the current decision models with richer temporal reasoning for performing such ongoing analyses. Challenges include a more in-depth understanding of the cost of different types of turn-taking errors; the development of a wider array of graded strategies and behaviors for taking, releasing, or holding the floor, and for gracefully negotiating floor conflicts; and finally, the ability to reason about uncertainty in the world as well as in the system's own processing delays in order to resolve tradeoffs between taking timely action and delaying for additional evidence that promises to enhance the accuracies of decisions.

Much also remains to be done with the corresponding generation of subtle verbal and non-verbal cues for enhanced signaling and naturalness of conversation, including the use of fillers, restarts, backchannels, and envelope feedback. We are excited about tackling these and other challenges on the path to fielding systems that can engage in fluid multiparty dialog.

Acknowledgments

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

References

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S., 2005. Integrating vision and speech for Conversations with Multiple Persons, in *Proc. of IROS'05*

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

Bohus, D., and Horvitz, E., 2010a. Computational Models for Multiparty Turn Taking, Microsoft Research Technical Report MSR-TR 2010-115.

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

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'03*.

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

Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice, *in Speech Communication*, vol. 53, issue 1.

Kronlid, F., 2006. Turn Taking for Artificial Conversational Agents, *in Cooperative Information Agents X*, LNAI 4149, Springer-Verlag

Matsusaka, Y., Fujie, S., and Kobayashi, T., 2001. Modeling of conversational strategy for the robot participating in the group conversation, in *Proc of EuroSpeech'01*.

Raux, A., and Eskenazi, M. 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system, *in Proc of SIGdial-2008*.

Raux, A. and Eskenazi, M., 2009. A Finite-State Turn-Taking Model for Spoken Dialog Systems, in *Proc. HLT'09*.

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

Schegloff, E. 2000. 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'06, Panel on Prosody of Dialogue Acts and Turn-Taking*

Selfridge, E., and Heeman, P., 2010. Importance-Driven Turn-Bidding for Spoken Dialogue Systems, *in Proc. of ACL-2010*, Uppsala, Sweden

Skantze, G., and Schlangen, D., 2009. Incremental dialogue processing in a micro-domain, *in Proc. of EACL-2009*.

Situated Interaction, 2011. Project web page: http://research.microsoft.com/~dbohus/si.html

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.

Traum, D., 1994. *A Computational Theory of Grounding in Natural Language Conversation*, TR-545, U. of Rochester.

Traum, D., and Rickel, J., 2002. Embodied Agents for Multi-party Dialogue in Immersive Virtual World, in *Proc. AAMAS'02*.

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

Yang, F., and Heeman, P., 2010. Initiative Conflicts in Task-Oriented Dialogue, *in Computer, Speech and Language,* vol. 24, issue 2.

Appendix A. Details on derivation of operational definition of turn-initial overlaps.

As described in Section 5.2, we operationally define *turn-initial* overlaps as detected user utterances that have an actual onset of less than 0.3 seconds from the beginning of a system utterance. Figure 5 shows the histogram of the onset time for user speech with respect to system utterances (start of system utterance is at 0 seconds), for overlapping utterances, where this onset is between -2 and +5 seconds. If multiple user utterances overlap with a single system utterance, only the first user utterance, i.e. the first overlap, is considered in computing this histogram. As Figure 5 shows, the onset distribution has a bi-modal character. We believe that the two modes may reflect two different phenomena in terms of the floor transition. The early-onset mode corresponds to situations in which a user starts to speak right around (before or immediately after) the time the system also started speaking; this indicates a situation where there is contention for the floor and the system cannot assume it has successfully acquired the floor. In contrast, user utterances starting at later times represent cases where the floor did first transition to the system and the user is aware of this transition. In producing an utterance the user is attempting to barge-in and take the floor back from the system (unless the user utterance is a backchannel). The threshold of 0.3 seconds on the onset for turn-initial overlaps was selected based on the shape of this distribution.

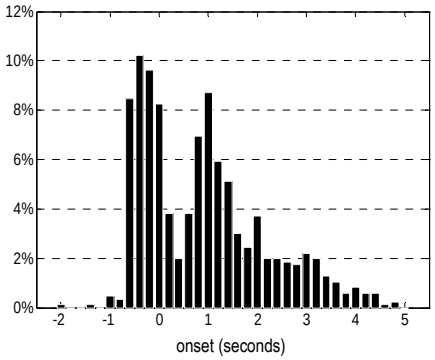


Figure 5. Histogram of onsets for first overlaps.

