Facilitating Multiparty Dialog with Gaze, Gesture, and Speech

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+(01) 425 706 5880
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA, 98052
+(01) 425 706 2127
horvitz@microsoft.com

ABSTRACT
We study how synchronized gaze, gesture and speech rendered by an embodied conversational agent can influence the flow of conversations in multiparty settings. We review a computational framework for turn taking that provides the foundation for tracking and communicating intentions to hold, release, or take control of the conversational floor. We then present details of the implementation of the approach in an embodied conversational agent and describe experiments with the system in a shared task setting. Finally, we discuss results showing how the verbal and non-verbal cues used by the avatar can shape the dynamics of multiparty conversation.

Categories and Subject Descriptors
H.1.2 **[Models and Principles]**: User/Machine System – *Human Information Processing*; H.5.2 **[Information Interfaces and Presentation]** User Interfaces – *Natural Language*; I.4.8 [**Scene Analysis**]: Tracking, Sensor Fusion

General Terms
Algorithms; Human Factors

Keywords
Multiparty turn taking; multiparty interaction; floor management; behavioral models; gaze; gesture; speech; spoken dialog; situated interaction; multimodal systems.

1. INTRODUCTION
As the verbal communication channel is fundamentally serial, people engaged in conversation need to coordinate with one another on turn taking. They do this with the presentation and recognition of non-verbal and verbal cues such as establishing or breaking eye contact, head and hand gestures, changes in speech prosody, and verbal affirmations [7,8,13,14,15,19]. These cues are produced and attended to in an effortless manner, and are used, along with predictions about the natural or intended end of another's turn, to guide the dynamics of an evolving conversation.

ICMI-MLMI'10, November 8-12, 2010, Beijing, China.

The work described in this paper is part of a larger effort aimed at endowing situated spoken dialog systems with the core competencies required to engage in fluid, multiparty turn taking [4]. Key research challenges in this area include: tracking in real-time the conversational dynamics and the floor management actions that control these dynamics, making turn-taking decisions, and rendering these decisions into appropriate behaviors that accurately convey the system's turn-taking intentions and allow it to effectively shape the conversational dynamics.

We focus here on the implementation of a behavioral control model that leverages verbal and non-verbal cues to enable an embodied agent to influence the flow of conversation in multiparty settings. We conduct group interaction experiments in a shared task setting with a system that implements this model, and empirically investigate the degree to which an avatar's gaze, gesture and speech allow it to shape the conversation on a turn-by-turn basis. We also discuss how a number of contextual factors, like dialog act type, previous speaker context, the presence of deictic markers, time elapsed, etc. relate to the system's ability to shape the flow of the interaction.

2. RELATED WORK
The process of turn taking in human-human interaction has attracted attention from researchers in the sociolinguistics and conversational analysis communities. Sacks, Schegloff and Jefferson [13] introduce a model for turn taking centered on a notion of *turn-constructional-units*, separated by *transition-relevant-places* which provide opportunities for speaker changes. Other researchers have highlighted the importance of non-verbal signals such as gaze and gesture in turn taking. For instance, Duncan [7] proposes that turn taking is regulated via both verbal and non-verbal cues and highlights correlations between the direction of the participants' eye gaze and turn taking. Wiemann and Knapp [19] report results from a quantitative analysis of verbal and non-verbal cues across a variety of dyadic settings. Goodwin [8] also investigates various aspects of the relationship between turn taking and attention.

One of the first implementations of a multimodal turn-taking model in a conversational agent was done by Thorisson [17], using a layered architecture with several update loops operating at different speeds. More recently, Raux and Eskenazi [12] describe and perform experiments with a turn-taking model for dyadic interactions based on a six-state non-deterministic finite-state-machine. Moving beyond dyadic interactions, Traum and Rickel

[18] describe a turn management system that supports dialogue with multiple virtual humans in immersive environments.

Given the prominent and multiple roles of gaze, a variety of models for controlling gaze in physically or virtually embodied conversational agents have been previously proposed [1,5,9,10, 11]. Gaze in these systems is frequently driven by a computed notion of "salience" in the scene and is based on the goal of maximizing realism or similarity to human behavior. For instance, Bennewitz et al. [1] propose a model which directs the attention of a robot in multiparty settings based on a measure of interest that takes into account the proximity of a detected face to the robot, and the probability that the face is correctly detected. Others, like Picot et al. [11] and Itti et al. [9], have used image saliency and cognitive attention maps in an effort to produce a biologically motivated models that closely mimic the behavior as well as limitations of the human visual attention system. Cassell [5] has shown that beyond turn taking, other aspects of the interaction, such as information structure correlate with gaze direction. Mutlu [10] has also investigated with a wizard-of-Oz study the degree to which the gaze of a robot can be used to shape engagement and participant roles in multiparty interactions.

Building on these earlier works, we have described and demonstrated in [4] a computational framework for managing multiparty turn taking in situated spoken dialog systems. The framework, which we shall briefly review in Section 3, subsumes models for tracking multiparty conversational dynamics, for making floor control decisions, and for rendering these decisions into appropriate behaviors. We shall focus on the behavioral subcomponent of this model, and empirically show how, by orchestrating gaze with other verbal and non-verbal turn-taking cues, we can enable a conversational agent to effectively participate in multiparty interactions.

