Analyzing the Nature of ECA Interactions in Children with Autism

Emily Mower[1], *Chi-Chun Lee*[1], *James Gibson*[1], *Theodora Chaspari*[1],
Marian Williams[2], *Shrikanth Narayanan*[1]

University of Southern California (USC)
[1] Signal Analysis and Interpretation Laboratory, USC, Los Angeles, California, USA
[2] Keck School of Medicine, USC, Los Angeles, California, USA
{mower, chiclee, jjgibson, chaspari}@usc.edu, mwilliams@chla.usc.edu, shri@sipi.usc.edu

10.21437/Interspeech.2011-748

Abstract

Embodied conversational agents (ECA) offer platforms for the collection of structured interaction and communication data. This paper discusses the data collected from the Rachel system, an ECA developed at the University of Southern California, for interactions with children with autism. Two dyads each composed of a child with autism and his parent participated in an experiment with two modes: interactions with and without the ECA present. The goal of this work is to assess the naturalness of the data recorded in the ECA interaction. This analysis was carried out using a classification framework with a prediction variable of the presence or absence of the ECA in the interaction. The results demonstrate that it is possible to estimate whether or not a parent is interacting with the ECA using their speech data. However, it is not generally possible to do so for the child suggesting that the Rachel system is eliciting communication data that is similar to that elicited through interactions between the child and his parent.

Index Terms: Embodied conversational agent, multimodal interface, audio-video recording, autism, children's speech

1. Introduction

Autism spectrum disorders (ASD) are pervasive developmental disorders in which the core symptoms include qualitative abnormalities in social communication [1]. Previous research has suggested that embodied conversational agents (ECA) can be used to elicit conversational data from children with autism [2, 3]. ECAs provide a structured method for eliciting communicative behavior, especially from children [4], using a controlled interaction scenarios that facilitate the comparison of communication behavior, both between and within subjects. This paper presents an analysis of the similarity in a child's speaking patterns when interacting with an ECA and with his parent.

Autism affects the social communicative behaviors of a child resulting in: spoken language delays, difficulty in initiating and sustaining communication, and the use of stereotyped and repetitive language [5]. ECAs provide a platform through which to analyze the social communicative abilities of children. However, a potential limitation of these ECA interfaces is that the collected interaction data may be biased by the manner in which it was collected; the child-ECA interactions may not be as natural as data collected in a more free-form setting. This paper will demonstrate that when properly designed, these ECAs can be used to collect data that is similar to that collected in interactions between a child and parent. This will be demonstrated by comparing the audio and lexical content of the child's and parent's speech both when the ECA is present and absent.

The interaction patterns resulting from ECA interactions have been studied in previous work. In [6, 7] the authors analyzed a real-estate ECA to study how to facilitate natural (adult) human-computer interaction. In [8] the authors analyzed the interplay between eye gaze and speech activity for adults interacting with a virtual agent and a human interlocutor. In [9] the authors performed a statistical analysis of speech pattern differences relating to whether the child was interacting with a human or virtual agent. This paper is an extension of the work presented in these studies seeking to use the differences in speech resulting from either a human or ECA interaction partner to predict the identity of the partner (human or ECA) approximating the perceptual difference in the observed interaction patterns.

The Rachel system is an ECA environment created at the University of Southern California to collect communication data from children with autism and their parents [10]. It consists of an ECA, Rachel, and a series of activities both Rachel- and parent-moderated. The child and parent interact with Rachel across four separate sessions. The use of an ECA to collect social communicative behavior assumes that the elicited behavior is representative of the child's communication abilities. This paper tests this assumption by automatically classifying if the child or parent is speaking in a Rachel- or parent-moderated interaction (an interaction in which Rachel is not present) using data extracted from the audio and the lexical content of the parent's and child's speech.

The results demonstrate that in the majority of cases the parent's audio data provides discriminatory information regarding the presence or absence of Rachel while the child's audio data does not. This indicates a similarity in the child's communicative behavior across the two interaction conditions, with and without the ECA present, along the analysis dimensions.

2. Design

The Rachel experiment is a four-session study (approximately twice a week) with activities designed to assess the children's emotional reasoning ability. Each session follows the same protocol: 1) the child and parent enter the experimental room and are (re-)introduced to Rachel; 2) Rachel leads the child through a briefing and warm-up game; 3) Rachel leads the child through a series of emotional problem-solving tasks; 4) Rachel leads the child through a story telling task and short debrief; 5) the parent leads the child through a story telling task; 6) the parent follows the child's lead through a game that complements the Rachel tasks. In the Rachel-moderated portion, the parent and child interact with each other and Rachel (activities 1–4). In the parent-moderated task, the parent and child interact without Rachel (activities 5 and 6).

