# Recognizing Character-directed Utterances in Multi-child Interactions

Hannaneh Hajishirzi, Jill F. Lehman, Kenichi Kumatani, Leonid Sigal, Jessica Hodgins
hannaneh.hajishirzi,jill.lehman,kenichi.kumatani,lsigal,jkh@disneyresearch.com
Disney Research Pittsburgh
4720 Forbes Avenue, Pittsburgh, PA
Pittsburgh, PA

## ABSTRACT

We address the problem of identifying when a child playing an interactive game in a small group is speaking to an animated or robotic character versus conferring with his friend. This judgment about addressee is critical for turn-taking. We explore a machine learning approach using a Support Vector Machine (SVM) to integrate audio and visual features that we believe can be sensed accurately. We extend the basic model by including a simple form of group information, limited speech recognition, and limited game state to improve classification accuracy. Our results demonstrate high accuracy in detecting when the character is being addressed. This model improves our understanding of children's group behavior in interacting with an agent.

## Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]; H.5.3 [**Group and Organization Interfaces**]

## General Terms

Human factors, Experimentation

## Keywords

Human character interactions, Machine learning experiments

## 1. INTRODUCTION

We are interested in autonomous, language-based interaction between animated or robotic characters and small groups of children. The interaction can be brief but should be fun. The age group studied is four to ten year olds, entailing high variability in articulation, vocabulary and behavior. 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. The focus here is the use of machine learning to perform addressee identification and understand which features play important roles. In particular, we want to determine whether a child is speaking to the character or conferring with other participants.

*HRI'12,* March 5–8, 2012, Boston, Massachusetts, USA.
ACM 978-1-4503-1063-5/12/03.

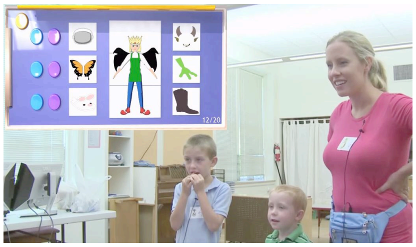

**Figure 1: Family playing Mix-and-Match**

## 2. USER STUDY

Twenty-seven compensated children and seven adult volunteers participated in groups of two to four players. Participants stood side-by-side, facing a large flat-screen display, about six feet away. Audio and video were captured, the former with both close-talk microphones and a linear microphone array collocated with the display.

Two games were available and could be played multiple times. The games required verbal interactions that included greetings, responses to yes/no questions, and referring phrases to choose from among three or six objects (Figure 1). Because participants were free to confer with each other prior to choosing, non-task utterances were also likely to include the same referring vocabulary.

The character's behavior was controlled via a Wizard of Oz set-up. The wizard's interface allowed only a small number of classifications: long silence, unclear speech, multiple people speaking, clear reference to an object not on the board, or a choice of one of the objects shown. The first four options caused the character to use specific reprompts, while a clear reference to a pictured item resulted in character action that changed the game state. Thus, in addition to acting as a voice activity detector, speaker identifier, addressee identifier, and speech recognizer, the wizard's behavior implicitly defined the character's turn-taking strategy as part of the overall natural language processing task.

All speech during game play was transcribed and annotated with the judgment CHAR or NCHAR to indicate if it was directed to the character. Annotators also labeled a small vocabulary of gestures (head shake yes, head shake no, pointing, emphasis) and movements (head turn away, head turn back, head incline). Gesture and orientation labels were based both on available technologies for sensing the relevant features and prior work in addressee identification and turn-taking in adults, e.g. [5, 4]. In particular, we follow [1] by substituting head orientation for eye gaze in designating addressee.

Children did most of the talking (1371/1895 utterances) during game play, addressing the character 71% of the time and gesturing during about 12% of their utterances. Children's NCHAR utterances were primarily requests for help in naming an object in the game or negotiations over the next choice. Adults acted largely in a support role, addressing the character only 12% of the time, and gesturing about the same amount (14% of utterances). In a physical set-up like ours, Bakx and colleagues [1] described the effect of "situational attractors" on facial orientation, noting that their adult dyads faced the system about 60% of the time they were actually speaking to each other. We observed similar behavior in our adults, who were oriented toward the screen during about 68% of NCHAR interactions, but more violation of the conversational convention in the children, who oriented toward the screen 82% of the time they were speaking to another person.

