Managing Chaos: Models of Turn-taking in Character-multichild Interactions

Iolanda Leite
Disney Research, Pittsburgh
and INESC-ID, Lisbon
iolanda.leite@ist.utl.pt

Hannaneh Hajishirzi
Disney Research, Pittsburgh
and University of Washington
hannaneh@uw.edu

Sean Andrist
Disney Research, Pittsburgh
and University of
Wisconsin–Madison
sandrist@cs.wisc.edu

Jill Lehman
Disney Research, Pittsburgh
jill.lehman@disneyresearch.com

ABSTRACT

Turn-taking decisions in multiparty settings are complex, especially when the participants are children. Our goal is to endow an interactive character with appropriate turn-taking behavior using visual, audio and contextual features. To that end, we investigate three distinct turn-taking models: a baseline model grounded in established turn-taking rules for adults and two machine learning models, one trained with data collected *in situ* and the other trained with data collected in more controlled conditions. The three models are shown to have different profiles of behavior during silences, overlapping speech, and at the end of participants' turns. An exploratory user evaluation focusing on the decision points where the models differ showed clear preference for the machine learning models over the baseline model. The results indicate that the rules for language interactions with small groups of children are not simply an extension of the rules for interacting with small groups of adults.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: [Audio input/outputs]; H.5.2 [**User Interfaces**]: [Natural Language]

Keywords

Multiparty turn-taking; multimodal inference; child speech behavior; overlapping speech; character-child interaction.

1. INTRODUCTION

Turn-taking is an essential mechanism in human language interactions [19]. Our ability to determine whether and when to respond to each other is largely unconscious and effortless, despite relying on a mix of contextual, verbal, and gestural cues that unfold over time. Smooth and natural exchanges are also desirable when humans talk with virtual characters or social robots [5, 7], and inappropriate turn-taking behavior by the character may undermine the user's sense of engagement. Adult behavior suggests three essential rules for managing the turn, which can be paraphrased: take the turn promptly when addressed, don't talk when somebody else is talking, and don't leave long, uncomfortable silences. Most adults follow these rules easily in dyads and, with some effort, fairly successfully in group situations. A character implemented with a turn-taking model that captures these rules, then, will be able to interact effectively with adults in a wide variety of circumstances. The same model might be ineffective in character-child interactions, however, as children have been shown to deviate from adult turn-taking rules in significant ways, particularly when they are in groups [13, 12].

The domain we work in–language-based games played by small groups of children–is an example of a situation in which the adult model seems inadequate. When hosting an interactive game, an animated or robotic character has one overarching goal: to make play effective and fun. The character must know when it has the opportunity or obligation to speak for play to be effective, but it must also accommodate a certain amount of "inappropriate" turn-taking behavior (e.g., talking at the same time) for play to be fun. If two children have overlapping speech, to whom should the character respond? The first speaker? The loudest speaker? The speaker who has played the least? Should the character wait until nobody is speaking to take the turn? Should it wait even if it is responding to the earlier speaker? Game play is not the only environment in which these questions arise; a character who answers children's questions at a museum exhibit or one who guides small group exploration in a classroom must address them as well.

In our game a human wizard answers such questions implicitly, making turn-taking decisions as a side-effect of performing the character's speech processing. His goal, like the character's, is to make game play effective and fun. Thus, in making decisions to act or wait, he implements an unarticulated policy that both serves that goal and provides training data for learning autonomous behavior. Such a solution also has problems, however: the wizard may rely on features of the environment (e.g., eye gaze) that he can sense but the character cannot. Moreover, in our highly dynamic environ-

ICMI'13, December 9–13, 2013, Sydney, Australia

ACM 978-1-4503-2129-7/13/12 ...$15.00.

ment, the wizard's reaction time can be both variable and long enough to introduce significant noise into the data.

In the remainder of this paper we briefly review existing approaches to turn-taking, then compare the performance of three models with respect to the audio-visual data collected from groups of children playing our game. The baseline model makes turn decisions while trying to enforce the three adult rules. The second model uses features that encode the participants' contextual, verbal, and gestural cues over time to learn turn decisions based on the wizard's behavior *in situ*. The third model also uses multimodal features and learns to make turn decisions, but is designed to overcome the wizard's reaction time lag by training on *post hoc* human annotations collected at strategic moments in the game. We find that the three models have different profiles of behavior during silences, overlapping speech, and at the end of participants' turns. We then report on a user study which shows preference for the learned models' performance over the baseline's, and discuss future work.

2. RELATED WORK

Seminal research on turn-taking in adults can be found in the areas of linguistics and conversation analysis [19, 21, 11]. This work has shown that there is an inherent structure behind turn-taking and that people employ several verbal and non-verbal cues to ensure minimal overlaps in a conversation. However, the analysis of turn-taking in children suggests that these regularities are not always observed, or at least they differ from the ones studied in adults [12, 13].

