Robo Fashion World: A Multimodal Corpus of Multi-child Human-Computer Interaction

Jill Fain Lehman
Disney Research, Pittsburgh
4720 Forbes Avenue, Suite 110
Pittsburgh, PA 15217, USA
1-412-688-3405
jill.lehman@disneyresearch.com

ABSTRACT
We present a retrospective view on our experience with small groups of more than 175 children (ages 4 to 10) playing versions of a language-based game hosted by an animated character. After describing the task, the audio-visual annotations used for modeling, and the regularities we see in the children's communicative behavior, we conclude with observations and challenges for other researchers interested in autonomous HCI and HRI with this population.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – Natural Language; D.2.2 [**Software Engineering**] Design Tools and Techniques – State diagrams; D.2.11 [**Software Engineering**] Software Architectures – Languages

General Terms
Human Factors, Design

Keywords
Multiparty interaction; Child-Character dialog; Spoken dialog; Multimodal interaction; Addressee identification.

1. INTRODUCTION
The work described here is part of a larger effort on language-based character interaction with children. Although the overall project's goal is to create and evaluate audio-visual technologies that support a wide variety of engaging scenarios, the focus in this work is to explore small group multimodal interaction in the context of a game that might, for example, provide an amusing diversion during wait times in an attraction line or hotel lobby. Despite the surface characteristics of the task, we believe our data and observations have wider applicability; the behaviors we discuss can arise whenever groups of children interact with an autonomous conversational agent that controls access to a resource, including access to a "correct" answer in an educational environment. Our game is one instance of such an interaction, but collaborative problem solving, group work in early education, and game-like environments for learning pronunciation or vocabulary in second-language curricula seem likely to produce others.

In the remainder of this paper we describe our task environment, Robo Fashion World (RFW), as well the data collected over the past three years. We continue by analyzing the children's behavior in terms of various human-labeled and machine-labeled audio-visual features. We conclude by raising a number of issues that present challenges to researchers interested in multi-child interaction.

ICMI'14, November, 2014, Istanbul, Turkey.

2. DATA COLLECTION AND CODING

2.1 Robo Fashion World
RFW is a game in which children dress up an on-screen fashion model by choosing from visually-available clothing items and silly accessories (see Figure 1). Edith, the animated robot character at the left of the figure, hosts the basic version of the game by mediating the interaction and making the costume changes.[1] 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.

During the data collections, a human performed all of Edith's language understanding tasks in a Wizard-of-Oz design. The 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 is unclear, a silence that is too long, or multiple voices speaking at once. Edith then autonomously selected a sequence of actions to perform. In the case of a clear choice of accessory, that sequence always followed the same pattern: she acknowledged the choice, pushed the orange button to dress the character with the selected item, and then passed the turn back to the group with words and/or gestures. Children could, and often did, call out different requests at the same time. The wizard's decision about which event to signal (or to wait for a clear answer) was based on the general criteria that *the game should move along effectively and be fun*, implicitly defining Edith's turn-taking style.

In addition to the basic mode of play described above, sessions in 2012 and 2013 included a second host, Charlotte, who looked similar to Edith except for the color of her clothing and the shape

[1] RFW grew out of an earlier game called Mix-and-Match [8]. Although the basic game play, choosing a visually-available item, was the same, we do not include data from the earlier study here because the host character, narrative of the interaction and surface characteristics of the environment were entirely different.

of her silhouette.[2] Those sessions also included a set of more flexible behaviors for the host when releasing the turn, in an effort to use proxemics, gesture and voice to effect more balanced turn-taking among the children. Groups in those years played two games, one with each physical host in either the basic or flexible conditions [1].

Figure 1: Edith hosts Robo Fashion World

2.2 Participants

The data collected from 2011 to 2013 represents 177 children (ages 4 to 10), eight parents, and three experimenters. Parents who joined the children in play were volunteers instructed to support their children in whatever way felt natural in their families. Groups were formed based on scheduling availability and, therefore, were as likely to combine strangers as they were to include family members or friends. Children were compensated for their participation.

