When the Conversation Starts: An Empirical Analysis

David Novick[1], Guillaume Adoneth[2], David Manuel[2], and Ivan Grís[1]

[1] Department of Computer Science, The University of Texas at El Paso, El Paso, Texas
novick@utep.edu, ivangris4@gmail.com

[2] French Air Force Academy, Salon-de-Provence, France
guillaume.adoneth.ea2010@gmail.com, david.manuel.ea2010@gmail.com

Abstract. We describe an exploratory study of what happens when a person enters a room where people are conversing, based on an analysis of 61 episodes drawn from the UTEP-ICT cross-cultural multiparty multimodal dialog corpus. We examine the reliability of coding of gaze and stance, and we develop a model for room-entry that accounts for observed differences in the behaviors of conversants, expressed as a state-transition model. Our model includes factors such as conversational task, not considered in existing social-force models, which appear to affect conversants' behaviors.

Keywords: Embodied conversational agent, gaze, stance, state-transition model

1 Introduction

As developers of systems for real-time interaction with embodied conversational agents strive for increased realism in multi-agent settings, they confront the issue of how these agents should react when another agent, and especially a real human agent, enters the room. Are the conversants silent, chitchatting, or engrossed in heated interaction? Do they in some way acknowledge the entry of another person into the room? Existing approaches for addressing this problem rely on generalized models of interaction based on notions of social force. Such models, however, may not suffice to describe interaction in context with realistic social dynamics among the conversants.

Our analysis of actual conversations suggests that other factors not present in applications to date of the social-force model affect the patterns of behaviors of conversants when another person enters the room. These factors include (a) whether the conversants are speaking and (b) whether their conversation relates to their nominal task.

In this paper, then, we describe an exploratory study of what happens when a person enters a room where people are conversing, based on an analysis of 61 episodes drawn from the UTEP-ICT cross-cultural multiparty multimodal dialog corpus [1]. We examine the reliability of coding of gaze and stance, and we develop a model for room-entry that accounts for observed differences in the behaviors of conversants. We conclude with a discussion of limitations of our study and avenues for future research.

2 Background

As real-time immersive environments with embodied conversational agents (ECAs) grow more interesting and more complex, they have evolved from single-ECA, walk-up-and-use systems, to multi-ECA training systems for negotiation, to multi-ECA systems in relatively realistic environments. Recent systems present the question, then, of how the ECAs should behave when the human user/agent comes into their space. Because people in real life rarely stand around in groups waiting for someone else to arrive before beginning their conversation, the issue really becomes how ECAs should react if they are already conversing.

To produce realistic behaviors for ECAs when another person enters a conversation, researchers have built on models of social force (e.g., [2]), which account for proxemics in dynamic contexts. Modeling proxemics only, Jan and Traum [3] produced group behaviors upon entry of anther person into a conversation, including an interesting case where agents had different cultural norms for proxemics. The authors judged their simulation produced the expected behavior for the agents in all of the three cases they modeled.

Where Jan and Traum looked purely at the spacing of agents, Pedica and Vilhjálmsson [4] extended their work through a model of territorial behavior model that extended the social-force model with agent goals and a more detailed account of proxemics. In particular, the model included orientation of each agent's body relative to the other agents. This research added significantly to the study of entry into a conversation because it modeled the agents' behaviors in terms of component goals, such as "move to destination" and "avoid obstacles."

In both [3] and [4], the agents' behaviors in the simulations came from general models of proxemics and social force, building on Kendon's analyses of f-formation [5]. In neither study did the authors observe and analyze actual instances of a person joining a group or of a person entering a room with people in it. This made validation of their models somewhat problematic. In [3], the simulation of the test cases produced behaviors that were judged by the authors to conform to the expectations that underlay the model. In [4], the model was evaluated through a perception study. The simulations resulting from the model were viewed by 171 subjects, who compared the naturalness of four pairs of videos where one reflected the model and the other was a base case without model-influenced behaviors. For three of the four pairs of videos, subjects found the videos based on the models to be more natural than the base-case videos.

In neither [3] or [4], then, were the simulations validated with respect to actual instances of conversation-joining episodes, probably because corpora of multiparty interaction typically do not include this sort of situation. Accordingly, the question is open as to what people actually do when, while engaged in conversation, another person enters their space. From episodes recorded as part of the process of corpus collection of the UTEP-ICT cross-cultural multiparty multimodal dialog corpus, we examine the reliability of coding agents' gaze and stance, suggest an empirically based model of entry into a room with conversing agents, and argue that models based on social force alone do not account for factors that significantly affect conversants behaviors in entry situations.