Copyright © 2011 ISCA

The Rachel system has a different challenge level associated with each session. The first level asks children to play an emotion spotting game. The second level is an emotional story telling task in which children tell stories based on sets of images. The third level utilizes the same emotional imagery but removes some facial expressions and the children tell a story and identify the logical emotional faces given the story. The final level utilizes the same emotional imagery but includes incorrect facial expressions given the stories in the previous two sessions; the children explain why certain faces are incorrect. In all sessions Rachel leads a story telling task using a book of pictures (no words); Rachel starts the story and the child finishes.

The parent-child interaction included two components: a story telling task and an activity. The parent plays the same role as Rachel in the story telling task, initiating the story telling. The activities differ in each session. The activity for the first session is imaginative play using: Mr. Potato Head, blocks, and play doh. The activity for the second session is emotional Jenga in which the parent and child use blocks marked with emotional words to play Jenga and act out the emotions if the block is so marked. The activity for the third session is a drawing activity. The parent and child are given markers and are asked to draw a time when they were happy and explain their drawings to each other. The final session's activity is a follow-up interview in which the parent and child explain both what they liked and did not like about the Rachel system.

Each session was recorded using a suite of audio-visual sensors. The behavior of the ECA was logged for post-hoc analysis. The ECA was controlled using the Wizard of Oz (WoZ) paradigm in which a hidden experimenter controls the output of the ECA. Additional information can be found in [10].

3. Description of Subjects

The Rachel system has been evaluated on two children each over four sessions. The subjects met the inclusion criteria: of a diagnosis of autism using the Autism Diagnostic Observation Scale (ADOS), that the child and the parent both speak English, that the child is from 5-13 years of age, and that the child has received a score on the Expressive Communication subtest of the Vineland Adaptive Behavior scales of at least an age equivalent of 2 years, 6 months. The first child was a 12-year-old boy with an expressive language score of 6 years, 7 months (Vineland). He was accompanied by his mother during the first session and his father in the remaining sessions. His younger brother also attended the final session. The interactions between this child, his parent, and Rachel will be referred to as the "subject one experiments." The second child was a 6-year-old boy with an expressive language score of 2 years, 9 months (Vineland). He was accompanied by his mother in all four sessions. The interactions between this child, his parent, and Rachel will be referred to as the "subject two experiments." The inclusion criteria and subjects are described more fully in [10].

4. Methods

The purpose of the studies presented in this paper is to understand how the data collected from the Rachel-moderated interactions differ from that of the parent-moderated interactions (without Rachel). The hypothesis is that the Rachel interactions elicited communication data similar in form to that of the parent-moderated interactions. The parent-moderated interactions may also not be fully representative of the child's natural interaction patterns. However, observation is necessary to understand the child's communicative ability and as such the system must operate under this limitation.

This paper will test the hypothesis in two parts. The first study will analyze the similarity of the children's communication patterns collected during the Rachel-moderated and parent-moderated interactions. The expectation is that the communication patterns will be similar because previous research has demonstrated the efficacy of computer avatar-based interactions for children with autism. The second study will assess changes in the parent's communication patterns as a function of the interaction session. The expectation is that the parents will become increasingly interactive during the Rachel-moderated activities and that the parents will take on the role of interaction moderator when Rachel is not present. Thus, their speech patterns would be markedly different with respect to the nature of the interaction. The speech patterns are quantified using audio and lexical features (Section 4.1). Due to the interactive nature of this corpus, there was overlapped speech and very short utterances. We only used utterances that were longer than 0.5 seconds and contained no overlapped speech (67.22% of the data).

These hypotheses are tested using a feature selection and classification approach. Feature selection (Section 4.2) highlights features that differ between conditions. Classification (Section 4.3) provides the validation of both the importance of these features and their ability to discriminate between the target classes. The combination of these techniques demonstrates the feasibility of using the selected features and the feasibility of discriminating between the conditions.

