An Acoustic Analysis of Child-Child and Child-Robot Interactions for Understanding Engagement during Speech-Controlled Computer Games

Theodora Chaspari¹,², Jill Fain Lehman¹

¹ Disney Research, Pittsburgh, PA, USA
² University of Southern California, Los Angeles, CA, USA
{theodora.chaspari,jill.lehman}@disneyresearch.com

Abstract

Engagement is an essential factor towards successful game design and effective human-computer interaction. We analyze the prosodic patterns of child-child and child-robot pairs playing a language-based computer game. Acoustic features include speech loudness and fundamental frequency. We use a linear mixed-effects model to capture the coordination of acoustic patterns between interactors as well as its relation to annotated engagement levels. Our results indicate that the considered acoustic features are related to engagement levels for both the child-child and child-robot interaction. They further suggest significant association of the prosodic patterns during the child-child scenario, which is moderated by the co-occurring engagement.

1. Introduction

Engagement is essential to building rapport and enhancing the game experience during child-robot interactions. It refers to the “process of starting, maintaining, and ending the perceived connection between interactors” [1] and is related to enjoyment [2], learning success [3], and social rapport [1]. In order to make human-computer interaction more natural and effective, it is important to efficiently recognize and model engagement, as well as to identify the factors that are able to promote it.

Automatic recognition of engagement involves the use of vocal [4, 5], visual [4, 6], and physiological cues [7], such as prosodic patterns, backchannels, gait, posture, smiles, and electrodermal activity data. Fusing this multimodal information and taking the context of interaction into account can yield person-dependent models able to efficiently recognize and promote engagement [6]. Behavioral coordination has been further associated with increased engagement [4, 8], through the synchronization of a multifaceted set of gestures, gaze, language and vocal expressions between the human and the robot.

Previous studies have explored a variety of social, developmental and cognitive benefits arising from human-human coordination. Parent-infant synchrony has been related to self-control [9], attachment [10], pro-social behavior [11] and language outcomes [12]. In a context similar to that reported here, increased vocal and lexical coordination between children was found to be positively associated with engagement levels [13]. The work in the present paper is a validation of this earlier pilot study, which was performed on a small corpus of game-based interactions between children, and for which the child-robot case had not been analyzed. Evidence of mutual coordination has also been investigated during human-robot interaction in terms of affect exchange [8], body movements and language use [14]. Traditionally in social science, researchers have used statistical models of bivariate time series [15] or cross-correlation [9] to quantify the degree of behavioral coordination. Recent methods have expanded to dynamical systems [16] and cross-recurrence analysis [17]. Speech and language-related studies have examined vocal proximity indices [18] and signal-derived similarity measures of acoustic spaces [19].

We explore the association between engagement and the behavioral coordination of two children as well as a child and a robot during a game-based interaction. While engagement is a multimodal process that combines visual, verbal and non-verbal cues, due to the speech-based nature of the game, we focus on the way acoustic patterns of interlocutors are associated with engagement. Acoustic features include speech loudness and fundamental frequency (F0), selected in a knowledge-driven way. In order to preserve the interpretability of our analysis, we use a linear mixed effects (LME) model for exploring the interplay between the speakers’ prosodic patterns and the degree to which this is moderated by engagement. Our results indicate the existence of a significant association between the considered prosodic features and the annotated engagement levels in both child-child and child-robot interactions. They further suggest significant positive association between the children’s acoustic patterns indicating the presence of acoustic coordination between the two. The later is not apparent in the child-robot scenario due to the random behavior of the robot. This acoustic coordination is further moderated by engagement, i.e. higher coordination is related to higher engagement levels in the child-child case. These findings are discussed in relation to automatic adaptation of the robot’s behavior to the child’s vocal patterns and provide a foundation for robot/computer-assistive games by promoting children’s engagement during such interactions.

2. Data Description

“Mole Madness” (MM) is a language-based, speech-controlled interactive game played by two children, or a child and a robot [20]. It is a computer-based game built to explore language-use, turn-taking and engagement during a fast-paced, speech-based task. Similar to Super Mario Bros® games, MM includes a mole character moving horizontally or vertically through obstacles and rewards using the keywords “go” and
“jump”, respectively. Each player is assigned to one of the keywords at a time and this role alternates between rounds.

