# Creating Prosodic Synchrony for a Robot Co-player in a Speech-controlled Game for Children

Najmeh Sadoughi[1,2], André Pereira[1], Rishub Jain[1,3], Iolanda Leite[1], Jill Fain Lehman[1]
1. Disney Research, Pittsburgh, USA
2. University of Texas at Dallas, Texas, USA
3. Carnegie Mellon University, Pittsburgh, USA
nxs137130@utdallas.edu

## ABSTRACT

Synchrony is an essential aspect of human-human interactions. In previous work, we have seen how synchrony manifests in low-level acoustic phenomena like fundamental frequency, loudness, and the duration of keywords during the play of child-child pairs in a fast-paced, cooperative, language-based game. The correlation between the increase in such low-level synchrony and increase in enjoyment of the game suggests that a similar dynamic between child and robot co-players might also improve the child's experience. We report an approach to creating on-line acoustic synchrony by using a *dynamic Bayesian network* learned from prior recordings of child-child play to select from a predefined space of robot speech in response to real-time measurement of the child's prosodic features. Data were collected from 40 new children, each playing the game with both a synchronizing and non-synchronizing version of the robot. Results show a significant order effect: although all children grew to enjoy the game more over time, those that began with the synchronous robot maintained their own synchrony to it and achieved higher engagement compared with those that did not.

## 1. INTRODUCTION

Human-human interaction is more than just exchanging messages explicitly. There are other para-linguistic cues that facilitate the interaction by creating a sense of rapport and bonding between the interlocutors. Among these cues are nonverbal behaviors such as gestures, postures, and prosody [3, 13]. Studies have shown that there are subtle cues of synchrony between prosodic features of speech from both parties in a conversation [13, 15, 25], and that coordination between dyads is associated with a positive effect on engagement, even very early in life [9,17]. The positive effect of synchrony on interactions between people suggests that we might make human-robot interaction more natural and engaging by incorporating similar mechanisms in a robot partner and making it responsive not only to the content, but also the para-verbal behavior of the user.

We explore this possibility in the context of a simple, language-based video game called *Mole Madness* (MM). Previous work examining the behavior of child-child and child-robot co-players in MM showed two related phenomena. Chaspari et al. [6] investigated children's level of engagement during game play with each other[1]. It was found that the association of speech features between the two children - including fundamental frequency (F0), loudness, and word duration - was stronger during the times that the children were judged to be highly engaged in the game. When the analysis was extended to child-robot play with the same set of children [7], there was no coordination of prosodic features and lower engagement scores overall.

The work reported here tests the hypothesis that children's experience with the robot can be improved by better approximating the synchrony found in more engaged child-child pairs. In particular, we use the data from children studied in [6] to train a *dynamic Bayesian network* (DBN) that captures the strong correlations for F0 and loudness. The model is used during gameplay to select the robot's prosodic features adaptively, based on the child's values. Although rate of keyword speech did not reach statistical significance in [7], the trend in the data was strong enough to suggest adding that type of verbal synchrony to the robot's behavior as well. Our proposed system uses a $k$ nearest-neighbor approach in the multi-dimensional space of F0, loudness, and repetition to choose an utterance for the robot when a game action is required. In an experiment with a new set of children, each child played multiple game levels with both synchronizing and non-synchronizing versions of the robot, with order of condition counterbalanced across children. The results showed both the dyadic nature of synchronization and its profound effects. Children who played with the synchronizing version of the robot first maintained their prosodic synchrony to it even when it stopped maintaining its prosodic and verbal synchrony to them, and in this order engagement rose steadily across the session. In contrast, children who began with the non-synchronizing robot co-player never reached the same level of engagement over time, despite the robot's eventual adoption of synchronous speech.

After reviewing related work by others (Section 2), we describe prior work with *Mole Madness* in detail, both to explain the properties of the data we used to train our robot's



*HRI '17, March 06 - 09, 2017, Vienna, Austria*



ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020244

[1]Technically, the scales measure child's degree of willingness to continue to play, as judged by adults with extensive experience with young children.

(a) *Mole Madness* (b) Sammy-Child play

**Figure 1: Snapshots of the game and participants.**

new adaptive behavior and to create a baseline for analyzing synchrony (Section 3). We then turn to the method for training (Section 4), describe deployment of the models in a real-time adaptive implementation (Section 5), and present results from a new set of child co-players (Section 6).

