

Using Group History to Identify Character-directed Utterances in Multi-child Interactions

Hannaneh Hajishirzi, Jill F. Lehman, and Jessica K. Hodgins
hannaneh.hajishirzi, jill.lehman, jkh@disneyresearch.com

Abstract

Addressee identification is an element of all language-based interactions, and is critical for turn-taking. We examine the particular problem of identifying when each child playing an interactive game in a small group is speaking to an animated character. After analyzing child and adult behavior, we explore a family of machine learning models to integrate audio and visual features with temporal group interactions and limited, task-independent language. The best model performs identification about 20% better than the model that uses the audio-visual features of the child alone.

1 Introduction

Multi-party interaction between a group of participants and an autonomous agent is an important but difficult task. Key problems include identifying when speech is present, who is producing it, and to whom it is directed, as well as producing an appropriate response to its intended meaning. Solving these problems is made more difficult when some or all of the participants are young children, who have high variability in language, knowledge, and behavior. Prior research has tended to look at single children (Oviatt, 2000; Black et al., 2009) or multi-person groups of adults (Bohus and Horvitz, 2009a). We are interested in interactions between animated or robotic characters and small groups of four to ten year old children. The interaction can be brief but should be fun.

Here we focus specifically on the question of deciding whether or not a child’s utterance is directed to the character, a binary form of the addressee identification (AID) problem. Our broad goals in

this research are to understand how children’s behavior in group interaction with a character differs from adults’, how controllable aspects of the character and physical environment determine participants’ behavior, and how an autonomous character can take advantage of these regularities.

We collected audio and video data of groups of up to four children and adults playing language-based games with animated characters that were under limited human control. An autonomous character can make two kinds of AID mistakes: failing to detect when it is being spoken to, and acting as if it has been spoken to when it has not. The former can be largely prevented by having the character use examples of the language that it can recognize as part of the game. Such exemplification cannot prevent the second kind of mistake, however. It occurs, for example, when children confer to negotiate the next choice, respond emotionally to changes in the game state, or address each other without making eye contact. As a result, models that use typical audio-visual features to decide AID will not be adequate in multi-child environments. By including temporal conversational interactions between group members, however, we can both detect character-directed utterances and ignore the remainder about 20% better than simple audio-visual models alone, with less than 15% failure when being spoken to, and about 20% failure when not addressed.

2 Related Work

Our models explore the use of multimodal features that represent activities among children and adults interacting with a character over time. Prior research has tended to look at single children or multi-person

groups of adults and has typically used a less inclusive set of features (albeit in decisions that go beyond simple AID).

Use of multimodal features rests on early work by Duncan and Fiske who explored how gaze and head and body orientation act as important predictors of AID in human-human interactions (Duncan and Fiske, 1977). Bakx and colleagues showed that accuracy can be improved by augmenting facial orientation with acoustic features in an agent’s interactions with an adult dyad (Bakx et al., 2003). Others have studied the cues that people use to show their interest in engaging in a conversation (Gravano and Hirschberg, 2009) and how gesture supports selection of the next speaker in turn-taking (Bergmann et al., 2011). Researchers have also looked at combining visual features with lexical features like the parseability of the utterance (Katzenmaier et al., 2004), the meaning of the utterance, fluency of speech, and use of politeness terms (Terken et al., 2007), and the dialog act (Matsusaka et al., 2007). However, all use hand-annotated data in their analysis without considering the difficulty of automatically deriving the features. Finally, prosodic features have been combined with visual and lexical features in managing the order of speaking and predicting the end-of-turn in multi-party interactions (Lunsford and Oviatt, 2006; Chen and Harper, 2009; Clemens and Diekhaus, 2009).

Work modeling the temporal behavior of the speaker includes the use of adjacent utterances (e.g., question-answer) to study the dynamics of the dialog (Jovanovic et al., 2006), the prediction of addressee based on the addressee and dialog acts in previous time steps (Matsusaka et al., 2007), and the use of the speaker’s features over time to predict the quality of an interaction between a robot and single adult (Fasel et al., 2009).