Our data contain 62 children (48.4% girls) playing MM in pairs (referred as “child-child interaction”) with mean duration of 391 sec. Their ages ranged between 5-10 years old and each pair had an average age difference of 5.6 months.

Besides playing together, 61 of the children also played one-on-one with Sammy, a back-projected robot head developed by Furhat Robotics [21] (referred as “child-robot interaction”). Sammy’s vocal behavior consists of a set of prerecorded utterances from a female human voice, including a variety of “go” and “jump” expressions with varying prosody, prolongation, and frequency. When Sammy had to play a move corresponding to “go” or “jump”, one of these prerecorded versions was randomly selected. The mean duration of the child-robot interactions was 330 sec.

Data were recorded with two high-definition cameras and two high-precision omni-directional microphones. Snapshots of the child-child and child-robot interactions as well as a screenshot of the game are shown in Fig. 1.

### 3. Methods

#### 3.1. Engagement Annotation

Engagement annotation was performed by three female coders who have experience with children. Coders were asked to rate the “willingness” of the child to continue with the current activity or move to something else. The original video was split to show either the right or the left participant, so each engagement score (ES) was assigned based on the audiovisual record.

#### 3.2. Acoustic Feature Extraction

The keywords (“go” and “jump”) were manually segmented from the audio file. Loudness and fundamental frequency (F0), were computed by openSMILE [23] over each keyword. F0 is measured in Hz, while loudness is computed as the normalized speech intensity raised to the power of 0.3. These are averaged over each 10sec segment that corresponds to an engagement annotation interval. Similar features have been used in previous studies to quantify prosodic patterns [5, 13, 18].

#### 3.3. Linear Mixed Effects Model

Due the multilevel nature of our data, we employ a linear mixed effects (LME) model to quantify the relation between the children’s acoustic measures and the annotated engagement levels. LME is extensively used in social sciences and addresses the problem of violating independence assumptions arising from nested data structures, which is not handled appropriately by traditional ANOVA and multiple regression methods.

We first describe the LME formulation for the two-child game interaction. Let \( Y_{ij} \) be a child’s acoustic measure (i.e. loudness or pitch) from the \( i^{th} \) pair during the \( j^{th} \) time segment and \( X_{ij} \) be the corresponding measure from the other child of the pair. We also denote \( ES_{ij} \) the annotated mean engagement score averaged over the two children for the same time segment. The association between these quantities can be written using the LME formulation as follows

\[
Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 ES_{ij} + r_{ij} 
\]

\[
\beta_0 = \gamma_{00} + u_{0i} 
\]

\[
\beta_1 = \gamma_{10} + u_{1i} 
\]

\[
\beta_2 = \gamma_{20} 
\]

Combining (1)-(4) results in

\[
Y_{ij} = \gamma_{00} + u_{0i} + \gamma_{10} X_{ij} + \beta_2 ES_{ij} + \gamma_{11} X_{ij} ES_{ij} + r_{ij} 
\]  

Based on (5), a child’s acoustic score \( Y_{ij} \) over a time segment is expressed as the sum of a grand-mean acoustic score \( \gamma_{00} \) and a pair-specific mean \( u_{0i} \). Moreover, it depends on the other child’s acoustic score \( X_{ij} \) and the mean engagement value \( ES_{ij} \). The coefficient \( \gamma_{10} \) captures the association between the children’s acoustic features, while \( \beta_2 \) the relation between acoustic measures and engagement levels. Finally, \( \gamma_{11} \) quantifies the effect of engagement score on the association between those two acoustic measures, i.e. positive value of \( \gamma_{11} \) suggests that higher acoustic synchrony is related to increased engagement. The residual term is denoted by \( r_{ij} \).