3. MULTIPARTY TURN TAKING

3.1 Preliminaries

In the proposed framework [4], we view turn taking as a collaborative, interactive process by which participants in a conversation take coordinated actions to ensure that only one participant (generally) speaks at a given time. The participant who is ratified to speak by means of this collaborative process is said to have the *conversational floor* (henceforth *floor*). Furthermore, the collection of verbal and non-verbal signals that participants use to regulate turn taking are reified into a set of four *floor management actions: Hold, Release, Take* and *Null*. The participant who currently has the floor continuously performs either a *Hold* action, indicating that the participant is in the process of maintaining the floor, or a *Release* action, indicating that the participant is in the process of yielding the floor. The participants that do not have the floor continuously perform either a *Take* action, indicating an attempt to acquire the floor, or a *Null* action, indicating that a participant is simply observing the conversation, without issuing any claims for the floor. Under these assumptions, floor shifts happen as the result of cooperative, joint floor management actions taken by the participants. Specifically, the floor transitions from one participant to another if and only if a *Release* action by one participant is met with a *Take* action by another participant.

This approach to multiparty turn taking framework includes components for sensing conversational dynamics, for making floor control decisions, and for rendering these decisions into appropriate behaviors (Figure 1). In the following two subsections, we briefly review representations and concepts that play a key role in the sensing and decision making components of the proposed framework. Then, in subsection 3.4 we focus in more depth on the behavioral model, as it plays a central role in the experiments and the analysis discussed in the rest of the paper.

3.2 Sensing

The sensing subcomponent in the proposed model is responsible for tracking the conversational dynamics, *i.e.,* detecting spoken signals, inferring the source and targets of each detected signal, as well as the floor state, actions and intentions of each participant engaged in the conversation.

For each detected signal $s \in S$, the source is assumed to be one of the observed actors or background, and is represented by a multinominal variable $SS(s)$. A more complex representation is used for the signal target $ST(s)$ [4], allowing us to capture the participant role each actor in the scene can have with respect to the given utterance, as outlined by Clark and Carlson [6]:

- *Addressee*: a participant engaged in the conversation that the signal is addressed to.
- *Side Participant*: a participant engaged in the conversation the signal is not addressed to.
- *Overhearer*: other actor known to the speaker who is not engaged in the conversation but will receive the signal.
- *Eavesdropper*: other actor not known to the speaker who is not engaged in the conversation but will receive the signal.

In addition, for each participant in the conversation $p \in C$, the model tracks whether or not the participant currently has the floor $FS(p)$, whether or not the participant intends (desires) to have the floor $FI(p)$, and which floor management action the participant is currently performing $FA(p)$. For the case of floor releases, we also model the set of participants the floor is being released to. The triumvirate of these variables allows us to model a variety of floor and turn-taking phenomena [4].

Currently, we use handcrafted models that fuse audio and visual information [4] to infer speech source and target, as well as floor state, actions, and intentions. We believe data-driven solutions that jointly consider sets of variables and all participants in the scene can exploit more detailed audiovisual information to enhance the accuracy of these inferences.

3.3 Turn Taking Decisions

The decision-making component in the proposed model is responsible for (1) deciding when the system should generate new discourse contributions ($CONTRIB$) and (2) selecting the floor management actions to be performed by the system (SFA) at any point in time. These decisions are currently handled via a set of rules that take into account the current turn taking context (*e.g.*, the inferred floor state, actions and intentions for each participant) as well as high level dialog information (*e.g.*, the current set of planned system outputs, etc.). The decision to produce a new contribution is made when a participant releases the floor to the system, and no planned outputs are already available. Other policies (*e.g.*, in which the system takes the initiative and attempts a floor *Take* while another participant has the floor) can however also be implemented with ease within the proposed framework.

We note that the turn-taking decisions are decoupled from input processing. The dialog manager processes all inputs as soon as they are detected, but generates contributions only when it

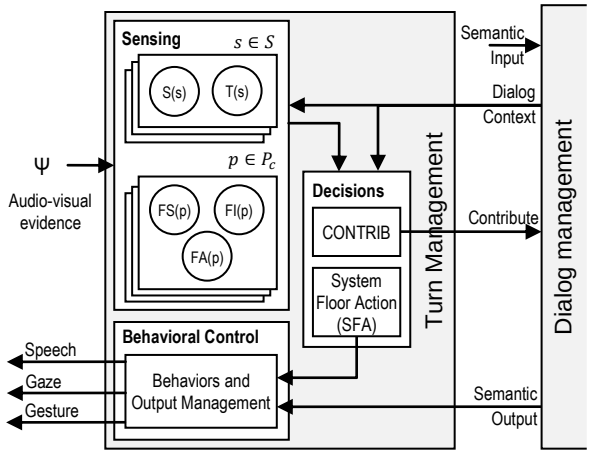


Figure 1. Turn taking components and overall architecture.

receives the corresponding signal from the turn-taking model. This decoupling enables flexible turn-taking behaviors in multiparty settings, *e.g.* allowing the system to monitor a side interaction between two participants, extract information from it, but only contribute when the floor is released back to it.

3.4 Behavioral Control

The behavioral layer is responsible for rendering the floor management actions taken by the system into a stream of coordinated verbal and non-verbal behaviors. These behaviors signal to the other participants the system's turn-taking intentions and help shape the conversational dynamics according to the system's needs. Below we describe in more detail the current implementation for of each of the four floor management behaviors in the context of an embodied conversational agent with controllable head pose and limited facial expressions.