Just as there were some systematic differences in the game environment over the years, there were also some systematic differences in the participant populations. In 2011, 29 games were played by 28 unique groups, 24 of which contained an adult or experimenter. In 2012, ten unique groups played two games each, with only one group containing a parent and no experimenters involved in play. In 2013, all 22 groups were comprised of children and all groups played twice. The general trend across time, then, was to increase the average group size and reduce the involvement of adults.

Additional statistics for the children are given in Table 1. The mean age and division of genders across the years was fairly steady, although 2013 had a slight bias for older boys. Distribution of children across age groups (not shown) was also fairly even, with an average of 25 children per age overall. Nevertheless, because the largest numbers of games and children came from 2013, their data tends to dominate the results.

Table 1. RFW participation, children only

Year	# of Games	# of Participants (Average #/group)	Mean (SD) Age; % female
2011	29	65 (2.4)	6.8 (1.9); 52%
2012	20	35 (3.3)	6.8 (1.9); 51%
2013	44	77 (3.5)	7.0 (2.1); 43%
All	93	177 (3.1)	6.9 (1.9); 48%

2.3 Procedure

RFW was one of a number of activities in which the children participated. During the consent process children were able to meet each other, play together for a few minutes, decorate their badges with stickers, and get fitted with microphones. Children might then participate in individual activities before coming back together and being escorted as a group to RFW.

In the room, children arranged themselves side-by-side behind a line marked by multi-colored segments on the floor. Their attention was drawn to the line and children were asked to name their segment colors to be sure they knew them in case Edith referred to them directly ("You, on the red line…"). After this step, the experimenter signaled to the wizard to start RFW. A game lasted an average of eight minutes.

Participants stood approximately six feet away from the large screen where the game was displayed. Sessions were video recorded using one frontal and two lateral cameras. Sound was captured using individual close-talk microphones. In the 2012 and 2013 sessions, Kinect depth data was also recorded in order to support anticipated work on gesture recognition.

2.4 Annotating the Data

Our analysis of behavior is fundamentally influenced by the goal of evaluating and creating technology to support interaction between a character and multiple children. We are specifically interested in uncovering regularities in the behavioral data that can be used to replace Edith's wizarded language capability with a fully-autonomous one. Minimally, autonomy requires that the character know: (1) when it is and isn't being addressed, (2) if addressed, both what is being said and what is meant, and (3) when it is appropriate to take the turn and act. In previous work building Support Vector Machine (SVM) models for addressee identification [8] and turn-taking [9] in RFW and similar activities, we have found the following annotations to be particularly useful:

Speech: who is speaking, start and end times, and content are transcribed by hand using both audio and video, segmented into utterances based on 50 msec silences.

Partner: a binary form of addressee for transcribed utterances is labeled by hand, based on both audio and video. Each utterance is designated as directed to the host character (**char**) or not to the host character (**nchar**).

Power and *pitch*: are generated automatically. Pitch is calculated using the harmonic product spectrum. For both, values are calculated over 500 msec windows and nominalized to **high**,

[2] For simplicity, we will to refer to the host as Edith throughout, although it should be understood that unless otherwise specified, the actual host could have had either physical form.

medium, **low**, or **none** based on +/- one order of magnitude with respect to the mean across the child's entire sound file.

Character prompt: a **true**/**false** value is associated with each sound file in Edith's repertoire and mapped automatically to a *Speech* segment for Edith based on start and end times in the log files generated during the game. Prompts may be interrogatives ("I have never heard of this thing, darling, does it have another name?") or comments that mark the end of Edith's turn, after changing the board and turning to face the group ("What an eye you have!").

Head orientation: segments are labeled separately for **turn-away** and **turn-back**, by hand, based on both audio and video. Annotators are told to use **turn-away** only when the turn is associated with a meaningful attention shift to a person or object, and not for brief, incidental head movements. The **turn-away** includes the time looking away from Edith (and the board) and continues until the beginning of the continuous head movement that ends in head orientation and attention being back on the screen. The time from the beginning of the head turn that ends at the screen until orientation to the screen is complete constitutes the **turn-back**.

Gesture: segments are labeled as **pointing**, **headshake-no**, **headshake-yes**, **clapping** and **emphasis,** by hand, based on both the audio and video. **Emphasis** gestures are defined as hand or arm movements toward the screen that are neither **pointing** nor part of grooming motions.