3 Methodology

To model appropriate behaviors for ECAs when another agent comes into the room, we analyzed naturally occurring entry episodes from the UTEP-ICT cross-cultural multiparty multimodal dialog corpus. The corpus comprises 24 conversations, among a total of 72 conversants. Half the conversations involve dyads, and half quads. The conversations were evenly split among American, Mexican, and Arab conversants, interacting in within-culture groups. Figure 1 presents a typical scene from a conversation in a quad.

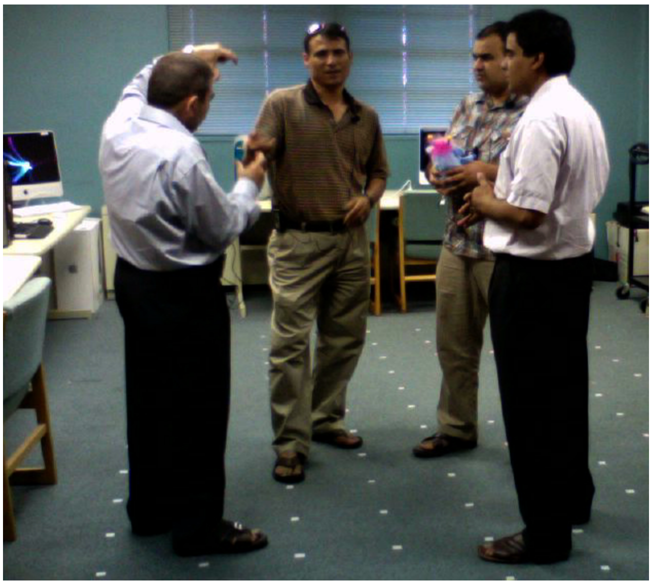

Fig. 1. Screen shot of conversation from the UTEP-ICT cross-cultural multiparty multimodal dialog corpus. One of the conversants holds a toy involved in two of the tasks.

Each group had five ten-minute conversational tasks. Tasks 1, 4, and 5 were mainly narrative tasks, where the participants can take turns relating stories or reacting to the narratives of others. Tasks 2 and 3 were constructive tasks, in which the participants pooled their knowledge and work together to reach a group consensus. Tasks 3 and 4 were designed to have a toy provide a possible gaze focus other than the subjects themselves, so that gaze patterns with a copresent referent could be contrasted with gaze patterns without this referent. Task 5 was meant to elicit subjective experiences of intercultural interaction. The interactions were recorded with six Apple iMac computers, placed around the periphery of a large open room that serves as a computer lab. We thus recorded six simultaneous views of the subjects as they conversed, making it possible, with rare exceptions, to code the subjects' proxemics, gaze and turn-state.

For the first task, an experimenter described the task to the group and then left the room. After the ten minutes (or if the conversation had appeared to conclude in less than ten minutes), the experimenter reentered in the room to give instructions for the next task. In cases where the participants had not concluded their task in ten minutes, the participants were typically still working on the task when the experimenter reentered the room. In cases where the participants had concluded their tasks, the participants were typically engaged in non-task social interaction when the experimenter reentered.

Thus for each group, the corpus contains recordings of four episodes in which the experimenter reenters the room. Because these entry interactions were ancillary to the main purpose of the corpus collection process, the subjects behaved spontaneously. Of the 96 possible cases of reentry by the experimenter, we found 61 usable episodes: 31 in conversations in quads and 30 in conversations in dyads.

For each usable episode, we annotated the conversants' gaze and stance, and we tracked the transition of the conversants' speech and gaze. The conversants behaviors were annotated using the Elan tool [6]. We coded gaze as toward the group (G), the experimenter (E), the toy (T), or Away; and we coded body stance similarly as G, E or Away). We also coded who was speaking. Finally, we developed a state-transition model that reflected the transitions between conversants' observed paralinguistic behaviors, noting the context for the transitions.

4 Results

We now turn to the results of our analysis, looking first at the coding of the conversation behaviors, second at our observations of typical behaviors, and third at the state-transition model.

4.1 Annotations

Each episode was independently annotated by two of the authors. We assessed the reliability of our annotations with Cohen's Kappa [7]. For this purpose, we considered that our annotations were similar when the annotators agreed both on the action and on its duration; instances were considered similar if difference in time between the two did not exceed 300 milliseconds.