The audio and lexical features are extracted from the speech and transcript data and feature selection is performed. The extracted features are used in two classification experiments using speaker-specific leave-one-utterance-out cross-validation (within a single session training and testing). This method was chosen because it permits the assessment of speech pattern similarity for an interaction on a given day. The goal of this paper is to demonstrate that an ECA can be used to elicit natural speech behavior similar in content to speech behavior elicited in a parent-moderated interaction motivating the selection of a cross-validation method that supports this type of comparison.

4.1. Feature Extraction

The lexical features were extracted from the transcripts at the speaker turn-level, defined as a period of time in which the speaker spoke without interruption. The lexical features include the number of words, length of utterance, number of laughs, richness of vocabulary, and the use of: backchannels (short feedback, e.g., 'mm-hmm', 'yeah'), proper nouns, pronouns, personal pronouns, prepositions, conjunctions, polite words, modal auxiliary verbs, words indicating family relations, question words, contractions, made up words, and informal words.

The number of words per utterance, the length of utterance, and the presence of laughter in the utterance were extracted directly from the transcripts. The richness of vocabulary was extracted using a list of words observed during the course of an interaction. For each turn, new words for a given speaker augmented the richness count. The use of backchannels, proper nouns, pronouns, personal pronouns, prepositions, conjunctions, words indicating politeness, modal auxiliary verbs, words indicating family relations, question words, contractions, made up words, and informal words were calculated by maintaining lists relevant to the vocabulary of the children.

The audio features were also extracted at the turn level and include: pitch, intensity, and the first 13 Mel Filterbank Coefficients (MFB) extracted using Praat [11]. These audio features have been demonstrated to be effective in behavioral studies [12, 13]. The features utilized in this study are statistical

Table 1: Correlation-based feature similarity for features extracted over the four sessions for the subject one and two experiments. Correlations in bold are significant at $\alpha \leq 0.05$

		Sub 1		Sub 2	
		Child	Parent	Child	Parent
Sub 1	Child	–	**0.29**	**0.25**	**0.30**
	Parent	**0.29**	–	0.11	-0.14
Sub 2	Child	**0.25**	0.11	–	**0.25**
	Parent	**0.30**	-0.14	**0.25**	–

functionals extracted over each speaker turn and include: mean, standard deviation, range, skewness, and kurtosis.

4.2. Feature Selection

The feature set included 75 audio and 20 lexical features. Feature selection was performed using the Wilcoxon Rank Sum Test, a nonparametric test for equivalence of the median. This test was chosen because the distributions of the features are not normal. The two groups were defined as speech extracted from a single speaker when a) Rachel was present (Rachel-moderated interaction) vs. b) when Rachel was not present (parent-moderated interaction). This feature selection technique identifies features with medians that are statistically significantly different in the two groups. All selected features were included in the final feature set.

The patterns of features selected for the parent and child are similar. This similarity can be quantified by aggregating the selected features over the four sessions creating a vector with a maximum count of four (the features were selected for each of the sessions). In the merged feature vector all audio feature classes (pitch, intensity, MFB) are represented for all four speakers. The similarity is calculated by comparing the aggregated feature vectors across each speaker (subject 1: parent and child, subject 2: parent and child) using a correlation metric calculated with Kendall's Tau (a nonparametric test for correlation). The results can be seen in Table 1.

In the subject one experiments, the child's feature vector is significantly correlated with his parent's and that of both the parent and child from the subject two experiments (Table 1). The parent's feature vector is significantly correlated with the child's but with neither the parent's nor the child's from the subject two experiments. In the subject two experiments the child's feature vector was correlated with his parent's and with the child's from the subject one experiments. The feature vector for the parent is significantly correlated with those of the children from the subject one and two experiments. The children's feature vectors are significantly correlated with each other and for subject one, those of the opposite parent while the parent's feature vectors are not correlated with each other. The Rachel- and parent-moderated tasks are geared towards the child which may explain the correlation between the children's feature vectors. The lack of correlation between the parent vectors may be indicative of the manner through which they chose to engage their child. However, the targeted nature of the engagement may explain the observed correlation between the feature vectors of the parents and the subject one child.

4.3. Classification

The data were classified using the binary classifier Support Vector Machines (SVM). A linear SVM separates two classes of data using a plane maximally far from the points closest to the separating plane. SVM can be augmented to include a nonlinear kernal function. This nonlinear component projects data into a higher dimensional space to separate data that are not linearly separable. In this paper Radial Basis Function (RBF) kernals are used with a sigma of 13, determined empirically. SVM has shown to be an effective classifier for behavioral data [13]. The results presented were also analyzed using the K-Nearest Neighbor (kNN) algorithm. In kNN, an unlabeled test point is classified by surveying its k nearest neighbors. The class most highly represented among those neighbors is assigned as the class label for the unknown test point.