For features that were created by hand, a second annotator coded approximately one quarter of the data using a rolling window across the sessions to make sure there were representative segments from each stage of the game. For *Speech* annotations, the second annotator coded only the presence of the utterance, not a duplicate transcription. To check reliability, data was aligned across coders such that (1) segments were aligned from start to finish in the session, (2) each segment from the second coder could align with at most one segment from the primary coder, and (3) two segments were considered aligned if the larger overlapped more than half of the smaller.

With respect to *Speech*, coders agreed about the presence of an utterance about 92% of the time; for aligned utterances, they agreed on *Partner* 95% of the time ($\kappa = .87$). Disagreements were predominantly for short utterances (e.g., fillers like "um" as a child paused for thought) where none of content, context, or body language strongly indicated addressee.

Coders had 84% of their *Head Orientation* segments in common, with 94% agreement on the label of the aligned segments ($\kappa = .90$). It seems clear that, in this case, the residual error is due to poor alignments or human error in using the pull-down menu of the coding tool rather than actual misjudgments about the direction of the turn.

Gesture had the lowest percentage of segments in common (79%), but high agreement for the labels of aligned gestures (90% of the time, $\kappa = .83$). Most of the disagreement was with respect to the coding of **emphasis**—both whether an emphatic gesture occurred (as opposed to a grooming motion, for example) and whether an aligned segment was actually an emphatic gesture or a point.

Finally, we note that the degree of agreement for each category is nearly identical to that category's statistic for the earlier data from Mix-and-Match, which used the same coding scheme with a different set of coders and a different multi-party interaction [8]. The consistent reproduction of generally high kappa values across tasks suggests that the coding criteria were well-specified and followed with reasonable accuracy. As a result, and despite the somewhat subjective criteria used for **turn-away** and **emphasis**, regularities in the data are likely to be a function of the children's behavior rather than an expression of the primary coder's potentially idiosyncratic judgment.

3. COMMUNICATION BEHAVIOR

In the previous section we enumerated three basic capabilities without which an autonomous character would be impossible. Here we examine the annotated behavioral data for each capability in turn: how children signal that they are addressing the host, what they say to indicate what they want, and how they manage turn-taking within the larger group that includes the character.

3.1 Signaling Edith as Addressee

The RFW corpus contains 9597 utterances, of which 9039 (94%) were spoken by children. When they participated, parents and experimenters acted largely in support roles, with 68% of their utterances labeled as **nchar**. In contrast, most of the children's utterances were directed to the host (6399/9039, 71%). Although six and seven year old participants addressed Edith somewhat more often than the others (79% and 76% of the time, respectively, versus 66% to 69% of the time for other ages), to whom a child spoke did not vary significantly with age.

While the age of the child cannot be used to predict the likelihood that an utterance is **nchar**, the duration of the utterance can. Utterances addressed to Edith were significantly longer, averaging 1308 msec, compared to 1107 msec for **nchar** (*Student-t*, $p << .001$).

Knowing that we would be unable to detect eye gaze at a distance, we arranged participants side-by-side to force a change in head orientation to make eye contact. This attempt to engineer the environment to create detectable **turn-away** and **turn-back** events at the beginning and end of side conversations was not entirely successful because the game board was too strong a "situational attractor" [3]. Speech to other participants co-occurred with a head turn only a third of the time, and adults were more than twice as likely to make the effort (57% overlapping head turns for **nchar** in adults versus 27% in children). Nevertheless, when a **turn-away** did overlap with one's own speech, it was almost three times more likely to overlap an **nchar** utterance.

Adults and children also differed with respect to their gestures. Children did most of the gesturing (80%) when both types of participants were in the group. Adults had virtually no **emphasis** gestures and, not surprisingly, when their **pointing** gestures occurred in the context of self-speech, the utterances were directed to the child rather than the character three quarters of the time. In contrast, children used **pointing** and **emphasis** gestures almost equally, and when either gesture overlapped with self-speech, the context was an utterance directed to the host three quarters of the time.