Regularities in turn-taking by human adults motivated the development of several computational models of turn-taking for embodied virtual agents [5, 3], social robots [4, 7] and dialogue systems [20, 22, 18, 24]. Fluent turn-taking behavior has been shown to improve the perceptions of life-likeness and fluidity of interaction [5], or even increase task efficiency [7].

Most of the existing turn-taking models and studies are focused on dyadic interactions with adults. For turn-taking in a multiparty setting, Bohus and Horvitz [3] proposed a decision-theoretic approach to balance the tradeoffs between waiting and taking the floor with the goal of minimizing the gaps in the interaction between a conversational agent and two adult participants. The model's cost function was refined with a *post hoc* data collection, in which a small group of annotators were asked to identify turn-taking errors of the system. The Ymir Turn Taking Model (YTTM) also works in multiparty settings [23]. YTTM includes parameters such as *urge-to-speak* and *yield tolerance* based on multimodal perceptions of each participant, and was evaluated using a virtual simulation of 12 agents interacting cooperatively with each other. In contrast to our approach, these two examples presume that turn-taking is cooperative and "polite," assuming, for instance, that no one interrupts the participant who is holding the floor.

Recognizing the end of a speaker's turn or the addressees in a conversation are examples of turn-taking related phenomena that have been investigated in multiparty settings. Matsusaka and colleagues [16] proposed a framework for simultaneously predicting dialog acts and addressee types. Work by de Kok and Heylen [10] employed a probabilistic approach to predict end-of-speaker turns using prosody, gestures, and focus of attention. More recently, Nakano and Fukuhara [17] analyzed multimodal data, including gaze and turn-taking behavior, to build a linear regression model that estimates conversational dominance in a group. They found that more dominant participants tend to take more (and longer) turns in a conversation. In a corpus of children playing an interactive game with a virtual character, Hajishirzi et al. [15] investigated the question of deciding whether or not a child's utterance was directed to the character using a set of visual, prosodic, and behavioral features.

Despite the large body of prior research on turn-taking in adults, only a few authors have explored turn-taking in groups of children [2, 14]. Blomberg et al. [2] deployed a robotic head that interacted verbally with groups of children and adults in a public space. Some simple turn-taking mechanisms were implemented in the robot; for example, when a visitor approached the installation, the robot would take the turn and ask the visitor a question. The analysis of the collected data raised the need for developing appropriate turn-taking and speech recognition mechanisms tailored for children. In this paper, we extend this line of research by developing and evaluating turn-taking models for an animated character that interacts with groups of children, where the timings, overlap tolerance and social cues for taking, holding or releasing the floor may differ from those dynamics in groups of adults.

3. TASK SCENARIO

Our work is done in the context of Robo Fashion World (RFW), an interactive game designed to facilitate the collection of audio and visual language data from young children in small groups. The data set reported here represents games played during the summer of 2011 by 65 compensated children (31 males, 34 females) and seven parent volunteers. All children were native English speakers and ranged in age from four to ten ($M = 6.8, SD = 1.9$ years). Participants were assigned to groups based on the convenience of scheduling for the family, resulting in 29 groups of one to four children with or without a parent or adult experimenter. The average group size was 3.2 (2.4 children), and about 60% of the groups combined children across families.

RFW is a game in which children dress up a fashion model in the center of the screen with silly clothing items and accessories available on a board (see Figure 1). The game is hosted by Edith, an animated robot character, who is responsible for mediating the interaction and making the costume changes. After a brief introduction that includes choosing a model, Edith explains the two main game actions: requesting a change to the model by naming one of the clothing items or accessories on the board and requesting a picture of the model to be printed and taken home after the game. Play then enters the *choice cycle* where, during each of 20 iterations, a valid reference to a board item is made, the model changes, and a replacement item appears on the board. Participants stood side-by-side approximately six feet away from a large screen where the game was displayed. The interactions were video recorded using one frontal and two lateral cameras. Sound was captured using both individual close-talk microphones and a linear microphone array located under the screen.

During game play, a human operator performed Edith's language understanding in a Wizard-of-Oz design. An interface allowed the wizard to signal a clear reference to each of the board items, a request for a picture, an utterance directed to Edith that was unclear, a pause that was too long,

Figure 1: A screenshot of Robo Fashion World.

or multiple voices speaking at once. Only one of the events could be chosen at a time, and choosing an event would always result in some action by the character. Thus, even though the interface has fewer than a dozen options, deciding when to take the turn is complicated: some events can co-occur (e.g., multiple children can call out different board items at once), a single, clearly-spoken object reference may be hard to resolve because children sometimes use their own vocabulary (*a king hat* rather than *a crown*), and object references that occur as part of side conversations need to be understood but not reacted to. The wizard's decision to signal an event or wait was based on his tacit understanding of the rules of turn-taking in this kind of interaction, the children's behavior, and the general criteria that the game should move along effectively and be fun. Log files containing the timing and content of wizard actions, the behaviors employed by the character as a result, and the changing state of the game board were generated automatically.