3.4.1 Hold Behavior

The *Hold* floor action and corresponding behavior indicate that the system is in the process of maintaining control of the floor.

Cassell et al. [5] have shown that, apart from turn taking, information structure—specifically the relative locations of the utterance theme and rheme—also correlates with gaze direction. The current implementation of the *Hold* behavior directs the avatar's gaze away from the addressees during the thematic part of the current output, and towards them during the rhematic part. An example of this behavior is illustrated in Figure 2.

During the rheme, if the utterance is addressed to a single participant, the system directs its gaze towards that participant (*e.g.*, Figure 4.A, track *k*, t_1 to t_2, and t_6 to t_7). If the output is addressed to multiple participants, the avatar starts by looking at one addressee, but establishes brief eye contact with each of the other addressees at different points in time. Specifically, the gaze stays with that initial addressee for a duration randomly sampled from the [0.3s-0.6s] interval. Next, the system establishes eye contact in turn with each of the other addressees, for durations sampled from the [0.7s-1.1s] interval. Once this scan is complete, the system re-establishes eye contact with the first addressee, and maintains it for a longer duration, sampled from the [4.0s-6.0s] interval. If the output has not yet concluded by this point, another scan of the addressees is initiated. In addition, no gaze shifts are performed during the last 0.4s of the spoken output. The resulting behavior is illustrated in Figure 2.A, t_1 to t_4 and in Figure 4.A, track *k*, t_{12} to t_{13}.

Although the *Hold* floor management action and corresponding behavior are usually performed when the system is speaking, this is not always the case. For instance, a system may try to hold the floor for a while and stall the conversation without speaking, while waiting for an answer from a backend component. In this case, the agent avoids eye contact: the *Hold* behavior directs its gaze away from all participants.

3.4.2 Release Behavior

The *Release* floor action and corresponding behavior indicate that the system is in the process of yielding the floor. When selecting a *Release* action, the decision making component in the proposed model also designates a set of floor release targets (FRT) *indicating the participants that the system is trying to release the* floor to. Typically (although not necessarily) this corresponds to the set of addressees for the output that preceded the floor release.

When activated, the *Release* behavior directs the avatar's gaze towards the participants in FRT. If FRT contains a single floor release target, the *Release* behavior directs the avatar's gaze towards that participant (*e.g.*, Figure 4.A, track *k*, t_2 to t_3). If FRT contains multiple participants, the eye contact is established with one of these participants, and maintained for a duration sampled from the [3.0s-5.0s] interval. If the *Release* behavior does not conclude by that point, which implies that none of the participants in FRT took the floor, the avatar shifts gaze towards another randomly chosen participant from FRT. Furthermore, if the participant that the avatar is gazing towards shifts his or her attention towards another participant in FRT, the *Release* behavior also directs the avatar's gaze towards the new participant.

We have also explored non-verbal cues for signaling floor releases. Specifically, in certain cases (commanded from the deliberative layer and described in more detail in the next section) the avatar will lift the eyebrows while gazing towards one of the participants in *FRT* in order to signal a floor release. An example of this behavior is illustrated in Figure 4.A, track *k*, t_{15} to t_{16}.

3.4.3 Take Behavior

The *Take* floor action and corresponding behavior indicate that the system is trying to acquire the floor. The *Take* behavior first directs the system's gaze towards the participant that currently holds the floor, and, once eye contact has been established, it triggers the next spoken output.

3.4.4 Null Behavior

The *Null* floor action and corresponding behavior indicate that the system is not issuing any claims for the floor. Like *Take*, the *Null* action is only performed when the system does not have the floor,

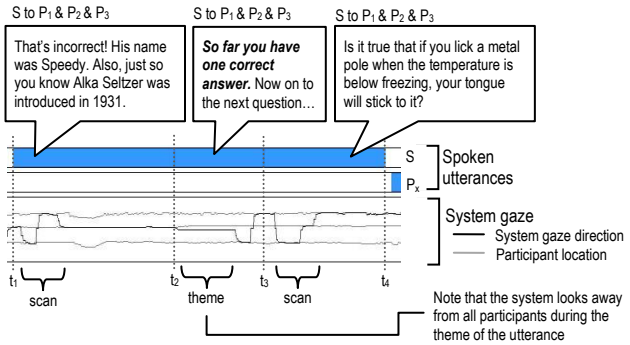


Figure 2. Example gaze behavior during *Hold* while addressing multiple participants (rheme region in italics).

which implies there is another participant P that has it. In this case, the gaze is directed as follows: if P is performing a *Hold*, (generally implying P is speaking), the gaze is directed towards P; alternatively, if P is performing a *Release*, the gaze is directed towards one of the participants that P is releasing the floor to.

3.4.5 Discussion

The models described above are informed by existing literature on the role of gaze in regulating turn taking, but are still relatively coarse. Numerous improvements are feasible. Examples include: generating stalls when the system has to hold the floor without any planned outputs; changing prosody on the fly to resolve overlaps (*e.g.*, if the system infers that a participant is attempting to take the floor while the system is trying to hold); generating backchannels with appropriate timing during the *Null* system behavior if a participant is speaking; stalling and drawing a participants' attention if eye contact cannot be established when the system is trying to take the floor, etc. While important technological challenges remain (*e.g.*, conversational speech synthesis with dynamic changes in prosody is an open problem) the framework we have proposed provides some of the core representations and mechanisms for implementing these more sophisticated behaviors. Finally, despite the simplicity of the current implementation, as we shall see in the following sections, the behavioral model described above can be used to successfully shape conversational dynamics in multiparty situations.