Horvitz and Bohus have the most complete (and deployed) model, combining multimodal features with temporal information using a system for multi-party dynamic interaction between adults and an agent (Bohus and Horvitz, 2009a; Bohus and Horvitz, 2009b). In (Bohus and Horvitz, 2009a) the authors describe the use of automatic sensors for voice detection, face detection, head position tracking, and utterance length. They do not model temporal or group interactions in determining AID, al-

though they do use a temporal model for the interaction as a whole. In (Bohus and Horvitz, 2009b) the authors use the speaker’s features for the current and previous time steps, but do not jointly track the attention or behavior of all the participants in the group. Moreover, their model assumes that the system is engaged with at most one participant at a time, which may be a valid conversational expectation for adults but is unlikely to hold for children. In (Bohus and Horvitz, 2011), the authors make a similar assumption regarding turn-taking, which is built on top of the AID module.

3 User Study

We use a Wizard of Oz testbed and a scripted mix of social dialog and interactive game play to explore the relationship between controllable features of the character and the complexity of interacting via language with young children. The games are hosted by two animated characters (Figure 1, left). Oliver, the turtle, is the main focus of the social interactions and also handles repair subdialogs when a game does not run smoothly. Manny, the bear, provides comic relief and controls the game board, making him the focus of participants’ verbal choices during game play. The game appears on a large flat-screen display about six feet away from participants who stand side-by-side behind a marked line. Audio and video are captured, the former with both close-talk microphones and a linear microphone array.

Oliver and Manny host two games designed to be fun and easy to understand with little explicit instruction. In Madlibs, participants help create a short movie by repeatedly choosing one everyday object from a set of three. The objects can be seen on the board and Oliver gives examples of appropriate referring phrases when prompting for a choice. In Figure 1, for example, he asks, “Should our movie have a robot, a monster, or a girl in it?” After five sets of objects are seen, the choices appear in silly contexts in a short animation; for instance, a robot babysitter may serve a chocolate pickle cake for lunch. In Mix-and-Match (MnM), participants choose apparel and accessories to change a girl’s image in unusual ways (Figure 1, right). MnM has six visually available objects and no verbal examples from Oliver, except in repair subdialogs. It is a faster-paced game

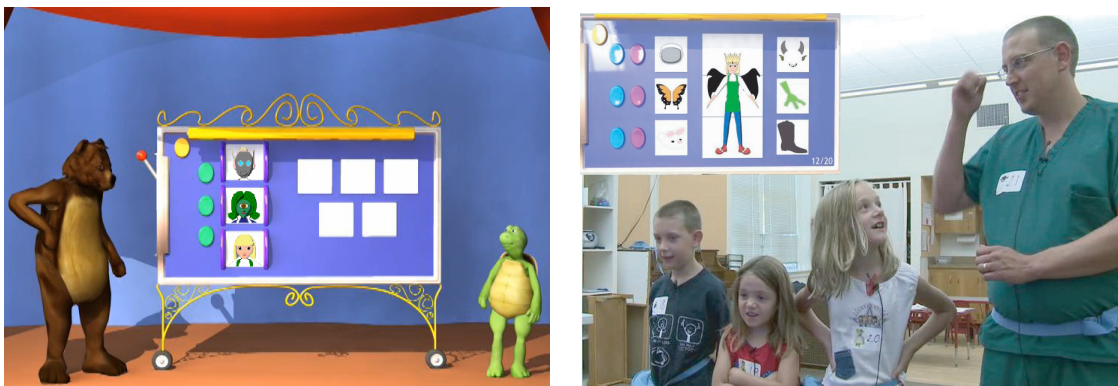


Figure 1: Manny and Oliver host Madlibs and a family play Mix-and-Match

with the immediate reward of a silly change to the babysitter’s appearance whenever a referring phrase is accepted by the wizard.

The use of verbal examples in Madlibs is expected to influence the children’s language, potentially increasing the accuracy of speech recognition and referent resolution in an autonomous system. The cost of exemplification is slower pacing because children must wait while the choices are named. To compensate, we offer only a small number of choices per turn. Removing exemplification, as in MnM, creates faster pacing and more variety of choice each turn, which is more fun but also likely to increase three types of problematic phenomena: out-of-vocabulary choices (“the king hat” rather than “the crown”), side dialogs to establish a referring lexical item or phrase (“Mommy, what is that thing?”), and the use of weak naming strategies based on physical features (“that green hand”).

The two games are part of a longer scripted sequence of interactions that includes greetings, good-byes, and appropriate segues. Overall, the language that can be meaningfully directed to the characters is constrained to a small social vocabulary, yes/no responses, and choices that refer to the objects on the board. The wizard’s interface reflects these expectations with buttons that come and go as a function of the game state. For example, *yes* and *no* buttons are available to the wizard after Oliver asks, “Will you help me?” while *robot*, *monster*, and *girl* buttons are available after he asks, “Should our movie have a robot, a monster, or a girl in it?” The wizard also has access to persistent buttons to indicate a long silence, unclear speech, multiple people speaking, or a clear reference to an object not on the board.