We can combine the gesture and head orientation features with the automatically generated features described in the previous section to learn an SVM model that categorizes pre-segmented utterances as **char** or **nchar**. Doing so in a naïve fashion produces an overall accuracy of 73%, with correct categorization of **char** utterances 77% of the time and correct categorization of **nchar** utterances

63% of the time.[3] We can increase this performance—which isn't a lot better than simply assigning all utterances to **char**—by adding both the speaker's history and the group history information to the feature vector (as in [8]). History features—i.e., feature vectors from prior time slices—capture the dynamic relationships between verbal and non-verbal cues that co-occur across time and participants in conversation. With four seconds of the speaker's history and one second of history from the other members of the group, overall accuracy reaches about 79% (85% for **char** and 64% for **nchar**). Because these results are based in part on human-coded data, they represent an upper bar on accuracy for the given feature set. In other words, if we can replace the human annotations with equally accurate machine-sensed values, then the feature set itself captures a significant amount, but by no means all, of the regularity available in the data. Human coders agreed 95% of the time about who was being addressed, but they were undoubtedly using the content of each utterance in their decision-making in addition to the features and context captured in our models.

3.2 What is Said and Meant

Based on the transcribed segments, we divided utterances into five categories: item references, picture requests, emotional-evaluative content ("that's my favorite" "he looks kinda weird with one eye"), unintelligible, and other ("yes" and "no" responses, fillers ("um," "uh") and incomplete phrases that could not be resolved as one of the other categories).

As shown in Figure 2, the distribution of each type of utterance differed as a function of *Partner*. Item and picture references dominated speech to Edith. Emotional and evaluative comments almost always indicated conversations among participants. Note that although item and picture references were not the most prevalent type of **nchar** utterance, they still constituted more than 20% of those utterances overall.

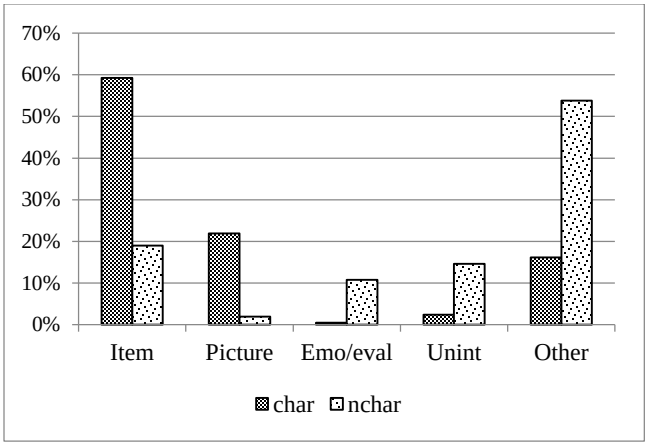

Figure 2: Distribution of utterance types as a function of *Partner* (%s are relative to the total for each kind of *Partner*)

We further distinguished between three types of item references: in-vocabulary, weak naming strategies, and out-of-vocabulary. In-vocabulary references constituted the largest group overall (4113/4471, 92%) as well as the largest group for each type of *Partner*. To be considered in-vocabulary, the utterance had to contain one of the words given to the illustrator to describe the item (e.g., "alien hands," "bunny slippers," "hairy monster feet") or an equivalent term from WordNet [5] that was likely to be known by young children. Out-of-vocabulary references were the next largest set: 6% overall and 6% of **char**, but only 4% of **nchar** references. The remainder—weak naming strategies that referred to board items by deictic expression ("that one"), by position ("now on the top left"), or by physical attribute of the icon ("the green one," "the gold thingy")—constituted only 2% of all item references overall. However, weak naming utterances were more prevalent than out-of-vocabulary references when participants spoke to each other (10% versus 4% of **nchar**) because children used such strategies to elicit the appropriate names for visually unclear or lexically unknown accessories.

Although the wizard was free to accept any understandable reference to a board item, an autonomous host must recognize speech based on a pre-determined lexicon and language model. **Char** and **nchar** utterances each had vocabularies of less than 900 words, with about half their words in common. Frequency histograms show predictable curves: about 35% of each set's words occur only once. In **char** utterances three quarters of the words occur ten times or fewer; **nchar** utterances had even less repetition, with about 82% of the vocabulary occurring less than or equal to ten times.