Parents who volunteered to play were instructed to support their children's participation in whatever way felt natural for their family. Experimenters who participated in groups with young children made game choices only when the group's turn-taking behavior warranted it. As a result, the activity was largely controlled by the children, who spoke to the character about 80% of the time (2025/2535 utterances) and took about 84% (2535/3034) of all conversational turns. The children also tended to do most of the gesturing, with 80% of the total clapping, head nodding, pointing, and emphasis motions. Parent utterances (157 total) were addressed equally often to the character or children; experimenter utterances (342) were addressed to the children about three quarters of the time. All participants displayed *situational attraction* [1], the tendency to orient physically toward the screen rather than each other during some or all of a non-character directed utterance.

Although audio and visual regularities at the level of single utterances are important when building a turn-taking model, the turn-taking style of the group as a whole is also critical. We formalize the differences among our groups by defining the *chaos factor* in a game as the percent of participant utterances that begin while another participant is talking. Chaos factors ranged from 3% to 36% (median of 12%). It is not surprising that chaos is highly correlated with both total group size ($r = .52$, $df = 27$, $p < .01$) and the number of children in the group ($r = .64$, $df = 27$, $p < .01$); the more participants (or children), the more opportunity there is for overlapping speech. It is a bit surprising, however, that groups that included parents, groups that included experimenters, and groups that included neither fell at both the high and low ends of the range. In other words, some groups self-organized their turn-taking to be quite orderly while others gravitated to more chaotic play, and which kind of turn-taking occurred could not be predicted by the presence or absence of an adult. The wizard was able to accommodate this variability; an autonomous Edith must be able to do so as well.

4. TAKE OR WAIT? THREE MODELS

To replace the wizard with a fully-autonomous character, Edith must know when a participant is speaking, when the speech is addressed to her, what was said (and meant), and when she should take the turn to respond. We conceive of the autonomous agent as a set of sensors and semi-independent modules that map audio and video inputs into features that combine with task context to make these high level decisions. Features have been chosen based on analysis of the participants' behavior, prior research on addressee identification and turn-taking in adults, and the constraint that features must be detectable in our physical environment. The detectability constraint means, for example, that we rely on head orientation as a substitute for eye gaze; although eye gaze is known to predict the release of a turn, it cannot be tracked in our groups.

Because we want to distinguish the adequacy of the feature set from the adequacy of particular feature detectors, our models are built initially using labels generated by human annotators. Annotated models represent the ability of a given set of features to capture regularity given human levels of sensing and create an upper bar for performance with automatic detectors. Table 1 describes each feature and its origin. Language annotations were derived using the video and close-talk microphone audio, and segmented into utterances based on pauses of at least 500 msec. Gesture and head orientation were derived from the front-camera video. Rather than impose an *a priori* duration or angle for human judgments of head orientation, annotators were told to use *turn-away/turn-back* labels, and to mark a turn when it was associated with meaningful interaction with a person or object, but not to mark brief, incidental head movements. Further details, including a discussion of inter-annotator reliability are given in [15].

Our goal is to develop a model that determines whether Edith should take the turn or wait at every moment of the interaction. We expect implemented feature detectors to run in a distributed fashion, however, so models based on feature vectors that combine output across modalities will require synchronization. In anticipation of this, we map the extracted features into 500 msec time slices, where the value of a feature at the synchronization boundary reflects what happened in the majority of the interval. Thus, a hand-annotated feature will be true in the vector exactly when the annotator's label extended over at least 250 msec of the interval. As the top of Figure 2 shows, the time slice mapping makes it appear to the model as if every utterance stops and starts on a 500 msec boundary. Power and pitch are also coerced to a single value for a time slice, mapping the average power/pitch in the interval to one of low, medium, or high. We define three turn-taking models on top of this common underpinning.

Feature name	Description	Source
Head Orientation	Facing Edith or away	Hand annotated
Gestures	Presence of head shakes, pointing or emphasis motions	Hand annotated
Voice Activity	Whether a participant is speaking	Hand annotated
Addressee Identification	Whether speech is to Edith	Hand annotated
Yes-No Words	Use of *yes*, *no* or synonyms in an utterance	Hand annotated
Valid Asset Word	Use of *picture* or word(s) that refer to a board item	Hand annotated + log files
Pitch	Pitch of the participant's speech signal	Automatically extracted
Power	Volume of the participant's speech signal	Automatically extracted
Prompt	Whether Edith's last utterance required a response	Log files

Table 1: Features used to train the turn-taking classifiers.

4.1 Baseline Model (M_B)

In a game where multiple children may be calling out their choices simultaneously, each trying to produce the change to the display that he or she prefers, the host can try to keep the game moving by simply responding to the first utterance in each choice cycle that it can understand. The baseline model, M_B, encodes this minimal functionality, taking the turn at the end of any utterance that is addressed to Edith, irrespective of other considerations. Keeping the game moving also requires that Edith takes control at some point during a long pause. M_B takes the turn after 3.5 seconds of silence, a value that has been used in a multiparty game with adults [3].