These buttons launch Oliver’s problem-specific repair behaviors. The decomposition of functionality in the interface anticipates replacing the wizard’s various roles as voice activity detector, addressee identifier, speech recognizer, referent resolver, and dialog manager in an autonomous implementation.

Although meaningful language to the characters is highly constrained, language to other participants can be about anything. In particular, both games establish an environment in which language among participants is likely to be about negotiating the turn (“Brad, do you want to change anything?”), negotiating the choice (“Billy, don’t do the boot”) or commenting on the result (“her feet look strange”). Lacking examples of referring phrases by Oliver, MnM also causes side dialogs to discuss how objects should be named. Naming discussions, choice negotiation, and comments define the essential difficulty in AID for our testbed; they are all likely to include references to objects on the board without the intention of changing the game state.

3.1 Data collection and annotation

Twenty-seven compensated children (14 male, 13 female) and six adult volunteers participated. Children ranged in age from four to ten with a mean of 6.4 years. All children spoke English as a first language. Groups consisted of up to four people and always contained either a volunteer adult or the experimenter the first time through the activities. If the experimenter participated, she did not make game choices. Volunteer adults were instructed to support their children’s participation in whatever way felt natural for their family. When time permitted, children were given the option of playing one or both

games again. Those who played a second time were allowed to play alone or in combination with others, with or without an adult. Data was collected for 25 distinct groups, the details of which are provided in Table 5 in the Appendix.

Data from all sessions was hand-annotated with respect to language, gesture, and head orientation. Labels were based on an initial review of the videos, prior research on AID and turn-taking in adults, and the ability to detect candidate features in our physical environment. A second person segmented and labeled approximately one third of each session for inter-annotator comparison. The redundant third was assigned randomly from the beginning, middle, or end of the session in order to balance across social interactions, Madlibs choices, and MnM choices. Labels were considered to correspond to the same audio or video sequence if the segments overlapped by at least 50%.

For language annotations, audio from the close-talk microphones was used with the video and segmented into utterances based on pauses of at least 50 msec. Typical mispronunciations for young children (e.g., word initial /W/ for /R/) were transcribed as normal words in plain text; non-standard errors were transcribed phonologically. Every utterance was also labeled as being directed to the character (CHAR) or not to the character (NCHAR). Second annotators segmented the audio and assigned addressee but did not re-transcribe the speech. Inter-annotator agreement for segmentation was 95% ($\kappa = .91$), with differences resulting from only one annotator segmenting properly around pauses or only one being able to distinguish a given child's voice among the many who were talking. For segments coded by both annotators, CHAR/NCHAR agreement was 94% ($\kappa = .89$).

For gesture annotations, video segments were marked for instances of pointing, emphasis, and head shaking yes and no. Emphatic gestures were defined as hand or arm movements toward the screen that were not pointing or part of grooming motions. Annotators agreed on the existence of gestures 74% of the time ($\kappa = .49$), but when both annotators interpreted movement as a gesture, they used the same label 98% of the time ($\kappa = .96$).

For orientation, video was segmented when the head turned away from the screen and when it turned

back. Rather than impose an *a priori* duration or angle, annotators were told to use the turn-away label when the turn was associated with meaningful interaction with a person or object, but not for brief, incidental head movements. Adults could also have segments that were labeled as head-incline if they bent to speak to children. Annotators agreed on the existence of these orientation changes 83% of the time ($\kappa = .62$); disagreements may represent simple differences in accuracy or differences in judgments about whether a movement denoted a shift in attention. Orientation changes coded by both annotators had the same label 92% of the time ($\kappa = .85$).

The annotated sessions are a significant portion of the training and test data used for our models. Although these data reflect some idiosyncrasy due to human variability in speech perception, gesture recognition, and, possibly, the attribution of intention to head movements, they show extremely good agreement with regard to whether participants were talking to the character. Even very young children in group situations give signals in their speech and movements that allow other people to determine consistently to whom they are speaking.

3.2 Analysis of behavior

As intended, children did most of the talking (1371/1895 utterances, 72%), spoke to the characters the majority of the time (967/1371, 71%), and made most of the object choices (666/683, 98%). Adults generally acted in support roles, with 88% of all adult utterances (volunteers and experimenter) directed to the children.