For our annotation of the conversants' gaze, we obtained a Kappa coefficient of 0.64, indicating substantial agreement. For our annotation of the conversants' stance, we obtained a Kappa coefficient of 0.57, indicating moderate agreement. The difference in the reliability of the annotation reflects the subjective experience of the annotators, who found it harder to code body stance than to code gaze, because the conversants' positions were often ambiguous. Accordingly, in the formal part of the state-transitional analysis that follows, we used gaze alone, rather than both gaze and stance.

4.2 Observations

As an introduction to our formal state-transition model, we begin with a more informal account of two aspects of recurrent behaviors among the conversants that we observed when the experimenter entered the room.

Initial Position in the Room and Body Movements

In the episodes we studied, we observed that head movements tended to precede body movement. In most of the episodes, subjects tended to move their head in the direction of the experimenter before moving their body, depending on their position.

If subjects could see the door (from which the experimented entered the room) in their original position, they turned their head in the direction of the experimenter when he entered, and then slightly moved their body. Most of the time, the conversants' stance did not point exactly at the direction of the experimenter but rather away, between the group and the experimenter. In general, the subjects' bodies seem to form an angle of between 10° and 60° with respect to the experimenter. The subjects' gaze but not their bodies were in the direction of the experimenter. The subject's movements of their torso were is quite small, rarely exceeding dozen degrees.

Typically, the subjects turned their body after turning their head and gazing at the experimenter, with the delay between head and body movement depending on the subject. We found no real pattern in the episodes. Some subjects turned their head and torso almost at the same time, while others turned only their heads. But in most cases, we observed a slight torso turn, occurring 300 to 1000 milliseconds after the head movement.

If subjects had their back to the door, in most cases they would turn around to be able to see the experimenter. But their bodies typically did not point at the experimenter. Again, the stance is directed away, with only the head directed at the experimenter. The subjects typically turned around by spinning around one of their feet: one foot is moved backwards, and the subject rotates around the other one. The foot moves and head turns appear to be separated by less than 300 milliseconds. In more than half of the cases we analyzed, the subjects' stance points at the group. The body stance is then directed away, and seldom ends pointing directly at the experimenter. However, when conversations started, conversants turned to face their interlocutor.

Our observations tended to confirm the realism of the some of the behaviors simulated in [3] and [4]. In particular, if the conversants in a circle when the experimenter entered the room, they tended to open the circle around the experimenter. This means that subjects with their back to the door often move as they are turning around.

In general, then, the subjects tended to look at the experimenter but maintained their stance so that their body was not clearly directed towards him.

Speakers and Listeners

When a conversant in the group was speaking, the conversants reacted to the experiment's entry in ways that differed depending on the nature of their discussion. We were able to classify these as task discussion and non-task discussion.

When the group's topic involves the task they've been assigned, the conversants speaking tended to keep speaking when the experimenter entered the room. While continuing to speak, they tend to turn around to look at the experimenter and then turn back to face the rest of the group to complete their utterance. With the entrance of the experimenter, though, the discussion ends quickly.

In contrast, when the topic is about something other than the assigned task, the conversants tended to look at the experimenter, staying focused on him, and the speaker quickly stopped speaking.

4.3 State-Transition Model

Based on our observations of the conversants' behaviors upon the experimenter's entry in to the room, we propose a state-transition model that reflects differences in behaviors given (a) whether a conversant is speaking when the experimenter enters the room and (b) the nature of the conversants' conversation. The model is depicted in Figure 2.

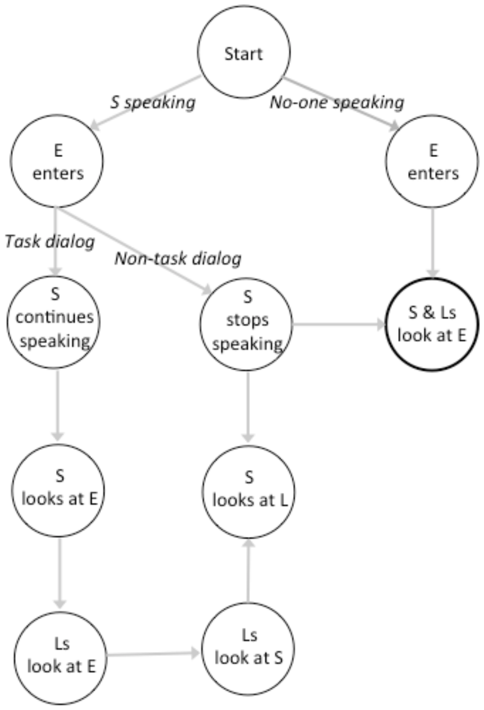

Fig. 2. State-transition model of conversants' behaviors when the experimenter enters the room. E is the experimenter, S is the speaker, and Ls are the listeners (or listener, in the dyadic case).