Despite the simplicity of the algorithm, the kind of behavior an implementation of M_B would exhibit *in situ* may seem variable and complex. If the group is chaotic and speech recognition is poor, Edith is likely to appear quite passive, essentially waiting until only one speaker can be heard clearly. Children might notice this and respond by self-organizing their behavior to be less chaotic or might grow frustrated because the character seems unresponsive to too many valid choices. In less chaotic environments, non-verbal cues could enable Edith to grab the turn despite some degree of overlap among participants, giving the impression that she is just as boisterous as the rest of the group. Whether all combinations of participants can play effectively and have fun under such a model is unclear.

4.2 Wizard's Model (M_W)

In constrast to M_B's purely rule-based design, M_W is purely data-driven, based on the turn-taking decisions made by the wizard. We created a Support Vector Machine (SVM) binary classifier [8] using the features described in Table 1 and take-or-wait labels derived from the wizard's actions in the log files. Because turn-taking within a group unfolds over time, we explored a set of models with access to group information over varying amounts of past history. Each SVM in the set was trained on an extended feature vector that included the features for all the participants in the group in the current time slice as well as the features for all participants from some number of previous time slices. Performance plateaued after four time slices, thus M_W was trained using the LibSVM implementation [6] with a Radial Basis Function (RBF) kernel, mapping features for every participant over a two second history to one of the two outputs: *take* or *wait*. Because Edith's speech and actions are not currently interruptible, time slices during which Edith executed her turn were not used during training or subsequent tests. In total, M_W was trained at 13,067 decision points.

The wizard was able to create a turn-taking environment in which children were effective and had fun, so M_W holds some promise that it can reach the same goal by encoding the wizard's behavior. It may fail to meet that goal, however, for two reasons. First, the model may be unable to capture the inherent motivations behind the wizard's turn-taking decisions because the feature set is inadequate. It does not, for example, encode how many turns each child has taken at a given point in the session, a factor the wizard might have used in deciding to wait and see if delay would encourage a younger child to participate. The list of such possible features is long; our intent here is less to argue that our feature set is complete than to explore whether a set that has been well-documented as important in adults will be useful for children as well.

The second limitation of the model is more problematic. The wizard had the relatively complex task of mapping the behavior of the group–the shouting, the movement, the affect, etc.–into a single interface event. As a result, his response time was variable and often crossed multiple 500 msec boundaries. Although we can tell from the log files when Edith acted, we cannot tell when the wizard formed the intent to take the turn; but it is exactly the features that existed at the moment of intent that the model should associate with the *take*. Features that exist one or two seconds later may be quite different, particularly in a chaotic environment. We can see the negative effect of the lag between intent and action by measuring the model's internal consistency with K-fold cross-validation, testing performance in each session with a model trained on the other 28. When we compare the predictions of M_W against the wizard's actual data, performance is poor: Max F1 = 0.40 and the Area Under ROC Curve (AUC) = 0.51.

Our final model was designed to overcome the inherent lag in the wizard's response and give a clearer picture of the adequacy of the feature set for turn-taking in small groups of children.

4.3 Annotators' Model (M_A)

M_A, the annotators' model, reflects a hybrid approach, sitting somewhere between the purely rule-based (M_B) and purely data-driven (M_W) ends of the spectrum. In essence, we asked a set of annotators to behave as wizards in a version of Robo Fashion World where time could be stopped at theory-driven moments and the only interface option was to answer the question: should Edith take the turn now?

In particular, annotators watched brief segments taken from the choice cycle portions of all sessions, with video of the children from the front camera shown side by side with what the children saw on the display. The annotators

were able to watch the segment only once and were directed to answer immediately the forced-choice question that appeared when the video stopped. The end points of the segments were selected from theory-driven moments for taking the turn and waiting, and were based on the boundaries in the original hand annotations, *not* on the synchronization boundaries that define time slices for the models. In particular, annotators were questioned at decision points that were generated in the following manner:

- **End of Utterances (EOU)**: a decision point was generated at the end of every child's utterance plus 50 milliseconds (to eliminate the perception that the utterance was clipped). There were 2,015 segments of this type.
- **Middle of Utterances (¬EOU)**: to build an accurate model, both positive and negative instances are needed in the training set. Thus, we included video segments for which the expected turn-taking decision would be *wait* by randomly selecting one time slice in the middle of every utterance that was longer than 500 msecs such that the ¬EOU interval did not co-occur with an EOU from another utterance. There were 1,642 decision points of this type.
- **Silences (SIL)**: for each silence larger than one second (the average silence length across all sessions), we randomly selected one point in the first half of the silence and another point in the second half of the silence. The task contained 1,292 video segments of this type.

The 4,949 video segments described above always began immediately after a board change and ended at one of the generated decision points. Four annotators with prior experience in human behavior analysis, three female and one male, were recruited for this task. Video segments were distributed so that each was rated by two annotators and each pair of annotators had the same number of segments in common. There was no contact among annotators after training; inter-annotator reliability across the full data set was significant ($\kappa = 0.739, p < 0.001$).