The majority of children's CHAR utterances (71%) were object choices. Although the wizard in our study was free to accept any unambiguous phrase as a valid choice, an automated system must commit to a fixed lexicon. In general, the larger the lexicon, the smaller the probability that a reference will be out-of-vocabulary, but the greater the probability that a reference could be considered ambiguous and require clarification. The lexical entry for each game object contains the simple description given to the illustrator ("alien hands," "pickle") and related terms from WordNet (Fellbaum, 1998) likely to be known by young children (see Table 3 in the Appendix for examples). In anticipation of weak naming strategies, MnM entries also contain salient

visual features based on the artwork (like color), as well as the body part the object would replace, where applicable. Entries for Madlibs objects average 2.75 words; entries for MnM average 5.8. With these definitions, only 37/666 (6%) of character-directed choices would have been out-of-vocabulary for a word-spotting speech recognizer with human accuracy. However, Oliver’s use of exemplification has a strong effect. In Madlibs, 98% of children’s choices were unambiguous repetitions of example phrases. In MnM, 92% of choices contained words in the lexicon, but only 28% indexed a unique object.

Recognition of referring phrases should be a factor in making AID decisions only if it helps to discriminate CHAR from NCHAR utterances. Object references occurred in 62% of utterances to the characters and only 25% of utterances addressed to other participants, but again, Oliver’s exemplification mattered. About 20% of NCHAR utterances from children in both games and from adults in Madlibs contained object references. In MnM, however, a third of adults’ NCHAR utterances contained object references as they responded to children’s requests for naming advice.

Language is not the only source of information available from our testbed. We know adults use both eye gaze and gesture to modulate turn-taking and signal addressee in advance of speech. Because non-verbal mechanisms for establishing joint attention occur early in language development, even children as young as four might use such signals consistently. Although we use head movement as an approximation of eye gaze, we positioned participants side-by-side to make such movements necessary for eye contact. Unfortunately, the game board constituted too strong a “situational attractor” (Bakx et al., 2003). As in their kiosk environment, our adults oriented toward the screen much of the time (68%) they were talking to other participants. Children violated conversational convention more often, orienting toward the screen for 82% of NCHAR utterances.

Gesture information is also available from the video data and reveals distinct patterns of usage for children and adults. The average number of gestures/utterance was more than twice as high in adults. Children were more likely to use emphasis gestures when they were talking to the characters; adults hardly used them at all. Children’s ges-

tures overlapped with their speech almost 80% of the time, but adult’s gestures overlapped with their speech only half the time. Moreover, when children pointed while talking they were talking to the characters, but when adults pointed while talking they were talking to the children. Finally, adults shook their heads when they were talking to children but not when they were talking to the characters, while children shook their heads when talking to both.

To maintain an engaging experience, object references addressed to the character should be treated as possible choices, while object references addressed to other participants should not produce action. Interactions that violate this rule too often will be frustrating rather than fun. While exemplification in Madlibs virtually eliminated out-of-vocabulary choices, it could not eliminate detectable object references that were not directed to the characters. In both games, such references were often accompanied by other signs that the character was being addressed, like orientation toward the board and pointing. Using all the cues available, human annotators were almost always able to agree on who was being addressed. The next section looks at how well an autonomous agent can perform AID using only the cues it can sense, if it could sense them with human levels of accuracy.

4 Models for Addressee Classification

We cast the problem of automatically identifying whether an utterance is addressed to the character (and so may result in a character action) as a binary classification problem. We build and test a family of models based on distinct sources of information in order to understand where the power is coming from and make it easier for other researchers to compare to our approach. All models in the family are constructed from Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and use the multimodal features in Table 1 to map each 500 msec time slice of a child’s speech to CHAR or NCHAR. This basic feature vector combines a subset of the hand-annotated data (Audio and Visual) with automatically generated data (Prosodic and System events). We use a time slice rather than a lexical or semantic boundary for forcing a judgment because in a real-time interaction decisions must be made even when

Audio	speech: presence/absence
Prosodic	pitch: low/medium/high speech power: low/medium/high
System event	character prompt: presence/absence
Visual	orientation: head turn away/back gesture: pointing/emphasis

Table 1: Basic features

lexical or semantic events do not occur.