# 3. ADDRESSEE CLASSIFICATION

We cast the problem of automatically identifying whether an utterance is addressed to the character (and so should result in character action) as a binary classification problem. We represent each time slice of a child's participation with a set of features and learn one or more Support Vector Machine (SVM) classifiers to map the time slice to CHAR or NCHAR. Performance depends on the size of the time slice, the feature set, and the topology of the model. We report on four models, all of which use a time slice of 500 msecs and the following basic features, either derived from the hand-annotated data or generated automatically:

- Child speech (hand): **present** or **absent** over most of the time slice, independent of the content of the utterance
- Animation (hand/computer): **prompt** or **not**, whether the animation is generating speech or sound effects that are prompting for a response
- Orientation (hand): **head turn away** and **head turn back**
- Sound (computer): **pitch** and **volume**, averaged over the time slice

**Basic SVM:** This model is an SVM classifier [3] trained to predict binary CHAR/NCHAR values based on the basic feature vector at each time slice. It represents the ability to predict the addressee independent of speech recognition and focused on only the current time slice (500 msecs) of the child's behavior.

**SVM-group:** Speech not directed to the character is usually directed toward another person in the group, typically an adult. To take advantage of this regularity, we add the feature vector for the group leader to the feature vector for each child at each time slice. This model would require the additional ability to identify the group leader.

**SVM-group-words:** This model considers the effect of accurate speech recognition over a small, task-independent vocabulary. We add a Content Marker feature to the vector, capturing whether the participant's speech contained a small set of discourse markers (e.g., *um, ok*) or WH question words (e.g., *what, where*) in the transcribed data.

**SVM-group-words-history:** Both the Animation feature and the leader's vector have the potential to be most useful when they are considered over time. For an initial test of temporal effects, we change to a two-layer topology. In the first layer, we use the SVM-group-words model to compute the CHAR/NCHAR score for the time slice. For the second layer, we train a new SVM whose features include values returned by the first-layer SVM for $k = 1$ previous and $l = 2$ next time steps. At test time, we use the learned SVM models in both layers to assign CHAR/NCHAR labels. This model would increase the delay in a real-time application of the classifier.

| Approach | Max f1 | AUC | TPR | TNR |
|---|---|---|---|---|
| SVM | 0.88 | 0.48 | 0.80 | 0.61 |
| SVM-group | 0.89 | 0.54 | **0.90** | 0.61 |
| SVM-group-words | **0.90** | 0.59 | 0.89 | 0.67 |
| SVM-group-words-history | 0.90 | **0.63** | 0.88 | **0.72** |

**Table 1: Results of applying our method under different conditions**

## 3.1 Results

We used the LibSVM implementation [2] under all conditions, holding out one child's data at a time during training, and balancing the data set (balancing factors 2 and 5 in LibSVM) to compensate for the uneven distribution of CHAR and NCHAR utterances in the corpus. Table 1 reports average values over all sets of remaining children, for each condition in terms of Max $F_1$, area under the precision-recall curve (AUC), true positive rate (TPR), and true negative rate (TNR).

Using only the basic features (SVM), $F_1$ and TPR are high for our interactions, but the TNR is relatively low and therefore the area under the precision/recall curve is low as well. Adding the basic features of a group leader to the child's data (SVM-group)gives increased performance across all measures. SVM-group-words shows how even a small amount of accurate speech recognition could further improve $F_1$, by helping to achieve a more balanced true positive/true negative rate. Finally, for classifications that could be done off-line or with an extra one second delay to buffer the look-ahead, the best AUC value is achieved.

Our current model improved our understanding of the effect of children's visual and audio features in interacting with an agent. We plan to advance the two-layer SVM model by adding the pairwise relations among all the people in the group (thus, eliminating the need to identify the leader). In addition, we intend to use automatic sensors for speech activity, orientation, and gesture recognition with our audio and video data, and compare the results to what has been achievable with human annotation.

# 4. REFERENCES

[1] I. Bakx, K. van Turnhout, and J. Terken. Facial orientation during multi-party interaction with information kiosks. In *INTERACT*, pages 701–704, 2003.

[2] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intellig. Sys. and Tech.*, 2:27:1–27:27, 2011.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[4] N. Jovanovic, H. op den Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *EACL*, pages 169–176, 2006.

[5] M. Katzenmaier, R. Steifelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *ICMI*, pages 144–151, 2004.