As we did with M_W, we can ask how well a model trained with the annotators' data captures their consistency. To train M_A we used the same implementation of LibSVM, the time slices during which annotators made their turn-taking decisions (two decisions per time slice), and a history of two seconds. In 29-fold cross-validation against the annotators' ground truth, M_A faired considerably better than M_W: Max F1 = 0.77 and AUC = 0.58.[1]

In summary, then, M_B implements a policy for turn-taking that is completely consistent with respect to the feature space but which may or may not correspond to what any human would do when hosting the game. We expect that if M_B is implemented with human levels of feature detection it would be adequate to keep the game moving without turn-taking inappropriately or adding to the general level of chaos, but may or may not result in play that is fun. M_W is intended to implement the decision making of the wizard who performed the sensing functions in our sessions. M_W is based on behavior that allowed a fun and effective experience but does not seem to reflect that behavior accurately because of the variable lag introduced by the wizard's reaction time. Finally, M_A is an attempt to capture what the wizard would have done if he had really been acting as the host of the game, rather than merely as a set of sensors limited to making categorical judgments through an interface. M_A is trained on less human data overall, but specifically on those moments where turn-taking decisions are likely to be made. In the next section we compare and evaluate the models' behaviors.

[1]We also ran a model using only the decision points where annotators agreed, resulting in: Max F1 = 0.82 and AUC = 0.67. We chose to continue working with the larger model (all annotator points) because it more accurately reflects the annotators' indications that there are, in fact, moments where either choice is acceptable.

5. COMPARING THE MODELS

To choose a single model to control Edith's turn-taking, we use the set of 13,067 time slices from our 29 sessions as a common basis for comparison. This data set represents all the time slices in which Edith does not hold the turn or, alternatively, all the time slices in which the wizard had to decide. To make the comparison, the feature vectors for each participant at each time slice are mapped to a take-or-wait decision for each model. Effectively, this procedure asks each model, "If the character is placed at a moment in time in which this set of features is available to it, what will it do?"

5.1 Analyzing the Models' Behaviors

We generate a full set of decisions for M_W after training on all the intervals using two seconds of history. The resulting behavior of the model can be understood as a straightforward generalization of the wizard's behavior. Similarly, we generate M_A's decisions after training on all the annotators' data with two seconds of history. Since M_A is trained on annotators' decisions for a carefully selected third of the intervals, it generalizes its basis more broadly than M_W.

Generating decisions for M_B also requires some generalization. A straightforward application of the algorithm is problematic because it would generate a *wait* during every interval between the time M_B would *take* and the wizard did take the turn. In Figure 2, for example, M_B would take in timestamp 9 but be forced to wait in timestamp 10. The role of M_B, as a baseline, is to respond unambiguously to the first person addressing the character in the choice cycle by taking the turn immediately after a character-directed utterance, regardless of whether there is another person talking. Since a one second delay is still considered an "immediate" response [9], we generalize this intent by allowing M_B to take the turn with a delay of up to one second (two timestamps), but only if nobody else is talking during that time.

We analyze the performance of the three models as a function of the context in which a *take* can occur. The synchronization boundary of every time slice falls into one of four categories, examples of which are given in the Category band of Figure 2:

1. **Non-chaotic end-of-utterance (EOU&¬CHAO)**: the interval follows the end of an utterance with no overlapping speech (timestamp 9). A *take* at the end of an EOU&¬CHAO interval will give the appearance of the character smoothly taking the turn.

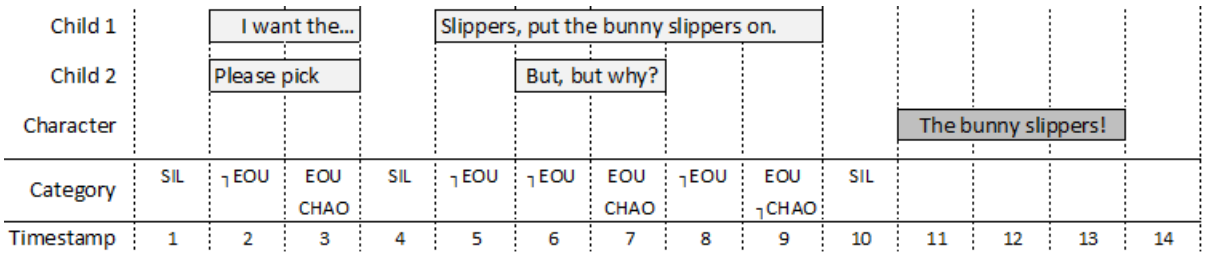

Figure 2: Participant and character behavior mapped to time slices.

2. **Chaotic end-of-utterance (EOU&CHAO)**: the interval boundary follows an end of utterance and contains on-going speech from another participant (timestamps 3 and 7). A *take* at the end of an EOU&CHAO might be perceived as grabbing the turn, interrupting, or boisterous, depending on context.