4. SYSTEM

We now briefly describe a situated spoken dialog system [2] that implements the turn-taking framework outlined above and that provided the test-bed for experiments we report in Section 5.

4.1 Hardware and Software Architecture

The system takes the form of an interactive multimodal kiosk (Figure 3) that displays a rendered head with controllable pose and limited facial gestures, and can interact via natural language with one or with multiple users.

A wide-angle camera is used in conjunction with face detection and head pose tracking software to detect and track multiple participants in the scene, as well as their focus of attention. A four-element linear microphone array captures the audio signal and performs sound source localization. The Windows 7 speech recognizer, configured with simple grammars, is used to perform speech recognition. A conversational scene analysis component fuses the resulting signals, and runs inferences about attention, engagement, turn taking, as well as the long-term goals and activities of various actors in the scene, etc. The results of this real-time scene analysis (some of them illustrated in Figure 2) are passed to a reactive control layer, which orchestrates the avatar's behaviors based on the semantic outputs planned by a finite-state-based multiparty dialog management component. A more detailed description of these components is available in [2].

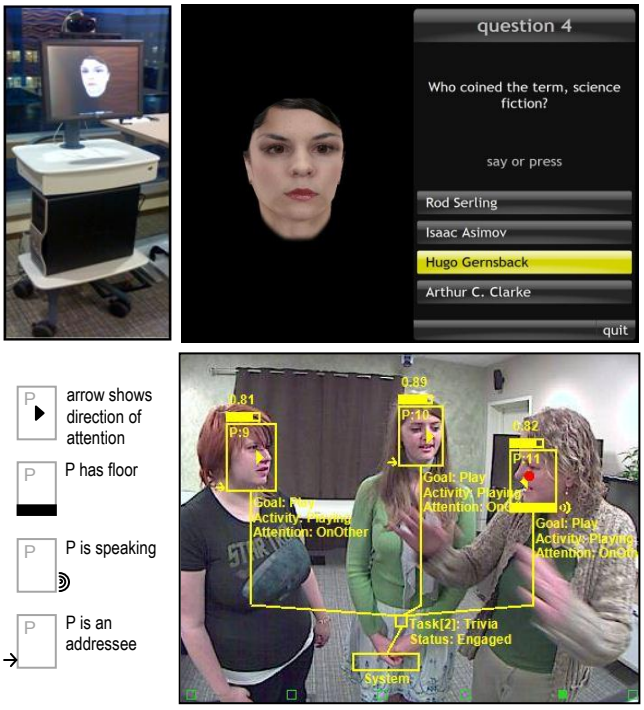
4.2 Questions Game Application

The experiments reported in the sequel were conducted with a questions game application, implemented using this framework. Figure 4 illustrates an excerpt from a multiparty interaction with this application. Videos of this excerpt, as well as other interactions with this system are available online [16].

Interactions with the questions game application start with an opening phase in which the avatar automatically engages approaching users [3] and asks them if they would like to play a game. Once engagement is established, the system asks a series of trivia questions (*e.g.*, the *Question.Direct* dialog act at time t_1 in Figure 4). When asking each question, the system also displays the set of possible answers on the screen, as shown in Figure 3. In these experiments, for each question, the system randomly decides whether to address it to one (randomly selected) participant, or to all engaged participants. For instance, in the segment shown in Figure 4, the question at time t_1 was addressed only to P_{16}; as discussed earlier, this was signaled by directing gaze towards P_{16} throughout the duration of the system's *Hold* and subsequent *Release* floor actions (see tracks *f* and *k* in Figure 4, from t_1 to t_4.)

In multiparty situations, participants may talk to each other, or even to themselves as shown in this example: at t_3, P_{16} is repeating the system's question to himself in a low voice. By inferring the speech source and addressees for each detected utterance, the system can monitor such side-exchanges and wait until the floor is being released back to it. This happens once an utterance is addressed to the system as in the situation at time t_6 in Figure 4.

Once an answer is received, and the recognition confidence score exceeds a grounding threshold, the system will seek the agreement of one other randomly selected participant via a *Confirm.Seek-Agreement* act. Two renderings of this act are available. One is verbal, *e.g.*, *"Do you also think that's true?"* or *"Is that correct?"* (*e.g.*, at time t_6 in Figure 4.) The second is non-verbal and relies on a simple facial animation: the avatar lifts its eyebrows while gazing and releasing the floor towards the addressee (*e.g.*, at time t_{12} in Figure 4.) Alternatively, if the recognition confidence score for an answer falls below a grounding threshold, an explicit confirmation is directed towards the participant that produced the answer, *e.g.*, *"Beethoven, right?"* (*Confirm.Value* dialog act, not



P_9, P_{10}, and P_{11} are all engaged in playing a trivia game with the system. The system is currently looking towards P_{11}, as shown by the red dot. The participants' focus of attention is directed towards each other. P_{11} has the floor and is currently speaking to P_9 and P_{10}.

Figure 3. System running the questions game application, and real-time scene analysis.