The same equations hold for the robot-child interaction with the following modification: \( Y_{ij} \) and \( X_{ij} \) are the acoustic features from the child and the robot, respectively, while \( ES_{ij} \) represents the child’s engagement. All LME models include the 10sec segments during which there exists at least one keyword from both interacting partners. The input data of the model were normalized to have a zero-mean.
Patterns of annotated engagement vary between the three coders (Fig. 2). While the average ES value is around 4 for all annotators, the second one depicts the most balanced use of the scale. In contrast, the distribution from the first annotator shifts left and from the third one is skewed to the right. These differences appear more pronounced during the child-child interaction case. Due to these differences, we examine three separate LME models, one per annotator. We further explore the absolute difference of engagement levels between two children over the same 10sec segment (Fig. 3). The corresponding histograms indicate a significant portion of segments over which the two children depict different engagement levels. Taking this into account, we run the LME models for the child-child interaction in two ways: first we consider all segments (referred as “All Segments”) and then we only include the segments for which the absolute ES difference does not exceed a unit (referred as “Similar ES Segments”). The later allows us to examine cases where the two children appear similarly engaged or disengaged, providing further insights on how to apply the findings to the child-robot interaction scenario.

LME results indicate a significant positive association between the acoustic features of two children playing together (γ10) for the child-child interaction, as well as between a child’s acoustic measure and the corresponding engagement level (γ20), as shown in Table 1a. We further observe a significant effect of engagement to the association between the children’s acoustic measures (γ11), i.e., higher engagement corresponds to stronger relation of the acoustic scores. The aforementioned effect appears more significant for the “Similar ES Segments” compared to the “All Segments” case. This finding indicates higher acoustic synchrony when both children are engaged (Fig. 4).

A significant association between a child’s acoustic measure and his/her engagement levels (γ20) is found during child-robot interactions (Table 1b). However, the relation between the child’s and robot’s acoustic features (γ10) is not always significant, since the robot was not designed to adapt to the player’s behavior. This is further supported by the non-significant effect of engagement to the child-robot acoustic score association. Future work will involve appropriately adapting the robot’s vocal characteristics to the child in order to promote engagement during the interaction.

Figure 2: Histograms of engagement scores. Higher values correspond to higher perceived engagement by the annotators.

Figure 3: Histogram of the absolute difference of engagement scores (ES) between the two children of the same pair over each 10sec segment during the child-child interaction.

Figure 4: Predicted values of loudness and fundamental frequency (F0) based on the linear mixed effects (LME) model for low/medium/high engagement score (ES), corresponding to the 5th/50th/95th percentile of ES values, respectively. Results concern the child-child interaction case including the segments over which the absolute ES difference between peers is not more than a unit (“Similar ES Segments”).
Table 1: Linear mixed effects (LME) model estimates for predicting the acoustic features of one speaker (Spk1) based on the corresponding features from the second speaker (Spk2) and the annotated engagement score (ES). “All Segments” refers to all 10sec segments during which there exists at least one keyword from both interacting partners. “Similar ES Segments” includes only the ones over which the absolute ES difference between interactors was not more than a unit.

(a) Child-Child Interaction

<table>
<thead>
<tr>
<th>Feature</th>
<th>LME Estimate</th>
<th>All Segments</th>
<th>Similar ES Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Segments</td>
<td>LME Estimate</td>
<td>LME Estimate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotator 1</td>
<td>Annotator 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotator 1</td>
<td>Annotator 2</td>
</tr>
<tr>
<td>Loudness Spk1</td>
<td></td>
<td>1686</td>
<td>.01 (.04)</td>
</tr>
<tr>
<td></td>
<td>Intercept (γ_{00})</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Loudness Spk2 (γ_{11})</td>
<td></td>
<td>.21** (.03)</td>
</tr>
<tr>
<td></td>
<td>ES (γ_{20})</td>
<td></td>
<td>.11** (.01)</td>
</tr>
<tr>
<td></td>
<td>Loudness Spk2: ES (γ_{11})</td>
<td>.06* (.02)</td>
<td>.03 (.02)</td>
</tr>
<tr>
<td>F0 Spk1</td>
<td></td>
<td>18.26** (2.45)</td>
<td>13.89** (1.82)</td>
</tr>
<tr>
<td></td>
<td>Intercept (γ_{00})</td>
<td></td>
<td>.38** (.02)</td>
</tr>
<tr>
<td></td>
<td>F0 Spk2 (γ_{11})</td>
<td></td>
<td>.13 (.03)**</td>
</tr>
<tr>
<td></td>
<td>ES (γ_{20})</td>
<td></td>
<td>.05* (.02)</td>
</tr>
</tbody>
</table>