3. **Non-end-of-utterance (¬EOU)**: one or more people are in the middle of speaking. A ¬EOU *take* will give the appearance of interrupting if only one person is speaking (timestamps 5 and 8), but may seem less rude if multiple people are speaking (timestamps 2 and 6) because the person who technically holds the floor (the one who began speaking first) has already been interrupted.

4. **Silence (SIL)**: no speech during the interval (timestamps 1, 4, and 10). A *take* during silence can be associated with a reprompt by the character, which would either be perceived as helpful or impatient, depending on context.

Figure 3 contrasts the behavior of the models in terms of the percentage of takes in each of the categories, showing clear differences in turn-taking style.

EOU&¬CHAO: M_A and M_W are about equally likely to take the turn at the end of a non-chaotic utterance and more likely to do so than M_B. Recall that M_B always takes the turn after an utterance that the feature annotators said was character-directed. Therefore the extra *takes* by M_A and M_W must occur after utterances that the feature annotators labeled non-character-directed. Most non-character-directed utterances fall into one of two classes: evaluative/emotional comments such as "That's funny looking," which are directed to everyone (or to oneself but spoken aloud), and side conversation utterances about whose turn it is or what a board object should be called. In a previous data set [15] we found that even humans may disagree about

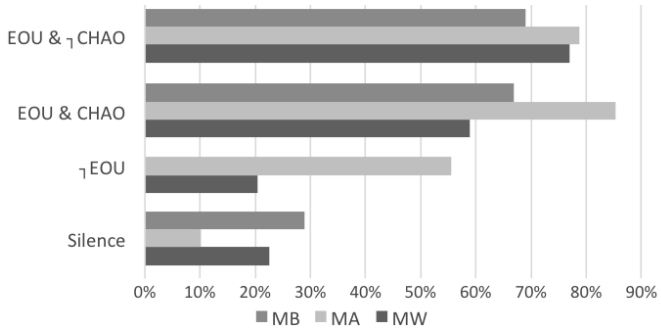

Figure 3: The percentage of takes by each model as a function of category.

addressee for evaluative comments. So both the wizard and the annotators whose judgments created M_A may simply have disagreed with the "ground truth" addressee feature for some of the evaluative comments.

The side conversation utterances are more problematic. In particular, naming conversations involve the use of terms that may be meaningful references to objects on the board. Where situational attractors like the board demand it, participants' body language at these moments may be indistinguishable from their body language during true requests for a board change. As M_W and M_A generalize across the feature space, there will be times that they incorrectly view some of these side utterances as requests, and take the turn to change the board. If this happens often enough, Edith may be viewed less as an engaging character and more as a computer. Alternatively, participants might compensate for Edith's inappropriate *takes* by adopting behaviors that are natural–lowering the voice, speaking behind a hand, or turning away from the screen–and lead to more successful play.

EOU&CHAO: The difference between this category and the previous one is that someone is still talking when an end of utterance occurs. Again, M_B takes if the utterance is character-directed, responding promptly to one person even though it means interrupting another. So the extra *takes* by M_A can be understood by the analysis given above. The behavior of M_W has changed, however; M_W is not taking the turn at the end of all character-directed utterances. It is not possible to know whether M_W takes less often because the wizard could not react quickly enough to take in a chaotic moment, or whether the wizard simply chose to let the chaos die down before making a decision. It is clear, however, that the annotators whose data created M_A were not willing to wait; they (and their model) are aggressive in chaotic circumstances, choosing either to respond to the request or to manage the chaos immediately.

¬EOU: Because letting the speaker finish is considered to be a basic rule of turn-taking, M_B never interrupts in the middle of an utterance unless it must respond to a direct request. As the figure shows, neither M_W nor M_A follows the rule, with M_A, again, taking the most aggressively. Some of M_W's *takes* are likely caused by overgeneralization of delayed *takes* in other circumstances. To understand the remainder of M_W's *takes* and all of M_A's *takes* in this category, consider timestamps 6 and 8 of Figure 2. At 6, one person has begun to talk when another chimes in; if the video is stopped here, a *take* can occur if the wizard or annotator wants to warn the children not to talk at once. At 8, only one person is speaking, but that person has already said enough to indicate his or her choice. The annotators tended to take the turn at such points, and M_A generalized

	Percent of agreement			Weighted sum
	Judge1	Judge2	Judge3	
MA	64%	61%	53%	+955 pts
MW	62%	52%	60%	+785 pts
MB	40%	39%	59%	-685 pts

Table 2: Judges' agreement with the models, with and without strength as a factor.

that tendency to other situations in which an item reference occurred early in the utterance.

Silences: The three models behave identically in the first second of silence–taking as if those intervals marked EOUs–but behave quite differently after that. In contrast to its behavior elsewhere, M_A almost never jumps in during silences longer than one second. This suggests that the annotators were willing to wait as long as it took for a child to take the turn or an adult in the group to prompt a child. M_W often takes the turn after the first second, but the number of takes decreases as the length of the silence increases, with only a few *takes* after 3.5 secs. This suggests that the times M_W does *take* may result from lags in the wizard's reaction time at the end of an utterance. Finally, M_B takes turns in all silences longer than 3.5 secs as suggested by [3].