In general, words that defined in-vocabulary item references were more prevalent in the **char** utterance set than the **nchar** set. The words "take" and "picture" were particularly predictive, being the third and fourth most prevalent words in **char** utterances (after "a" and "the") but 25th and 40th in the **nchar** list. Similarly, there was a small set of words that occurred often in **nchar** utterances and rarely in **char**. This set of *social words*—"he," "she," "it," "its," "we," "your," "his," "her," "my," "what," "what's, "that," "that's," and "this" and their homophones (e.g., "it's")—occurred in less than 2% of **char** utterances but almost a quarter of **nchars**.

In theory, the differences in the distribution of words that signal the different utterance types could be used to significantly improve our models of addressee identification. Currently, this improvement does not play out in practice, however. Despite the relatively small vocabulary, Sphinx performed poorly on the corpus, recovering only 24% of the words correctly.[4] Some of the error is no doubt due to the prevalence of overlapping speech discussed in the next section, but the discrepancy between the acoustics of adults' and children's voices contributes significantly as well [6] [7] [10]. Using the children's utterances to perform unsupervised MLLR adaptation of Sphinx's acoustic model brought word accuracy up to 33%. We then used the fact that most of the utterances in the task are picture references or in-vocabulary item references addressed to Edith to examine the results of a post-processing step that scanned the recognized words in each utterance for that relevant subset of vocabulary. This *functional recognition* still achieved only 63% accuracy in assigning each utterance to **char** or **nchar**; too low for practical use.

[3] The models assume that real-time information from different sensors will be synchronized at 500 msec boundaries. Results reflect classification of 500 msec timeslices using LIBSVM [4] with ten-fold cross-validation and a C-SVC radial basis model with gamma 1/num_features and cost 10.

[4] Sphinx3 Version 8 was used, with an adult acoustic model trained on MFCCs with a frequency range of 120-6800Hz, 8 Gaussians/state. The trigram language model was based on all utterances spoken by at least two children to avoid any bias when running hold-one-out training and testing sets.

Finally, we note that even with further improvement of ASR for picture and item references, the existence of such terms in a non-trivial percentage of **nchar** utterances suggests that using them to predict *Partner* might not be wise. About 20% of **nchar** utterances contained picture/item references (more generally: 10% of all utterances contained picture/item references and were **nchar)**. Absent other indicators (e.g., head orientation, gesture), the 20% are likely to result in false **char** classifications because of the overwhelming preponderance of such terms in **char**. But we know from the analysis above that children did not typically use either head turns or gestures to indicate that they were talking to other participants. Thus, with just functional recognition—that is, without actually understanding what is meant by the utterance in the conversational context—such utterances are likely to be interpreted by an autonomous character as requests for action, misinterpreting the child's intent and usurping the child's turn. Since the game is always 20 turns long and accessory choices cannot be undone and do not repeat, Edith's inability to distinguish between requests and side conversations about items will make her seem incompetent at best, malevolent at worst.

The situation is different with respect to *social words.* Although they are highly predictive of only one kind of *Partner*, misclassifications that are made because of their existence (i.e., false **nchar** decisions) are likely to have little effect. Even if a social word co-occurred with a picture/item reference and was misclassified as **nchar**, the participant's intended choice would simply be ignored, leaving the speaker to try again. Enjoyment would quickly turn to frustration if this event occurred too often, but the combination of picture or item word with social words occurred only 19 times in the 6577 utterances addressed to the character. Still, social words are short, making them difficult to recognize accurately. In addition, the information they carry is largely redundant with the historical information described above; adding a feature for social words increases the overall accuracy by only one percent.