We consider three additional sources of information: group behavior, history, and lexical usage. Group behavior – the speech, prosody, head orientation, and gestures of other participants – is important because most of the speech that is not directed to the characters is directed to a specific person in the group. History is important both because the side conversations unfold gradually and because it allows us to capture the changes to and continuity of the speaker’s features across time slices. Finally, we use lexical features to represent whether the participant’s speech contains words from a small, predefined vocabulary of question words, greetings, and discourse markers (see Appendix). Because the behavioral analysis showed significant use of words referring to choice objects during both CHAR and NCHAR utterances, we do not consider those words in determining AID. Indeed, we expect the AID decision to simplify the task of the speech recognizer by helping that component ignore NCHAR utterances entirely.

The full set of models is described by adding to the basic vector zero or more of group (g), word (w), or history (h) features. We use the notation $g[+/-]w[+/-]h[(\text{time parameters})/-]$ to indicate the presence or absence of a knowledge source. The time parameters vary and will be explained in the context of particular models, below. Although we have explored a larger portion of the total model space, we limit our discussion here to representative models (including the best model) that will demonstrate the effect of each kind of information on the two main goals of AID: responding to CHAR utterances and not responding to NCHAR utterances. There are eight models of interest, the first four of which isolate individual knowledge sources:

The *Basic* model ($g-w-h$) is an SVM classifier trained to generate binary CHAR/NCHAR values based solely on the features in Table 1. It represents the ability to predict whether a child is talking to the

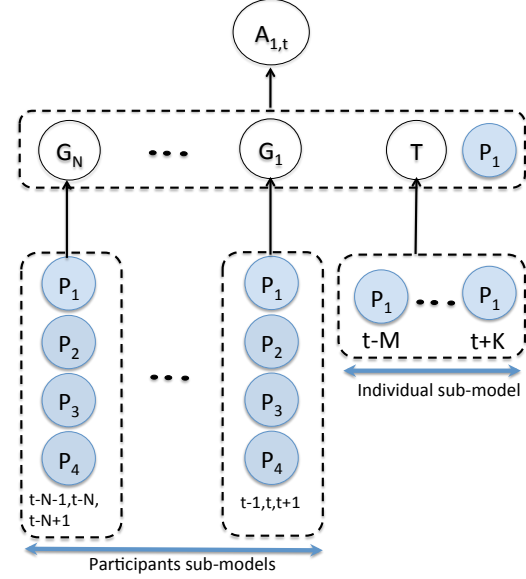


Figure 2: The two-layer *Group-History* model maps group and individual behavior over a fixed window of time slices to a CHAR/NCHAR decision at time t . The decision at time t ($A_{1,t}$) is based on the participant’s basic features (P_1), the output of the individual’s submodel (T) – which encapsulates the history of the individual for M previous and K subsequent time slices – and the output of N participant submodels, each of which contributes a value based on three time slices.

character independent of speech recognition and focused on only 500 msec of that child’s behavior.

The *Group* model ($g+w-h$) incorporates group information, but ignores temporal and lexical behavior. This SVM is trained on an extended feature vector that includes the basic features for the other participants in the group together with the speaker’s feature vector at each time slice.

The *History* model ($g-w-h(N,K)$) considers only the speaker’s basic features, but includes N previous and K subsequent time slices surrounding the slice for which we make the CHAR/NCHAR decision.¹

The *Word* model ($g-w+h$) extends the basic vector to include features for the presence or absence of question words, greetings, and discourse markers.

The next three models combine pairs of knowledge sources. The *Group-Word* ($g+w+h$) and *History-Word* ($g-w+h(N,K)$) models are straight-

¹A *History* model combining the speaker’s basic vector over the previous and current time slices ($N = 4$ and $K = 0$) outperformed a Conditional Random Fields (Lafferty et al., 2001) model with $N + 1$ nodes representing consecutive time slices where the last node is conditioned on the previous N nodes.

forward extensions of their respective base models, created by adding lexical features to the basic vectors. The *Group-History* model ($g+w-h(N,K,M)$) is more complex. It is possible to model group interactions over time by defining a new feature vector that includes all the participants’ basic features over multiple time slices. As we increase the number of people in a group and/or the number of time slices to explore the model space, however, the sheer size of this simple combination of feature vectors becomes unwieldy. Instead we make the process hierarchical by defining the *Group-History* as a two-layer SVM.

Figure 2 instantiates the *Group-History* model for participant P_1 playing in a group of four. In the configuration shown, the decision for P_1 ’s utterance at time t is based on behavior during N previous and K subsequent time slices, meaning each decision is delayed by K time slices with respect to real time. The CHAR/NCHAR decision for time slice t depends on P_1 ’s basic feature vector at time t , the output from the Individual submodel for time t , and the outputs from the Participants submodel for each of the time slices through t . A concrete instantiation of the model can be seen in Figure 4 in the Appendix.