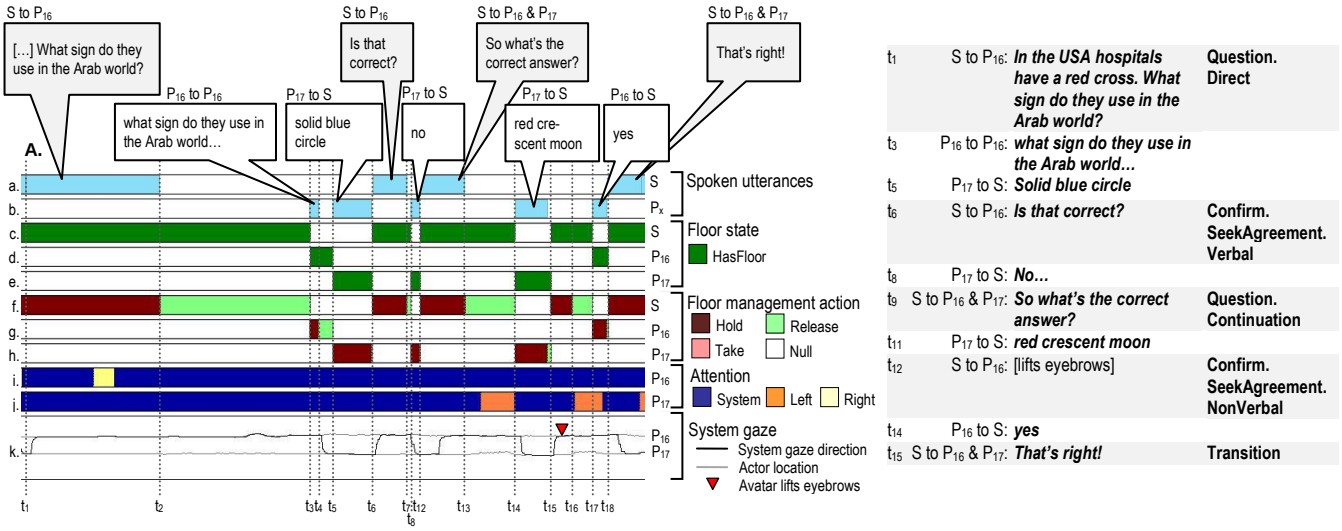


Figure 4. An excerpt from a multiparty interaction with the questions game system.

present in the example above).

If a negative response to the *Confirm.SeekAgreement* act is detected, the system tries to push the interaction forward with a *Question.Continuation* act which urges the participants to decide on a correct answer, *e.g.*, "*So what's the correct answer?*", (*e.g.*, at time t_9 in Figure 4.) The system also takes the floor and performs the *Question.Continuation* act if a long silence (>3.5 sec) is detected during which one participant releases the floor to another participant, but the second participant does not take it.

Once an answer is agreed upon, the system informs the users whether the answer was correct or not. If the answer was incorrect, a short explanation about the correct answer is also provided. Then, the system transitions to the next question. After six questions, the game concludes. At the end, the avatar informs the users about their performance and thanks them for playing.

5. EXPERIMENTS AND RESULTS

To validate the proposed framework and investigate how well the system is able to shape conversational dynamics using the turn-taking and behavioral model described earlier, we conducted a user study with the questions game system.

5.1 Experimental Setup

60 participants were recruited from the general population, 30 male and 30 female, with ages ranging between 18 and 61. The participants were recruited as pairs of people who were previously acquainted.

The experiment was conducted in a usability lab and consisted of 15 one-hour sessions. Each session involved four participants, *i.e.*, two pairs of two acquainted participants. In each session, using these four participants, we formed all possible subgroups of size two (six subgroups) and of size three (four subgroups). Each subgroup played one game with the system, and the order of the subgroups was randomized across sessions. This setup allowed us to collect a large set of multiparty interactions under diverse conditions (*e.g.*, all-male *vs.* all-female *vs.* mixed-gender groups; groups where people knew each other *vs.* groups where they did not; various age combinations, etc.).

The interactions were recorded through an overhead camera and microphone. In addition, the system logged its own camera view and the audio signal captured by the microphone array, as well as the frame-by-frame inferences and decisions it made at runtime. Several interactions are available for review and download [16].

5.2 Corpus and Annotations

A total of 150 multiparty interactions with the questions game system were collected: 90 two-party interactions and 60 three-party interactions. One of each had to be eliminated from the dataset due to systemic failures caused by malfunctions of the acoustic echo-cancellation on the microphone array.

The remaining corpus contains 148 interactions and 4605 spoken utterances that were detected by the system. Each utterance was manually annotated with source and addressee information. The source of each utterance was identified as either one of the participants, or as background noise. The set of addressees for each utterance was also identified. This set can contain one or multiple participants, including the system.

The annotations described above were performed on utterances detected by the system at runtime. Because the multiparty setting is especially challenging for a traditional voice activity detector, a number of utterances contained multiple (either subsequent or overlapping) utterances by different participants. In these cases, we identified the participant who spoke first in that segment. In addition, the recorded microphone array audio stream was used to segment and annotate utterances that were not detected by the system at runtime (*e.g.*, participants talking in a low voice or whispering to each other, etc.) This first set of annotations enabled the analyses described in the sequel. At the same time, a more thorough speaker diarization and annotation process, including efforts for computing inter-annotator agreement on these tasks are currently being planned.

5.3 Analysis and Results

The findings showed that the proposed models enable the system to successfully participate in multiparty interactions. Users rated the system's multiparty turn-taking abilities favorably. Participant responses to an overall, post-experiment subjective assessment questionnaire are discussed in more detail in [4].