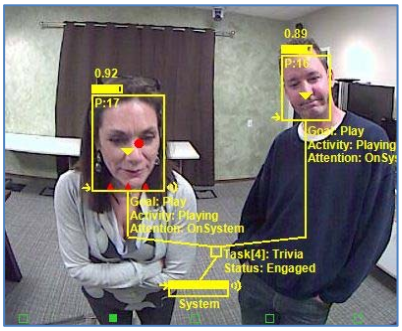
Appendix B. Sample responses from survey

Category	#	Example comment
Please describe what you liked best about interacting with the system		
Multiparty interaction prowess	21	- I enjoyed how it recognized who was speaking and actually looked at you - I liked how the avatar tracked the players; how it understood speech - It was great to play a game where you don't have to use your hands, just your mind. The way the avatar would recognize position of who spoke was nice. The blinking action at the avatar made her more realistic but she needed more than her face. - That it would look right at you and ask a question - I liked how the avatar made eye contact with each person playing the game
Overall experience with system	15	- It was very new and thus it was fun. I don't play computer games often and I did enjoy this one. Which is rare for me. - It was different than any other trivia game I've played in the past - I think this is a great way for a human to interact with a computer - It's cool interacting with the avatar
Rewarding task	14	- I liked the challenge of the questions - It's a great fun way to improve knowledge - New experience that I found enjoyable. I enjoyed thinking about choices and having an interaction with the avatar
Speech and language	11	- Voice recognition was fairly accurate, no need to repeat - The ability of it to understand what I was saying. Plus it's pretty cool. - I liked it because it wasn't really hard for the system to understand what we were saying. Even though we have an accent.
If there was one thing you could change about this system, what would it be?		
Avatar rendering	32	- The avatar should be more friendly – she came off a bit austere – she didn't smile even when we got 5 out of 6 questions right, it was only "pretty good". - The way it moves its lips needs to be better - The avatar seemed a little to "stiff". It needs to be more natural in movement and speech - The face was a "warmer face". Smiling perhaps.
Multiparty failures	13	- Extend the time limit when questions haven't been fully answered. It would sometimes say we were correct or false before we had confirmed our answer - Sometimes it skips and pauses and making it difficult to understand - Consistency in waiting and asking player to confirm answer instead of overhearing conversations and choosing an answer itself
Task domain	6	- It would be cool if it could remember our names. Also, 6 questions was a little short. I think 8 or 10 questions would be better. - I think the questions should be more pop culture related
Speech and language	5	- I enjoyed her. I would like her to understand a little easier. We had to repeat answers on occasion which wasn't too bad. Overall I really liked it. Perhaps it could ask our names and call us by name when speaking to us

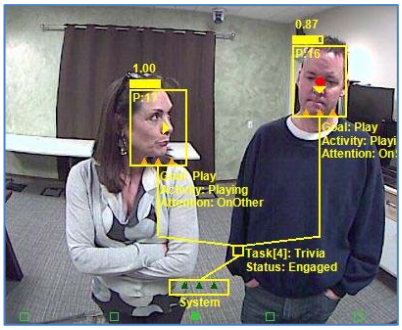
Appendix C. Excerpts from interactions with the system. We present and discuss two segments from an interaction with the questions game system. The segments illustrate challenges for diarization, tracking conversational dynamics (e.g. inferring speech source, target, floor actions, etc.) and decisions making for multiparty turn taking. The video for this entire interaction, as well as an additional interaction are available online at (Situated Interaction, 2011)