Table 2: SVM results (chance in parentheses) for the subject experiments (audio features). Subjects are referred to by subject number and ch (child) or pa (parent). Bold entries are significantly different from chance ($\alpha \leq 0.05$, diff. of proportions).

	Ses1	Ses2	Ses3	Ses4
S1-Ch	0.79 (0.79)	**0.77** (0.65)	0.80 (0.80)	**0.71** (0.56)
S1-Pa	0.89 (0.89)	**0.78** (0.57)	**0.85** (0.67)	**0.84** (0.55)
S2-Ch	0.76 (0.68)	0.70 (0.71)	0.60 (0.61)	0.78 (0.78)
S2-Pa	**0.82** (0.66)	**0.74** (0.56)	0.65 (0.63)	0.81 (0.72)

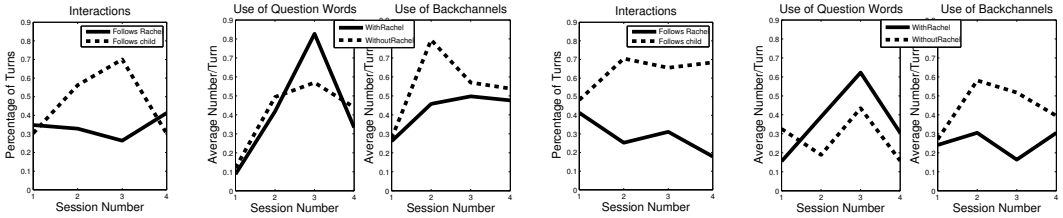
5. Results

The classification goal is to determine the interaction type (Rachel-moderated vs. parent-moderated) given data from a single speaker. The SVM classification (Table 2) of the parents' and children's data using audio features achieved an average accuracy for subject 1 of 76.88% (child) and 84.14% (parent) and for subject 2 of 71.17% (child) and 75.60% (parent). The chance classification accuracies are 70.13%, 67.01%, 69.41%, and 64.24%, respectively. The kNN results were lower overall, but followed the same trends as the SVM results. Thus, only the SVM results will be discussed. The lexical features achieved classification results that were not statistically significantly different from chance despite several lexical features with statistically significantly different medians according to the Wilcoxon test. There were also no benefits gained from augmenting the audio feature vector to include the lexical features.

The results suggest that for subject one experiments the child's speech could be used to identify the interaction type in two of the four sessions (sessions two and four) and that the parent's speech was differentiable in three of the four sessions (except session one). The subject two experiment results demonstrate that the child's speech was not differentiable and that the parent's speech was differentiable in two of the four interaction scenarios (sessions one and two, the sample size was low for session four). It should be noted that in session four of the subject one experiments the parent brought the child's brother to the experimental session. Therefore, the interaction dynamics in the Rachel-moderated section did not include as much parent feedback and encouragement as seen in sessions one through three. This may have accounted for the child's distinctive speech patterns in the Rachel-moderated vs. parent-moderated interaction scenarios in that session.

6. Discussion

The classification results suggest that the speech patterns of the parents generally differed between the Rachel- and parent-moderated interactions while those of the children did not. The validation method was within-session leave-one-utterance-out cross-validation ensuring that the classification was based on the parent's and child's speech style for a given day. If the ECA is to be used in diagnostic or intervention interactions the speech produced by the child when interacting with the ECA should be representative of the speech that he would produce on the given day. Therefore, the achieved result is desirable because it suggests that an ECA can be used to elicit speech from the children representative of their daily speech patterns.



(a) S1- Interaction (b) S1- Question words & backchannels (c) S2- Interaction (d) S2- Question words & backchannels

Figure 1: The parent interaction dynamics for the subject 1 (S1) and subject 2 (S2) experiments.

The parents' speech patterns differed depending on the interaction type (Rachel- or parent-moderated) suggesting that the presence of Rachel in the interaction altered the interaction style of the parents. However, the classification results do not suggest the reasons behind these differences. The interaction patterns of the parents can be better understood graphically (Figure 1).