Figure 3 makes it clear that the models perform quite differently. Indeed, the three models agree about what to do in only about 60% of the time slices. What is unclear is whether one of them performs better.

5.2 Which is better?

To investigate this question, we conducted a user study to compare turn-taking decisions at points where one of the models disagrees with the others. Our goal was to discover whether one model would gather a clear consensus.

5.2.1 Procedure

About 20% of the time slices where the turn-taking decisions differ were randomly selected for this study. In an interface similar to the one used to collect the data for M_A, a new group of three judges (two female and one male) watched each of 1010 video segments. Each segment started at the beginning of a choice cycle and ended at the 500 msec boundary where the models disagreed. Annotators who contributed to M_A made a simple yes/no decision after the question "Should Edith take the turn now?" Judges contributing to the evaluation were given the same prompt but asked instead to choose among five options: (1) *definitely take the turn*, (2) *take the turn*, (3) *either one is OK*, (4) *wait* or (5) *definitely wait*. The scale was used both to represent the judges' confidence in their decisions and to exclude from the analysis ambiguous moments in which individuals, themselves, could not decide.

All judges used the five-point scale with similar distributions of responses and few *either-one-is-OK* values, suggesting that all were comfortable with the distinctions being made and none was much more or less confident in his/her decisions than the others.

5.2.2 Ranking the Models

The first three columns of Table 2 present the percentage of agreement between the outcomes of the models and the decisions of the judges. Note that no model garnered overwhelming consensus, and no judge agreed with any one model's decisions more than two-thirds of the time. Judges 1 and 2 had similar profiles, preferring both M_A and M_W over M_B, while Annotator 3 had no clear preference.

To factor in strength of agreement, we ranked each model according to the following criteria: if the model's prediction was in line with the turn-taking decision provided by the judge, the model received +2 or +1 point if the answer was, respectively, a *definitely take/wait* or a *take/wait*. If there was a mismatch between the model's prediction and the judge's response, the model received a penalty of -2 or -1 following the same rule. If the judge's rating was *either-one-is-OK*, the response was ignored. The weighted sum column of Table 2 shows that with respect to the decisions where the models differed, the behavior of the two SVM models was preferred to the rule-based model, with M_A somewhat preferred to M_W.

Since the behavior of the models differs as a function of the contexts explained above, we also considered whether judges were agreeing or disagreeing with taking or waiting in each context. Table 3 summarizes the results. Because we sampled randomly from the disagreements, each category is represented according to its prevalence in the full set: 56 points from EOU&¬CHAO, 30 from EOU&CHAO, 543 ¬EOU, and 381 silences.

In end of utterance intervals, **EOU**, the judges do not agree with most of the extra takes by M_W and M_A, favoring the rule-based model, which waits at the end of non-character-directed utterances. It is important to stress, however, that M_B's decisions rely on the hand annotations for the utterance boundaries, and that automatic prediction of end-of-utterance (or end-of-turn) is still a very hard problem, especially in multiparty settings [10]. The SVM models are essentially learning to detect the end of an utterance from the presence of other features, particulary features in the two seconds of history. It is unclear whether an implemented version of M_B–one that could not rely on human levels of sensing–would retain its advantage in this category.

With respect to **¬EOUs**, the judges disagree with M_B's decision to wait during the disputed segments. Both M_W and M_A received positive points for taking the turn in the middle of the utterances; M_A jumped in more often, so it received the most points. Like the group of annotators whose data informed M_A, these judges are willing to take the turn as soon as they hear a valid reference to a board item, even if that means interrupting the speaker.

Finally, the judges are quite clear in their dislike for takes that occur in **Silences** greater than one second. M_A fares best since it has the fewest of these. M_B–which always takes the turn after silences longer than 3.5 seconds, as suggested for adults–is clearly the wrong behavior when interacting with children in our environment. There is, no doubt, a "magic number" after which even these judges would consider a silence to have gone on too long, but in our game that moment rarely occurred.

	M_W		M_A		M_B	
	T	**W**	**T**	**W**	**T**	**W**
EOU&¬CHAO	-39	29	-58	10	-3	65
EOU&CHAO	-20	-6	-11	3	23	37
¬EOU	215	-22	234	-3	0	-237
SIL	-230	858	-154	934	-829	259

Table 3: Weighted sums for each models *take* (T) or *wait* (W) decisions as a function of category.