The Individual submodel is an SVM that assigns a score to the composite of P_1 ’s basic feature vectors across a window of time (here, $M+K+1$). The Participants submodel is an SVM that assigns a score to the basic features of all members during each three slice sliding subwindow in the full interval. More intuitively: the Individual submodel finds correlations among the child’s observable behaviors over a window of time; the Participants submodel captures relationships between members’ behaviors that co-occur over small subwindows; and the *Group-History* model combines the two to find regularities that unfold among participants over time, weighted toward P_1 ’s own behavior.

The final model of interest, *Group-History-Word* ($g+w+h(N,K,M,Q)$), incorporates the knowledge from all sources of information. A Lexical submodel is added to the Individual and Participants submodels described above. The Lexical submodel is an SVM classifier trained on the combination of basic and word features for the current and Q previous time slices. The second layer SVM is trained on the scores of the Individual, Participants, and Lexical submodels as well as the combined basic and

Model	Max f1	AUC	TPR	TNR
Basic features				
g-w-h-	0.879	0.504	0.823	0.604
g+w-h-	0.903	0.588	0.872	0.650
g-w-h(8,1)	0.897	0.626	0.867	0.697
g+w-h(4,1,8)	0.903	0.645	0.849	0.730
Basic + Word features				
g-w+h-	0.904	0.636	0.901	0.675
g+w+h-	0.906	0.655	0.863	0.728
g-w+h(8,1)	0.901	0.661	0.886	0.716
g+w+h(4,1,8,4)	0.913	0.701	0.859	0.786

Table 2: Comparison of models

word feature vector for the child.

5 Results and Discussions

We used the LibSVM implementation (Chang and Lin, 2011) for evaluation, holding out one child’s data at a time during training, and balancing the data set to compensate for the uneven distribution of CHAR and NCHAR utterances in the corpus. As previously noted, we used a time slice of 500 msec in all results reported here. Where history is used, we consider only models with a single time slice of look-ahead ($K = 1$) to create minimal additional delay in the character’s response.

Table 2 reports average values, for each model and over all sets of remaining children, in terms of Max F_1 , true positive rate (TPR), true negative rate (TNR), and area under the TPR-TNR curve (AUC). TPR represents a model’s ability to recognize utterances directed to the character; low TPR means children will not be able to play the game effectively. TNR indicates a model’s ability to ignore utterances directed to other participants; low TNR means that the character will consider changing the game state when it hasn’t been addressed.

Table 2 (top) shows comparative performance without the need for any speech recognition. F_1 and TPR are generally high for all models. Using only the basic features, however, gives a relatively low TNR and an AUC that is almost random. The *History* model, ($g-w-h(8,1)$), increased performance across all measures compared to the basic features ($g-w-h$). We found that the *History* model’s performance was best when four seconds of the past were considered. Group information within a single time slice also improves performance over the basic features, but the *Group-History* model has the

best overall tradeoff in missed CHAR versus ignored NCHAR utterances (AUC). *Group-History*'s best performance is achieved using two seconds of group information from the past via the Participants submodel and four seconds of the speaker's past from the Individual submodel.

Comparing the top and bottom halves of Table 2 shows that all models benefit from accurate recognition of a small set of task-independent words. The table shows that word spotting improves both TPR and TNR when added to the *Basic* model, but tends to improve only TNR when added to models with group and history features. Improved TNR probably results from the ability to detect NCHAR utterances when participants are facing the characters and/or pointing during naming discussions and comments.²

Table 2 shows results averaged over each held out child. We then recast this information to show, by model, the percentage of children that would experience TPR and TNR higher than given thresholds. Figure 3 shows a small portion of a complete graph of this type; in this case the percentage of children who would experience greater than 0.6 for TPR and greater than 0.5 for TNR under each model. TPR and TNR lines for a model have the same color and share a common pattern.

Better models have higher TPR and TNR for more children. The child who has to keep restating his or her choice (poor TPR) will be frustrated, as will the child who has the character pre-emptively take his or her choice away by "overhearing" side discussions (poor TNR). While we do not know for any child (or any age group) how high a TPR or TNR is required to prevent frustration, Figure 3 shows that without lexical information the *Group-History* and *Group* models have the best balance for the thresholds. *Group-History* gives about 85% of the children a $TPR \geq 0.7$ for a $TNR \geq 0.5$. The simpler *Group* model, which has no 500 msec delay for lookahead, can give a better TPR for the same TNR but for only 75% of the children. When we add lexical knowledge the case for *Group-History* becomes stronger, as it gives more than 85% of children a $TPR \geq 0.7$ for a $TNR \geq 0.6$, while *Group* gives 85% of children about the same TPR with a $TNR \geq 0.5$.