We shall now perform a fine-grained analysis of the system's multiparty turn-taking abilities. Specifically, we focus on the

system's *Release* actions and investigate the system's ability to shape the conversational dynamics, that is, to influence who will speak next. To perform this evaluation, we focus our attention on the utterances that correspond to floor transitions from the system to a participant. We call these *first-response* utterances: they immediately follow a system output, are initiated while the system is performing a *Release* action, and account for 47.6% of the data. Other types of utterances include *overlaps* (18.3%), which are produced while the system is speaking, *continuers* (5.8%), which are initiated while the system is speaking but complete after the system has finished speaking, and *follow-ups* (28.2%) which are produced after first-responses, as illustrated in Figure 5. We eliminate from the analysis utterances spoken in the opening and closing phases of the game since engagement is in a transitional state during these periods and in many cases all users were not in view, etc. We therefore focus only on *first-responses to questions and confirmations*, which represent the bulk portion of the first-responses (90%).

We assume that the system has successfully shaped the multiparty conversational dynamics if, upon a floor release performed by the system, the participant that the system is currently gazing towards is the one who takes the floor and responds. Note that in single addressee situations, the participant who the system is currently gazing towards is the one who the system is releasing the floor to; similarly, in multiple addressee situations, the participant who the system is gazing towards is one of the participants who the floor is being released to.

To measure shaping success, we therefore define a *shaping score*, as the *probability that the current gaze target is the source of the first-response*. If the system exerted no influence and participants were to respond randomly, the baseline shaping score would be 0.5 in two-party interactions and 0.33 in three-party interactions. We thus report shaping scores separately for two-party and three-party interactions and, given the absence of a previous approach, we contrast our current solution against this agnostic baseline. We note that this experiment in effect creates a current baseline for the turn-taking model, which we hope to further improve upon.

Across all data, the shaping score is 66.3% in two-party interactions and 47.2% in three-party interactions. Both numbers represent statistically significant improvements over the corresponding baseline ($p<10^{-4}$.) These aggregate results are not very informative. As we shall shortly see, a number of contextual factors, such as dialog act type, current, and previous turn-taking context, and whether or not the system's output contains the deictic pronoun "you," can significantly influence the shaping score. Below, we investigate in more detail how the system's shaping ability varies with context, and discuss lessons learned from this analysis.

We begin with a breakdown by the *dialog act of the last system output, i.e., Question vs. Confirm*. As shown in Figure 6.A, the shaping score for questions is essentially no different than the random baseline (49.0% for two-party and 34.2% for three-party), while for confirmations it is much higher (81.7% and 58.5%) The observed differences between questions and confirmations, within

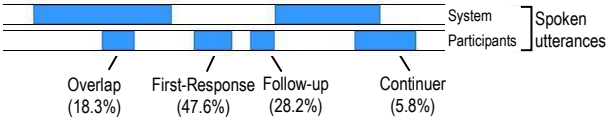


Figure 5. Different types of participant utterances.

two-party and three-party interactions are statistically significant ($p<0.05$ in a *t*-test, marked with a * sign in Figure 6.A). We also performed a logistic analysis of variance across the entire dataset: the dependent variable is the shaping score, and the independent variables are the interaction type (two- or three-party) and the feature under discussion, in this case the dialog act:

$$\text{LogOdds(ShapingScore)} \leftarrow \text{InteractionType} + \text{DialogAct}$$

The logistic ANOVA model indicates that the dialog act has a statistically significant effect, with $p<10^{-4}$ (shown next to the feature name in Figure 6.A.)

We believe the low shaping scores for questions is in line with the shared task nature of answering these questions. Participants were instructed to play the game cooperatively, and in this context responses are also shaped to a large extent by participants who actually know the answer. Furthermore, even though for 50% of the questions the avatar's gaze stayed with a single addressee, for the other 50% (which were addressed to all participants) eye contact was established with each participant as the system was producing the question. We believe this behavior further reinforced the overall shared nature of the question answering portion of the task, leading to the observed results. A breakdown based on whether the question was addressed to one or all participants shows no statistically significant differences (see Figure 6.K.) We are interested to investigate in future work how the structure of the task (shared problem-solving vs. a task where contributions have to be more clearly delineated) affects these dynamics. In contrast, for confirmations, which are always addressed to a single participant, the shaping score is much higher, which is in line with the discourse obligation created by the confirmation act towards the addressee.

Next we consider a feature of the turn-taking context: whether the *last output was addressed to the preceding speaker or to a different speaker*. Some cases in which the utterance preceding the last system output contained multiple overlapping speakers were excluded and this analysis was performed on the remaining (92%) cases. As Figure 6 shows, this factor also correlates with the shaping score: the score is larger when the last output is addressed to a different (new) speaker. Given the cooperative and shared nature of the overall game, this is consistent with the tendency participants might have to strive towards equally sharing the floor and maintaining engagement by alternating turns. While this effect is detected on confirmations (Figure 6.E) it is interesting to notice that the system attains relatively high shaping scores even when re-addressing the previous speaker: 74.5% for two-party and 51.8% for three-party interactions. This indicates that the system is indeed able to shape the conversation per its own intentions, independent of default turn-taking dynamics in the data.