1	S→P_1	Hi. Would you like to play a questions game?	Immediately after the system's greeting, the two participants also say *"Hi"* and *"Hello"*. Their greetings are detected as a single utterance by the system which partially overlaps with the beginning of the system's follow-up question (overlaps are underlined in the examples to the left). According to the current policy, the system does not release the floor on this interruption and continues with its question. The *"Yes"* responses from (4) and (5) are overlapping with each other and are detected by the system as a single utterance which is correctly decoded.
2	P_{17}→S	*Hi*	
3	P_{16}→S	*Hello*	
4	P_{17}→S	*Yes*	
5	P_{18}→S	*Yes*	
6	S→P_{16}	Here is the first question. The study of the size and shape of the skull as an indicator of a person's character is known as: telepathy, graphology, or phrenology?	The system correctly infers the floor was addressed to it, and therefore takes the floor and produces the first question
7	P_{16}→S	*I'll go… Graphology*	Given the pause between *"I'll go"* and *"Graphology"* the response in (7) is in fact detected as two separate utterances by the system. The first part is non-understood, hence assumed addressed to other and the system does not take the floor. The system correctly understands and takes the floor after *"Graphology"*, and moves to ask for confirmation (8). Next, while the system asks the other participant for confirmation, due to imperfections in echo cancellation, the system hears a noise at the beginning of its utterance, but ignores the detected "barge-in".
8	S→P_{17}	What do you think?	
9	P_{17}→P_{17}	*graph-*	P_{17} softly says to herself *"graph-"*. This utterance is not actually detected by the system.
10	P_{16}→P_{17}	*That doesn't sound right but*	Next, the system misunderstands the utterance in (10) as "that sounds right" and incorrectly infers that the utterance was addressed to it. It therefore takes the floor and continues. This leads to a turn-initial overlap with the *"Phrenology"* utterance immediately produced by P_{17} (11)
11	P_{17}→S	*Phrenology*	
12	S→P_{16},P_{17}	I'm sorry but that's incorrect. Popular during the 19th century but not discredited, phrenology is a theory that a person's character and intelligence can be determined by the size and shape of their skull.	
13	P_{16}→P_{17}	*phrenology*	The follow-up utterances and discussion between participants (13-16) overlap with portions of the system's explanation. They indicate the high cost of the misunderstanding and of the system's incorrect inference and decision to take the floor (admonished by the user in (16) ❶, as well as the shortcomings of the current policy to not release the floor for barge-ins detected during explanations. This example highlights the need for more robust inferences, but also better policies for releasing back the floor and for machinery that would allow the system to gracefully backing from detected floor conflicts.
14	P_{16}→P_{17}	*She did not wait for your answer*	
15	P_{17}→P_{16}	*I guess not*	
16	P_{16}→S	*You didn't wait for my answer*	
17	S→P_{16},P_{17}	Next question […]	The system then moves on to the next question

1	S→P_{16}	In the USA hospitals have a red cross, what sign do they use in the Arab World?	This segment starts with the system asking a question
2	P_{16}→P_{16}	Hospitals have a red cross what sign do they user in the Arab World…	P_{16} echoes back the question to himself in a low voice. The system detects this utterance, correctly inferred that the floor is not being released to it (the utterance leads to a non-understanding), and the system turns towards the other participant, which produces *"Solid blue circle"* (3). The utterance and corresponding floor release to the system are correctly understood, and the system turns back to P_{16} for confirmation. Next, the system detects an utterance from P_{17}, who has changed her mind (5). The system takes the floor and asks both participants (by directing gaze towards them) what the correct answer is (6), and releases the floor to both participants ❷. P_{17} speaks first towards P_{16} (7) – this utterance is low energy and not detected by the system, and then towards the system (8). The system recognizes the floor is released to it after (8), takes the floor and directs it back towards P_{16} non-verbally, by turning towards him and raising eyebrows. P_{16} responds immediately, and the system takes the floor and continues to the next question.
3	P_{17}→S	Solid blue circle	
4	S→P_{16}	Is that correct?	
5	P_{17}→S	No	
6	S→P_{16},P_{17}	So what's the correct answer?	
7	P_{17}→P_{16}	Red crescent moon	
8	P_{17}→S	Red crescent moon	
9	S→P_{16}	[turns to P_{16}, lifts eyebrows]	
10	P_{16}→S	Yes	
11	S→P_{16},P_{17}	That's right. Okay, so far you're two out of three […]	



❶ In the first segment, while the system is speaking to both participants (12), P_{17} leans in as she produces utterance (16)



❷ In the second segment, the system releases the floor to both participants after producing (6)

Illustrations of conversational scene analysis performed by the system in real-time, at runtime.