²Results showing the affect of including object choice words in the *w+* models are given in Figure 4 in the Appendix.

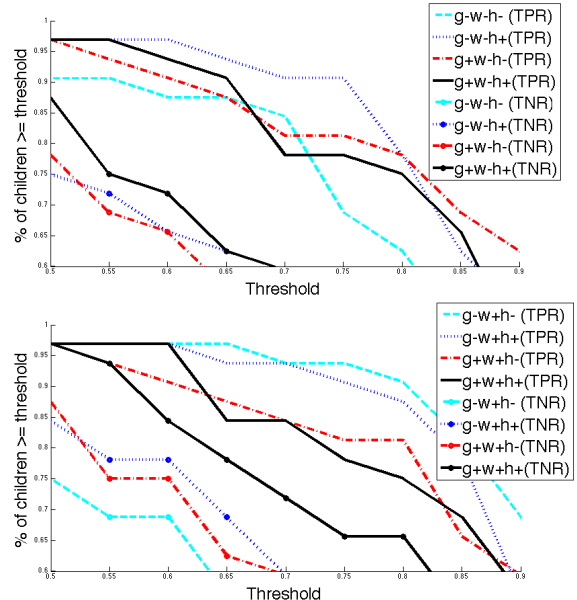


Figure 3: The percentage of children experiencing different TPR/TNR tradeoffs in models with (*bottom*) and without (*top*) lexical knowledge. The g-w-h- model does not fall in the region of interest unless lexical features are used.

6 Conclusions and Future Work

The behavior of the characters, types of games, group make up, and physical environment all contribute to how participants communicate over time and signal addressee. We can manipulate some relationships (e.g., by organizing the spatial layout to promote head movement or having the character use examples of recognizable language) and take advantage of others by detecting relevant features and learning how they combine as behavior unfolds. Our best current model uses group and history information as well as basic audio-visual features to achieve a max F_1 of 0.91 and an AUC of 0.70. Although this model does not yet perform as well as human annotators, it may be possible to improve it by taking advantage of additional features that the behavioral data tells us are predictive (e.g., whether the speaker is an adult or child). Such additional sources of information are likely to be important as we replace the annotated data with automatic sensors for speech activity, orientation, and gesture recognition, and embed addressee identification in the larger context of turn-taking and full autonomous interaction.

References

- I. Bakx, K. van Turnhout, and J. Terken. 2003. Facial orientation during multi-party interaction with information kiosks. pages 701–704.
- K. Bergmann, H. Rieser, and S. Kopp. 2011. Regulating dialogue with gestures towards an empirically grounded simulation with conversational agents. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 88–97.
- Matthew Black, Jeannette Chang, Jonathan Chang, and Shrikanth S. Narayanan. 2009. Comparison of child-human and child-computer interactions based on manual annotations. In *Proceedings of the Workshop on Child, Computer and Interaction*, Cambridge, MA.
- D. Bohus and E. Horvitz. 2009a. Dialog in the open world: Platform and applications. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 31–38.
- D. Bohus and E. Horvitz. 2009b. Learning to predict engagement with a spoken dialog system in open-world settings. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 244–252.
- D. Bohus and E. Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 98–109.
- C. Chang and C. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transaction on Intelligent Systems and Technologies*, 2:27:1–27:27.
- L. Chen and M. Harper. 2009. Multimodal floor control shift detection. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*.
- C. Clemens and C. Diekhaus. 2009. Prosodic turn-yielding cues with and without optical feedback. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 107–110.
- C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning Journal*, 20.
- S. Duncan and D. W. Fiske. 1977. *Face-to-Face Interaction: Research, Methods and Theory*. Lawrence Erlbaum.
- I. Fasel, M. Shiomi, T. Kanda, N. Hagita, P. Chadutaud, and H. Ishiguro. 2009. Multi-modal features for real-time detection of human-robot interaction categories. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 15–22.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- A. Gravano and J. Hirschberg. 2009. Turn-yielding cues in task-oriented dialogue. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 253–261.
- N. Jovanovic, H.J.A. op den Akker, and A. Nijholt. 2006. Addressee identification in face-to-face meetings. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 169–176.
- M. Katzenmaier, R. Steifelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 144–151.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 282–289.
- R. Lunsford and S. Oviatt. 2006. Human perception of intended addressee during computer assisted meetings. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 20–27.
- Y. Matsusaka, M. Enomoto, and Y. Den. 2007. Simultaneous prediction of dialog acts and address types in three party conversations. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 66–73.
- Sharon Oviatt. 2000. Talking to thimble jellies: children’s conversational speech with animated characters. pages 877–880.
- J. Terken, I. Joris, and L. de Valk. 2007. Multimodal cues for addressee hood in triadic communication with a human information retrieval agent. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*.