The interaction patterns from the subject one experiments can be seen in Figure 1(a). This figure shows the percentage of parent turns following an utterance of Rachel vs. following an utterance of the child. The results demonstrate that excepting the final session (during which the subject's younger brother was present) the parent follows the child more often as the sessions progress hinting that the parent is becoming increasingly involved in the dialog. Figure 1(b) supports this claim as well. As the session progresses, the parent asks an increasing number of questions and utilizes an increasing number of backchannels per turn while Rachel is present. Contrastingly, compared to the first session, the use of question words either stagnate or remain constant during the parent-moderated interactions. These findings also support increasing involvement and prompting in the Rachel-moderated interactions.

The interaction patterns from the parent in the subject two experiments can be seen in Figure 1(c). This figure shows that unlike the parent's behavior in the subject one experiments, the parent from subject two does not follow an increasing number of the child's utterances as the sessions progresses. Instead, in sessions two through four the percentage remains approximately constant. As seen in the subject one experiments, the parent's use of question words during the Rachel interaction increases as the sessions progress once again excepting session four (Figure 1(d)). Furthermore, over the course of the four sessions, the parent tends to use more backchannels in the interaction with Rachel as compared to the first session (Figure 1(d)), suggesting that the parent is becoming increasingly a part of the Rachel interaction environment.

The figures provide clues to the parents' interaction differences in the Rachel-moderated vs. parent-moderated interaction. The parents generally ask either a similar number or fewer questions and provide more backchannels in the parent-moderated vs. the Rachel-moderated interaction (Figures 1(b) and 1(d)). This may suggest that in the parent-moderated interactions the parents spend more time following the lead of their child, rather than querying for additional information, which may account for the differences in the parents' speech patterns.

7. Conclusions and Future Work

This paper provides a novel analysis of the nature of the interactions that develop between a parent, child, and an ECA, Rachel. The results suggest that the Rachel environment can be used to elicit natural communicative data from children with autism because their speech patterns do not differ widely between the Rachel- and parent-moderated interactions. This suggests that the data elicited using the Rachel ECA is representative of the child's communication abilities.

One of the limitations of this analysis is its reliance on a small number of subjects. The experiment is being extended to collect data from additional children with autism and their families. Future work includes analyzing the identified trends over a larger dataset.

Our future work includes using more sophisticated lexical modeling, such as topic modeling, to better quantify the word usage of the parent and child and the similarity between the parent and Rachel prompts. This type of modeling will provide more insight into the relationship between the word choice of Rachel and that of the parent and child. This modeling will also be extended to capture aspects of the child-specific grammars employed in the Rachel interactions. Future work will also extend the analysis to the video domain. The child and parent were recorded from three different angles. This recording presents the possibility for analyzing social cues that do not present in the audio or lexical data.

8. Acknowledgements

This work is supported by the National Science Foundation and Autism Speaks.

References

[1] A. P. Association, *Diagnostic and statistical manual of mental disorders (4th ed., text rev.)*. American Psychiatric Publishing, Inc., 2000.

[2] A. Tartaro and J. Cassell, *Using Virtual Peer Technology as an Intervention for Children with Autism*. New York: John Wiley & Sons, 2006, pp. 231–262.

[3] ——, "Playing with virtual peers: Bootstrapping contingent discourse in children with autism," in *International Conference of the Learning Sciences (ICLS)*, Utrecht, Netherlands, 2008.

[4] S. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, 2002.

[5] R. Landa, "Early communication development and intervention for children with autism," *Mental retardation and developmental disabilities research reviews*, vol. 13, no. 1, pp. 16–25, 2007.

[6] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[7] T. Bickmore and J. Cassell, "Small talk and conversational storytelling in embodied conversational interface agents," in *AAAI fall symposium on narrative intelligence*, 1999, pp. 87–92.

[8] G. Bailly, S. Raidt, and F. Elisei, "Gaze, conversational agents and face-to-face communication," *Speech Communication*, vol. 52, no. 6, pp. 598–612, Jun. 2010.

[9] M. P. Black, J. Chang, J. Chang, and S. S. Narayanan, "Comparison of child-human and child-computer interactions based on manual annotations," in *Workshop on Child, Computer and Interaction*, November 2009, pp. 1–6.

[10] E. Mower, M. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *International Conference on Multimedia & Expo (ICME)*, Barcelona, Spain, July 2011.

[11] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *ACII*, Amsterdam, The Netherlands, Sept. 2009.

[13] E. Mower, M. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.