A second turn-taking context feature that correlates with the shaping score is *whether the first-response is addressed to the system or to another participant*. The effect is detected on confirmations (Figure 6.F), questions (Figure 6.J), as well as across the entire dataset (Figure 6.C), and is in line with our expectations: if a first-response is addressed to the system, it is significantly more likely to be generated by the participant that the system was gazing towards.

Next, we take a closer look at confirmations, and investigate whether the presence of the *deictic pronoun "you"* in the confirmation prompt (*e.g., "Do you agree with that?" vs. "Is that correct?"*) correlates with the shaping score. As Figure 6.G shows the shaping scores are indeed higher when the deictic pronoun

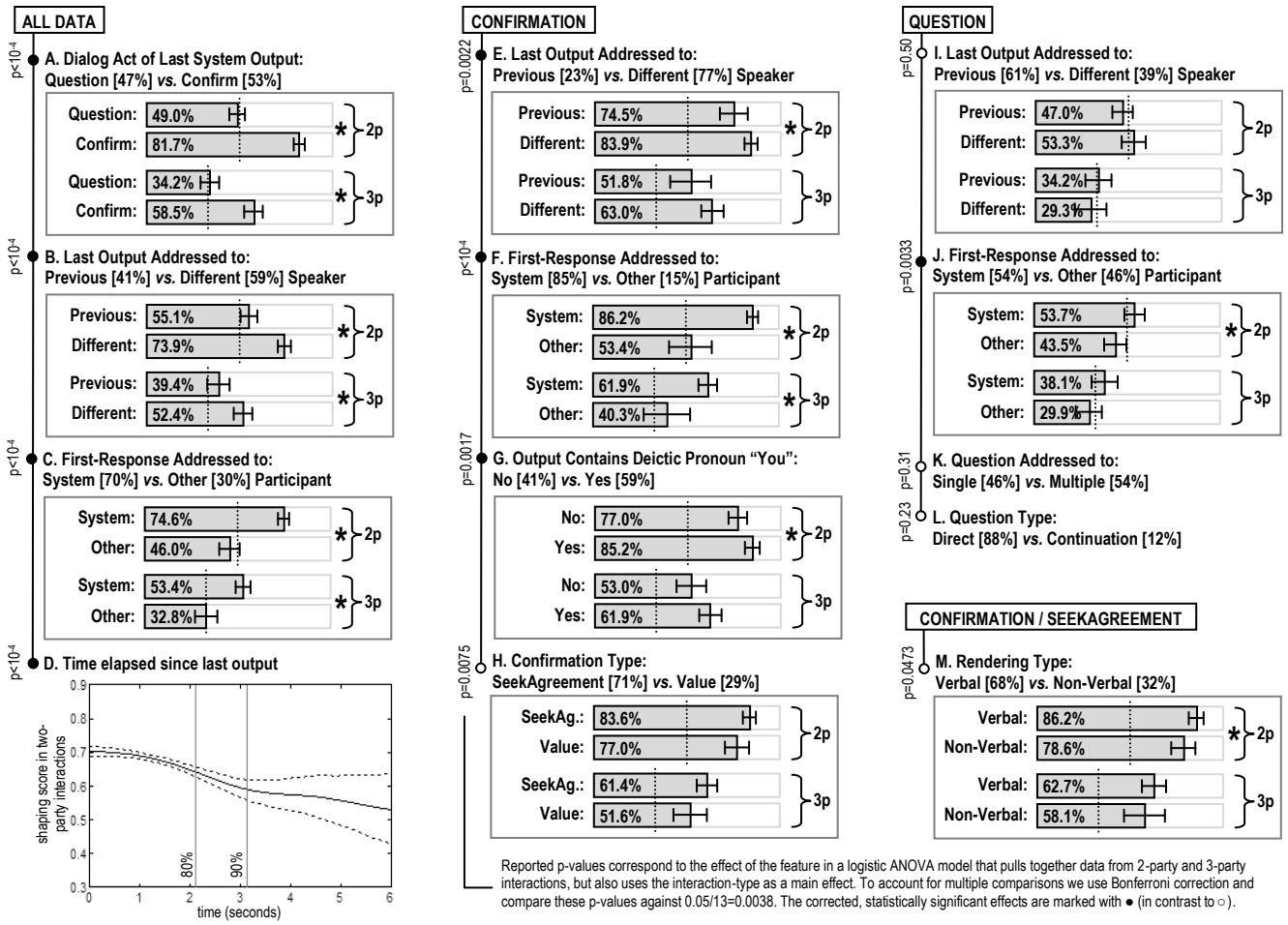


Figure 6. Breakdown of shaping score by dialog act and other contextual features.

"you" is present. For two-party interactions the score increases from 77.0% to 85.2%, while for three-party interactions it increases from 53.0% to 61.9%. While the latter difference is not statistically significant, the logistic ANOVA which pulls together the data from two-party and three-party interactions reveals a statistically significant effect (p=0.0017)

Similarly, a breakdown by *confirmation type* indicates slightly (but not statistically significant) higher shaping scores for *SeekAgreement* confirmations than for the *Value* confirmations. This result is in line with previous observations, in that *SeekAgreement* confirmations contained the deictic pronoun "you" and were addressed to a different speaker more often than *Value* confirmations.