3.3 Taking Turns

As explained in Section 2.1, the wizard implicitly controlled Edith's turn-taking behavior by choosing when and how to respond to participants' utterances. Recall as well that the wizard was instructed to keep the game moving and fun in order to create an engaging experience. The result was that Edith rarely expressed that there were "too many voices" and the participants, themselves, were left to decide how disorganized the turn taking could be. In [9] we used speech overlap to define the idea of a *chaos factor* for each session after noting three general phenomena that resulted from the lack of formal turn-taking structure in the 2011 interactions:[5]

1. *All groups have chaotic moments.* Overlapping speech was ubiquitous; every session had some.
2. *More people means more chaos*. The amount of overlapping speech was significantly correlated with both group size (Pearson $\mathbf{r} = .51$, $p < .01$) and the number of children in the group ($\mathbf{r} = .61$, $p < .01$).
3. *Individuals matter*. Despite the trend in (2), there was considerable variability in the amount of chaos at each group size. One group of four, for example, had overlap in only 15% of utterances, while another group of four had 57%. Similarly, one group of three had extremely orderly turn taking, with overlap in only 6% of utterances, while another group of three topped the 2011 list at 64%.

High chaos groups were also likely to have at least one child who didn't get to make many item choices. Thus the 2012 version of RFW introduced Charlotte and devised additional, more flexible character behaviors to try to enforce a fairer distribution of turns without reducing the fun. A second goal, of course, was to see whether fairer turn taking would also result in less chaos. In that pilot study we found that turn taking was more evenly distributed and the game no less fun with the more flexible host, but that the three chaos phenomena remained.

The data collection in 2013 used the same experimental design as 2012: each group played twice, once with a flexible host and once with an inflexible host, counterbalanced for group size, order of flexibility level, and assignment of behaviors to physical character. As shown in Table 1, this was a larger study than the pilot and we did not find that the fairness result persisted in either the 2013 or combined 2012-2013 data. It is unclear whether the change in outcome was due to the larger impact of individuals given the relatively small number of groups in 2012, the increase in average group size in 2013, or differences in the way the two wizards responded at the interface.

When we combine the data from all three years (61 groups), the pattern that emerges is quite clear and consistent with the earlier characterization: every session had chaotic moments, the more children in a session the higher the chaos factor was likely to be, but some groups thrived on chaos and others did not (the least chaotic group of four children had only 5% overlap while the most chaotic group had 74%). As expected, children's desire to make the next change to the model meant that overlap was more likely in **char** utterances than in **nchar**. Surprisingly, we did not find significantly less overlap in sessions that contained explicit turn management utterances ("It's not your turn" "why don't you pick one") compared with those that did not. Put another way: low chaos sessions were not the result of group members explicitly directing the turn taking.

Because the full data set contained a reasonable number of sessions both with and without adults, we looked at whether there was less chaos in groups that had an adult participant. We found a significant difference (*Student-t*, $p = .03$)—with less chaos in groups with adults and about 10% difference in the means—but caution that the result is confounded by the difference in mean group size between the two conditions as well.

4. REFLECTIONS AND CHALLENGES

Given that current technology does not support unrestricted human conversation, every effort at building autonomous language-based interaction is also an exercise in understanding how to constrain the nature of the task to make the language processing tenable. When we began this project we expected that a small amount of environmental engineering would go a long way. We designed RFW to use visual prompts of kid-friendly objects in order to tightly control the domain of discourse, and were successful in holding out-of-vocabulary references to a bare minimum. Nevertheless, even the small vocabulary relevant to effective game play was beyond ASR for these very young children. Moreover, as a situational attractor, that same game

[5] The values shown here differ from those published previously because in [9] we defined overlap with respect to the initial utterance. In other words, if utterance 2 began while utterance 1 was in progress, utterance 2 overlapped utterance 1. Here we use a slightly different definition in which both utterance 1 and utterance 2 are considered to suffer from overlap, a metric that is more informative for those concerned with ASR.

board was responsible for partially defeating our attempt to force detectable visual cues for addressee by placing participants side by side.

Similarly, we included adults as part of the group initially to add constraint in two ways. First, we assumed that if adults were part of the group Edith would not have to understand clarification requests ("what's that green thing?). But clarification side conversations tend to contain item references, creating a situation in which the words that should have helped determine to whom the child was speaking could not be used for that purpose without the unintended consequence of sometimes usurping the child's turn. This particular problem was exacerbated by the fact that children were even less likely to turn their heads during an **nchar** utterance when an adult was part of the group than when all participants were children. We assume that it is even easier to ignore the pragmatic convention of making eye contact when you can rely on an adult to understand what you mean.