(b) Child-Robot Interaction

<table>
<thead>
<tr>
<th>Feature</th>
<th>LME Estimate</th>
<th>All Segments</th>
<th>Similar ES Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Samples</td>
<td>LME Estimate</td>
<td>LME Estimate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotator 1</td>
<td>Annotator 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Annotator 1</td>
<td>Annotator 2</td>
</tr>
<tr>
<td>Loudness Spk1</td>
<td></td>
<td>1376</td>
<td>.68 (.03)**</td>
</tr>
<tr>
<td></td>
<td>Intercept (γ_{00})</td>
<td></td>
<td>.01 (.05)</td>
</tr>
<tr>
<td></td>
<td>Loudness Spk2 (γ_{11})</td>
<td></td>
<td>.05 (.04)**</td>
</tr>
<tr>
<td></td>
<td>ES (γ_{20})</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Loudness Spk2: ES (γ_{11})</td>
<td>-0.2(0.04)</td>
<td>-0.9 (.04)**</td>
</tr>
<tr>
<td>F0 Spk1</td>
<td></td>
<td>1376</td>
<td>229.87(7.18)**</td>
</tr>
<tr>
<td></td>
<td>Intercept (γ_{00})</td>
<td></td>
<td>.13 (.03)**</td>
</tr>
<tr>
<td></td>
<td>F0 Spk2 (γ_{11})</td>
<td></td>
<td>10.63(1.67)**</td>
</tr>
<tr>
<td></td>
<td>ES (γ_{20})</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F0 Spk2: ES (γ_{11})</td>
<td></td>
<td>.04 (.03)</td>
</tr>
</tbody>
</table>

5. Discussion

In this paper we modeled the interaction between acoustic patterns of children while playing a language-based computer game and its relation to the annotated engagement levels. We further extended our analysis to the child-robot interaction scenario, where each child played the same game with a robotic partner instead of a peer. Our results indicate significant associations between the acoustic measures of the two children, which appears stronger for the high engagement segments. Similar findings in the literature suggest that child-adult synchrony can be related to positive affect [24], attachment [10], and social behavior [9, 11]. These results are not as prominent for the child-robot interaction, since the robot’s behavior was not designed to follow the child’s acoustic patterns. Future work will attempt automatic robot adaptation to address this challenge.

A potential way to make the child-robot interaction more engaging would be to automatically synthesize the robot’s voice in order to achieve the desired prosodic patterns of loudness and pitch. For example, our results (Section 4) indicate that increased loudness from one child is related to increased engagement, which is further associated to increased loudness from the interacting peer. Therefore we will attempt to build a system that measures the children’s vocal patterns and synthesizes the robot’s voice in order to match those. A variety of prosody synthesis studies [25, 26] suggest the feasibility of this approach, which might be able to enhance rapport and promote engagement between the child and the robot [8, 14]. This can be useful in rendering robots social partners and peer tutors for children, as well as home companions [27, 28, 29].

6. Conclusions

We explored the acoustic patterns of children during a speech-based computer game and their relation to engagement. Our results indicate a significant association between the two children with respect to loudness and fundamental frequency, which is further moderated by the annotated engagement levels. This moderation effect appears larger if we only include instances over which both children have similar engagement levels. Such significant moderation was not apparent in the child-robot interaction, which can be justified by the fact that the robot’s behavior was random and not adapted to the human peer. These results provide a foundation for ways to build engagement and social rapport to enhance child-robot interactions.

Future work will expand the aforementioned analysis to visual cues and explore time-based models for predicting engagement and quantifying synchrony during such interactions.

7. Acknowledgments

The authors would like to thank Dr. Samer Al Moubayed and Dr. Iolanda Leite for their help in analyzing the data.

8. References