Furthermore, within the *SeekAgreement* confirmations, we analyzed the difference in shaping scores between the *verbal and non-verbal renderings* (Figure 6.M.) Recall that in a non-verbal rendering, the avatar turned towards a participant and signaled the need for a confirmation and the subsequent floor release by simply raising its eyebrows. Although the verbal rendering seems to lead to higher shaping scores, we found that, coupled with gaze, this gesture-based rendering of the floor release signal still strongly conveys the system's intentions and shapes the conversational dynamics: the corresponding shaping scores are 78.6% in two-party and 58.1% in three-party interactions.

Within questions, no statistically significant effects were detected based on whether the *question was addressed to one vs. multiple addressees* (Figure 6.K) or based on *question type, i.e., Direct* vs. *Continuation* (Figure 6.L.)

Finally, we also investigated the relationship between the *time elapsed since the last system output* and the shaping score. The logistic ANOVA indicates a statistically significant negative relationship ($p<10^{-4}$) across the entire dataset, indicating that, as time elapses after the system finished the output, the probability that the participant the system is gazing towards will respond is decreasing. This corresponds to the intuition that, in a shared task setting, if it appears that the participant to whom the floor is being released does not take the floor, the probability that another participant will step in to take the floor rises. A Parzen density estimation of the shaping score conditioned on elapsed time in two-party interactions is shown in Figure 6.D, together with the 80 and 90 percentile lines for the elapsed time distribution.

The analysis above indicates that several contextual factors correlate with and potentially affect the system's ability to steer the interaction in multiparty setting. In an effort to assess the combined predictive power of these features, we trained a logistic regression classifier to predict shaping success on a turn-by-turn basis, *i.e.*, to predict whether or not the participant the system is gazing towards will respond. The resulting model, including weights and p-values for each feature is shown in Table 1. Results

Feature	Weight	p-value
Constant Term	-0.4749	0.0012
InteractionTypeIsTwoParty	0.9114	< 0.0001
LastOutputIsConfirm	0.9985	< 0.0001
LastOutputAddressedToPrevSpeaker	-0.2655	0.0171
FirstResponseAddressedToOther	-0.7295	< 0.0001
LastOutputContainsYou	0.2435	0.0521
SecondsSinceLastOutput	0.0364	0.3684

	Error	Avg.LL	Mean SE
Majority baseline	40.9%	-0.6764	0.2417
Model (train)	32.5%	-0.5938	0.2045
Model (10-cv)	32.7%	-0.5992	0.2069

Table 1. Logistic regression: model and performance.

from a 10-fold cross-validation process indicate that the model compares favorably to a majority baseline, reducing the classification error rate on this task from 40.9% to 32.7%.

6. CONCLUSION AND FUTURE WORK

We reviewed a model for turn taking and described how we use this model to render synchronized speech, gaze, and gesture in an embodied conversational agent in order to shape addressee roles in multiparty interaction. Empirical results conducted in a shared task setting indicate that the methods enable the agent to manage turn taking in multi-participant settings and to exert influence on the flow of the conversation. For instance, for verbal confirmations within interactions involving two participants and the system, the participant that the system had released the floor to was the first to speak in 86.2% of the cases. We are interested in performing studies of human performance in similarly structured tasks for comparative analyses of the competencies of the conversational system.

Turn taking is a mixed-initiative, collaborative process, influenced by multiple factors. We have identified and discussed here correlations among several variables, including gaze, dialog act type, previous speaker context, presence of deictic markers, and elapsed time, and the system's ability to shape turn taking. Other relevant concepts that impinge either locally or globally on conversational dynamics include issues of dominance and social relationships, domain-specific time constraints, grounding acts, and emotional state. We note that the shared structure of the task considered here had an influence on the observed results. In future work, we plan to study how well the models generalize to tasks and situations with a different collaborative structure.

The analysis and machine learning experiments we have performed highlight the potential for using high-level, contextual features in conjunction with lower level audio-visual features for enabling more robust predictions and tracking of key variables in the multiparty turn-taking process (*e.g.*, speech source, or floor actions and intentions). We plan to explore models with these mixtures of features in future work.

ACKNOWLEDGMENTS

We thank Isabelle Bouanna, Anne Loomis Thompson, 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.

7. REFERENCES

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

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

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

[4] Bohus, D., and Horvitz, E., 2010. Computational Models for Multiparty Turn Taking, MSR-TR-2010-115, Microsoft Research

[5] Cassell, J., Torres, O., and Prevost, S., 1998. Turn taking vs Discourse Structure: How Best to Model Multimodal Conversation, Machine Conversations, 1998, pp. 143-154, Kluwer.

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

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

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

[9] Itti, L., Dhavale, N., and Pighin, F., 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention, in SIPE 48th Annual International Symposium on Optical Science and Technology, 2003, San Diego, CA.

[10] Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., and Hagita, N., 2009. Footing in Human-Robot Conversation: How Robots Might Shape Participant Roles Using Gaze Cues, in Proceedings of HRI-2009, San Diego, CA.

[11] Picot, A., Bailly, G., Elisei, F., and Raidt, S., 2007. Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent, Intelligent Virtual Agents, 272-282, Springer Berlin.

[12] Raux, A. and Eskenazi, M., 2008. A Finite-State Turn-Taking Model for Spoken Dialog Systems, in Proc. HLT'09, Boulder, CO.

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

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

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

[16] Situated Interaction Website, 2010 - http://research.microsoft.com/~dbohus/si.html

[17] 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.

[18] Traum, D., and Rickel, J., 2002. Embodied Agents for Multi-party Dialogue in Immersive Virtual World, in Proceedings of AAMAS'02, 766-773.

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