We wanted Edith to be an entertaining and engaging character. So the second constraining influence we hoped adults would provide was to create an implicit expectation of orderly turn-taking, and keep Edith from having to act the part of a disciplinarian to maintain some semblance of order. A study that controls for group size would be needed to determine whether adult presence in the group was the unambiguous cause of the less chaotic sessions. Nevertheless, less chaotic is not non-chaotic; the voices of groups of children who are engaged and having fun tend to collide, and such overlapping speech is a challenge even when processing adult speech [2].

Where does this leave us? Our experience with Robo Fashion World suggests three fruitful avenues for future research. For those who are interested in multimodal, language-based interaction with children per se, the question remains as to what other design decisions and environmental manipulations might help overcome limitations in the technologies that underlie multi-party tasks like ours. Would the convention for eye contact be more or less likely to be met if one of the group members was a robot? Given that positioning the children side-by-side did not guarantee head turns, would other configurations be better? Would it be more reliable to simply assume situational attraction will override head turns for **nchar** utterances, and move Edith far enough away from the game board to try to make **char** utterances force a detectable turn? Is current technology adequate for an autonomous Edith in games like RFW if only two children are involved? If the children are ten to twelve rather than four to ten?

For those interested in educational applications, where a formal approach to turn taking may be justifiable (and enforceable) on pedagogical grounds, what immediately comes to mind is the need for a systematic exploration of the effect of adults as group members. While the current predisposition is to treat educational games as an opportunity for independent learning, it seems possible that more complex tasks or larger groups might be possible with an adult present, even if only in a passive or supportive role.

Finally, to colleagues in the speech community we can say only that children's speech, particularly young children's speech, remains an unsolved problem. Empirical results that outline what complexity of language under what environmental conditions produce what degree of accuracy for what age groups would be a game-changing advance for all.

5. ACKNOWLEDGMENTS

Many people have contributed to the research reported here. Hanna Hajishirzi, Iolanda Leite, and Sean Andrist all made significant contributions to multiple aspects of the project, particularly to the models for addressee identification and turn-taking. Sadhwi Srinivas, Peijin Zhang, Pallavi Baljekar, and Nia Bradley contributed to the testing of and experimentation with the models. Yueran Yuan and Tushar Arora built Robo Fashion World with artwork provided by Sharon Hoosein and Emily So.

6. REFERENCES

[1] Andrist, S., Leite, I., and Lehman, J. 2013. In *Proceedings of the 12th International Conference on Interaction Design and Children*. IDC '13, ACM, 352-355.

[2] Anguera, M., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*. 20, 2 (Feb. 2012). IEEE, 356-370.

[3] Bakx, I, van Turnhout, K., and Terken, J. 2003. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of IFIP TC13 International Conference on Human-Computer Interaction*. Interact '03. IOS Press, 701-704.

[4] Chang, C. and Lin, C. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technologies*. 2, 3. ACM 1-27.

[5] Fellbaum, C. 1998 (editor). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

[6] Gerosa, M., Giuliani, D., Narayanan, S. and Potamianos, A. 2009. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. WOCCI '09. Article 7, 8 pages.

[7] Gustafson, J. and Sjolander, K. 2002. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In *Proceedings of* ICSLP 2002. ICSLP, 297-300.

[8] Hajishirzi, H., Lehman, J. F., and Hodgins, J. 2012. Using group history to identify character-directed utterances in multi-child interactions. In *Proceedings of the 13th Annual Meeting on Discourse and Dialogue*. SIGDIAL '12. ACL, 207-216.

[9] Leite, I., Hajishirzi, H., Andrist, S. and Lehman, J. 2013. Managing chaos: Models of turn-taking in character-multichild interactions. In *Proceedings of the 15th International Conference on Multimodal Interaction*. ICMI '13. ACM, 43-50.

[10] Russell, M. and D'Arcy, S. 2007. Challenges for computer recognition of children's speech. In *Proceedings of Speech and Language Technology in Education*. SLaTE-2007. ISCA, 108-111.