5 Implications, Limitations, and Future Work

The design of plausible ECAs in situations where another person enters the room can be built from the agents' gaze and stance. Our results suggest that head turns should be independent from body movements. Our results also suggest that an ECA's reaction should depend on its position relative to the direction from which the additional person enters, which here we describe as the door; we consider three cases.

- Case 1: The ECA is facing the door. In this simplest case, the ECA should maintain its stance and directly look at the person entering the room.
- Case 2: The ECA has its back to the door. In this case, the ECA should first turn its head to be able to see the additional person as he or she enters the room. Second, after a delay of about 300 milliseconds, the agent should, turn its body to face the person. We note that the agent's body should not point directly at the person but rather at an angle varying between 10 and 60 degrees relative to the person. Particular attention should be given to the agent's foot movement: one foot should move backwards, and then the ECA turns around.
- Case 3: The ECA has the door in view. In this case, when the additional person enters the room, the ECA can either turn its head and body or merely turn its head; our analysis did not identify factors to distinguish between these two behaviors. Regardless, the agent's change of stance is minimal; the torso movement involves only a few degrees of rotation (fewer than 10 degrees), and the ECA's body should not be pointed directly at the entering person.

5.1 Limitations

Our study has several limitations, including the nature of the relationship between the subjects and the person entering the room, the fact that we could not annotate about a third of the possible episodes.

While the conversational behaviors we observed were actual spontaneous interaction, they may not represent the full range of possible reactions to an additional person entering a room to join a conversation. In our study, the person entering the room had a special relationship with the subjects: the experimenter was a figure of authority, directing the conversational tasks. This may have affected the behaviors represented in the state-transition model because, as the experimenter is seen as an authority figure, the conversants might consider off-task conversation less important than the on-task conversation. Consequently, the reactions of the subjects might have been different if a different sort of person—a friend, a relative, a stranger, a co-worker—had entered the room. Nevertheless, the set of behaviors we observed do represent actual human reactions and constitute a reasonable initial basis for empirical study of the initiation of conversations.

While the corpus collection included five camera angles, in some cases the recordings of the experimenter's entry did not catch the conversants sufficiently for annotation of gaze and stance. We used the widest "master" angle for our analysis, but in every quad and in some dyads there was at least one episode where for this angle some people were out of camera range or had their back to the camera. As a result, we annotated only 61 of the 96 possible episodes. It would be possible to salvage additional episodes by looking at other camera angles.

5.2 Future Work

Our analysis did not consider differences across the cultural groups (American, Mexican, Arab), largely because the number of episodes within in each group would have been small enough to render the results likely unreliable. We are looking for analytical approaches that might be able find any intercultural differences from the data we have.

More broadly, we suggest that it would be useful to study conversational behaviors with different kinds of conversations and different kinds of participants, especially with respect to the relationship between the conversants and the additional person entering the room. In this regard, it would likely be helpful to try to model the conversants' behaviors in terms of the agent's component goals, following the example explored in [4].

In the meantime, we are implementing ECA behaviors in the Interactive Systems Group's immersion space based on the behaviors described in the state transition model. We expect that, as a person enters this room of our laboratory, conversing ECAs projected on the far wall will react realistically, reflecting the empirical analysis described in this paper.

References

1. Herrera, D., Novick, D., Jan, D., Traum, D.: "The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus." In: Proc. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, LREC 2010, Malta, May, 2010, 49-53 (2010).
2. Helbing, D., Molnár, P.: "Social Force Model for Pedestrian Dynamics." In: Physical Review 51 (5), 4282 (1995), cited in [4].
3. Jan, D., Traum, D.: "Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversation." In: Proc. of the ACL Workshop on Embodied Language Processing, 59-66, (2007).
4. Pedica, C., Vilhjálmsson, H.H.: "Spontaneous Avatar Behavior for Human Territoriality." In: Applied Artificial Intelligence 24(6), 575-593 (2010).
5. Kendon, A. Spacing and orientation in co-present interaction. Lecture Notes in Computer Science, 5967, 1-15 (2010).
6. Sloetjes, H., Wittenburg, P.: Annotation by Category - ELAN and ISO DCR. In: Proc. 6th International Conference on Language Resources and Evaluation (2008).
7. Carletta, J.: "Assessing Agreement on Classification Tasks: The Kappa Statistic." In: Computational Linguistics, 22(2), 249–254 (1996).