7 Appendix

Object Choice Words	
	antler, antlers, horn, horns, ear, ears, head, brown
	astronaut, astronauts, space, spaceman, spacemen, space-men, helmet, head
	bear, bears claw, claws, paw, paws, hand, hands, brown
	bunny, rabbit, bunnies, rabbits, slipper, slippers, foot, feet, white
Task-independent Words	
Discourse marker	hmm, mm, mmm, ok, eww, shh, oopsy
Question words	what, let, where, who, which, when
Greetings	hi, hello, bye, goodbye

Table 3: Excerpts from the dictionary for task-specific and task-independent words

Model	Max f1	AUC	TPR	TNR
Greeting, question & discourse words				
g-w+h-	0.904	0.636	0.901	0.675
g+w+h-	0.906	0.655	0.863	0.728
g-w+h(8,1)	0.901	0.661	0.886	0.716
g+w+h(4,1,8,4)	0.913	0.701	0.859	0.786
With object reference words added				
g-w+h-	0.894	0.576	0.777	0.768
g+w+h-	0.898	0.623	0.782	0.773
g-w+h(7,1)	0.910	0.642	0.838	0.783
g+w+h(4,1,8,4)	0.912	0.685	0.834	0.799

Table 4: The effect of adding object reference words

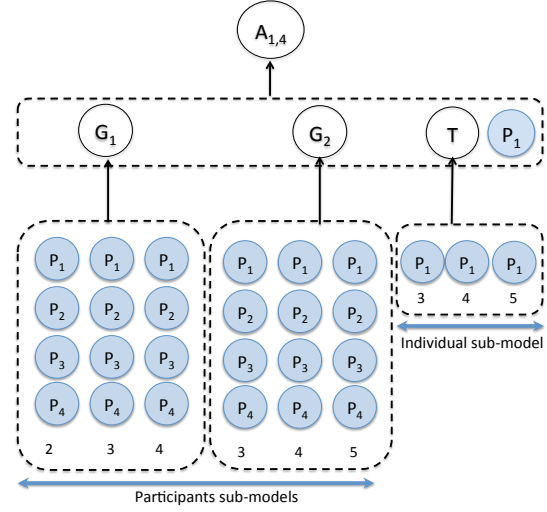


Figure 4: A concrete representation for the *Group-History* model with $N = 2$, $M = 1$, and $K = 1$ at time step $t = 4$. The value at $t = 4$ is delayed one time slice of real time.

Session Type	Group: participant(age)	Duration
full	p1(5), experimenter	9 min
full	p2(7), p3(6), p6(adult)	9 min
full	p4(7), p5 (4), p6(adult)	9 min
replay	p2(7), p3(6), p4(7), p5(4)	8 min
full	p7(10), experimenter	8 min
replay	p7(10)	6 min
full	p8(9), p9(8), experimenter	9 min
full	p10(10), p11(5), experimenter	11 min
full	p12(6), p14(adult)	11 min
full	p13(4), p14(adult)	11 min
full	p15(4), experimenter	8 min
full	p16(9), p17(7), experimenter	12 min
replay	p16(9), experimenter	3 min
full	p18(8), p19(6), p20(8), p21(adult)	12 min
full	p22(5), experimenter	9 min
replay	p22(5), experimenter	3 min
full	p25(6), experimenter	9 min
full	p26(8), p27(4), experimenter	11 min
replay	p26(8), experimenter	6 min
full	p28(7), p29(adult)	12 min
full	p30(5), experimenter	11 min
replay	p30(5), experimenter	4 min
full	p31(6), p32(5), p33(adult)	10 min
full	p34(4), p35(adult)	9 min
replay	p34(4), p35(adult)	4 min

Table 5: Details for sessions used in the analysis (does not include five sessions with corrupted data)