The judges overall preference for M_A is interesting for three reasons. First, it sanctions aggressive interrupting by the character in order to minimize stretches of chaos and keep the game moving with rapid changes to the visual display. Second, it tolerates stretches of silence up to at least five seconds without acting. Both of these characteristics tacitly acknowledge that the rules for language interactions with small groups of children are not simply some trivial extension of the rules for interacting with small groups of adults. The third reason the preference for M_A is interesting is that M_A had to generalize considerable more from its training data than M_W did. This suggests that focusing on the moments where the wizard would have, in theory, formed the intent to take or wait allowed the model to learn a consistent and appropriate set of audio and visual features. Whether those features can be sensed autonomously with the necessary accuracy to preserve M_A's behavior and whether the behavior, itself, will be fun and engaging for the children remains to be seen.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we addressed the issue of when an interactive character should take the turn in game play with small groups of children, with or without adults. We investigated three different approaches to making take-or-wait decisions based on multimodal features that encode the group's behavior over time. The three approaches resulted in three distinct turn-taking models: a rule-based model (M_B), a data-driven model trained with the turn-taking decisions made by the wizard during game play (M_W), and a hybrid model (M_A) based on annotators' post hoc decisions at theory-driven moments in the video record. Cross-validation of the two machine-learning models with respect to their own ground truths revealed that the annotators (and M_A) behaved more consistently than the wizard (and M_W). The most likely explanation for the poor performance of M_W is the variable delays between when the wizard made his decision to take the turn and when he signaled that decision in the interface. In an evaluation of the moments where the models differed, human judges preferred the behavior of the SVM-based models to the baseline.

Despite the similarities in the rankings of M_A and M_W, the turn-taking behavior of these models is quite different. Under M_A, Edith would take the turn aggressively, often interrupting a speaking child as soon as the child's game choice becomes clear. Under M_W, Edith would seem more polite and willing to wait for children to finish speaking before responding. Our next step is to implement the models in an autonomous version of the character to test whether children have as much fun with either model as they had when the wizard was leading the interactions. We are also planning to apply the same methodology to build a complementary *hold/release* model, so that Edith is able to stop an ongoing action and release the turn to the other participants if conditions suddenly change.

7. REFERENCES

[1] I. Bakx, K. Van Turnhout, and J. Terken. Facial orientation during multi-party interaction with information kiosks. In *INTERACT*, pages 163–170, 2003.

[2] M. Blomberg, G. Skantze, S. Al Moubayed, J. Gustafson, J. Beskow, and B. Granström. Children and adults in dialogue with the robot head furhat-corpus collection and initial analysis. In *WOCCI*, 2012.

[3] D. Bohus and E. Horvitz. Decisions about turns in multiparty conversation: from perception to action. In *ICMI'11*, pages 153–160. ACM, 2011.

[4] F. Broz, C. L. Nehaniv, H. Kose-Bagci, and K. Dautenhahn. Interaction histories and short term memory: Enactive development of turn-taking behaviors in a childlike humanoid robot. *CoRR*, abs/1202.5600, 2012.

[5] J. Cassell and K. R. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied AI*, 13(4-5):519–538, 1999.

[6] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[7] C. Chao and A. L. Thomaz. Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *J. of Human-Robot Interaction*, 1(1), 2012.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] J. Dabrowski and E. V. Munson. 40 years of searching for the best computer system response time. *Interacting with Computers*, 23(5):555 – 564, 2011.

[10] I. de Kok and D. Heylen. Multimodal end-of-turn prediction in multi-party meetings. In *ICMI-MLMI '09*, pages 91–98, 2009.

[11] S. Duncan. Some signals and rules for taking speaking turns in conversations. *J. of personality and social psychology*, 23(2):283–292, 1972.

[12] S. Ervin-Tripp. Children's verbal turn-taking. *Developmental pragmatics*, pages 391–414, 1979.

[13] C. Garvey and G. Berninger. Timing and turn taking in children's conversations. *Discourse Processes*, 4(1):27–57, 1981.

[14] J. Gustafson and K. Sjölander. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In *Proc. of ICSLP/Interspeech*, pages 297–300, 2002.

[15] H. Hajishirzi, J. F. Lehman, and J. K. Hodgins. Using group history to identify character-directed utterances in multi-child interactions. In *SIGDIAL*, pages 207–216, 2012.

[16] Y. Matsusaka, M. Enomoto, and Y. Den. Simultaneous prediction of dialog acts and address types in three-party conversations. In *ICMI '07*, pages 66–73, 2007.

[17] Y. Nakano and Y. Fukuhara. Estimating conversational dominance in multiparty interaction. In *ICMI '12*, pages 77–84, 2012.

[18] A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *NAACL'09*, pages 629–637. Association for Computational Linguistics, 2009.

[19] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.

[20] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa. Learning decision trees to determine turn-taking by spoken dialogue systems. In *INTERSPEECH*, 2002.

[21] E. A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29:1–63, 2000.

[22] E. O. Selfridge and P. A. Heeman. Importance-driven turn-bidding for spoken dialogue systems. In *ACL*, pages 177–185, 2010.

[23] K. Thórisson, O. Gislason, G. Jonsdottir, and H. Thórisson. A multiparty multimodal architecture for realtime turntaking. In *Intelligent Virtual Agents*, volume 6356 of *LNCS*, pages 350–356. Springer, 2010.

[24] G. Tur et al. The calo meeting assistant system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1601–1611, 2010.