## 2. RELATED WORK

Several authors have investigated synchrony between prosodic features of interactants and its effects on conversation. De Looze et al. [15] studied the relationship between mimicry of prosodic speech features and the level of involvement of people in a conversation. Mimicry strength was measured by correlations among several features throughout the conversation, and their results showed that level of involvement is positively correlated with the coordination of speakers' prosodic cues. Suzuki and Katagiri [25] conducted a similar experiment to see whether humans entrain to prosodic features while communicating with a computer. They manipulated prosodic features of the speech generated by the computer such as loudness and pause duration, and observed the user's response. The results indicated that users entrain to some extent to the speech provided to them, e.g., they produced louder sounds as the volume from the computer increased. Our approach extends this work not only by exploring other speech related features such as F0, but also by using a robot instead of disembodied computer-generated speech, and by examining the phenomena in interactions with children.

Most HRI studies about synchrony have been focused on rhythmic adaptation [4]. Michalowski and colleagues [18], for example, studied the effects of the synchronous and non-synchronous behaviors of a Keepon robot that was able to dance in coordination with music and children's movements. They found that children's interaction with the robot was positively affected by the robot's responsiveness to their actions. In a more recent laboratory study using the same robot [19], the authors found contradictory results when measuring children's retention (i.e., willingness to continue interacting with the robot) while dancing with a synchronous robot and a non-synchronous one. The authors attribute the mixed results to limitations in their rhythmic perception system, and discuss the challenges of measuring children's engagement in playful interactions. In the same line of research, Avrunin et al. [2] investigated people's impressions of agency and life-likeness of dancing robots. In their study, adult participants judged videos of dancing robots with regard to dance quality, life-likeness and entertainment value. While life-like motion was considered more entertaining, the results suggest that perfect synchrony (i.e., robot movement always matching the sound) is less life-like than a situation in which the robots are not always in sync. More recently, Hoffman and Vanunu [11] conducted a study where participants listened to music in the presence of a robot that was moving in sync with the music, off beat, or not moving at all. Despite not being aware of the beat precision, participants interacting with the synchronizing robot rated the songs more positively, and provided higher ratings in perception traits like human-likeness and similarity.

Another related area is psycho-motor alignment, wherein research typically focuses on the temporal relations between the actions of two or more agents (humans or robots) performing an activity together. In this domain, Prepin and Gaussier [22] proposed an architecture that enables a robot to move its arms in synchrony to the arms of a human user. The convergence of their reinforcement learning algorithm is a sign that the robot successfully learned to synchronize its movements to those of the user. Iqbal et al. [12] investigated psycho-motor entrainment in the context of human-robot teamwork. They presented an event-based model to enable robots to measure synchronous motion between humans, with the ultimate goal of enabling fluid joint action between robots and groups of people. Using data collected from mobile robots sensing pairs of humans marching synchronously and non-synchronously, the authors showed that their model can accurately detect synchronous motion. We use a method that is analogous to theirs but in a different modality and with continuous phenomena.

## 3. UNDERSTANDING SYNCHRONY IN MM

Our work extends previous results with children playing the same game. Indeed, the method for building a synchronizing version of the robot co-player, Sammy, relies on data collected during those earlier studies, and the non-synchronizing play observed in those games establishes a baseline against which to evaluate our results. In this section we review the prior work in detail to clearly distinguish both what is new and why.

### 3.1 The Game

*Mole Madness* (MM) is a speech-controlled, interactive

side-scroller in which two players move a mole through its environment, avoiding obstacles and gaining rewards [14]. Effective play requires coordinated use of the keywords "go" and "jump," which control horizontal and vertical motion, respectively (see Figure 1(a)). Each participant is responsible for one keyword/direction at a time, but switches roles between levels. The game is designed to be easy to learn by children as young as four, but still fast-paced enough to be fun for children who are nine or ten.

MM can be played by two children (CC), or by a child and Sammy (SC), a back-projected robot head designed by Furhat Robotics [1] that has been set in a cardboard body to sit next to the child in a more peer-like way (Figure 1(b)). An overall architecture controls the multiple parallel processes for the game, the robot, and the custom word spotter that performs keyword recognition. The game is programmed in Unity3D, and Sammy plays by accessing an $A^*$ search algorithm that returns a go/jump decision based on the next move along an optimal path. Sammy's vocal space of utterances consists mainly of a set of pre-recorded keyword files that vary with respect to prosodic features, durations, and frequency, although the robot also has some social speech that can be deployed at various points in the environment when gameplay allows.

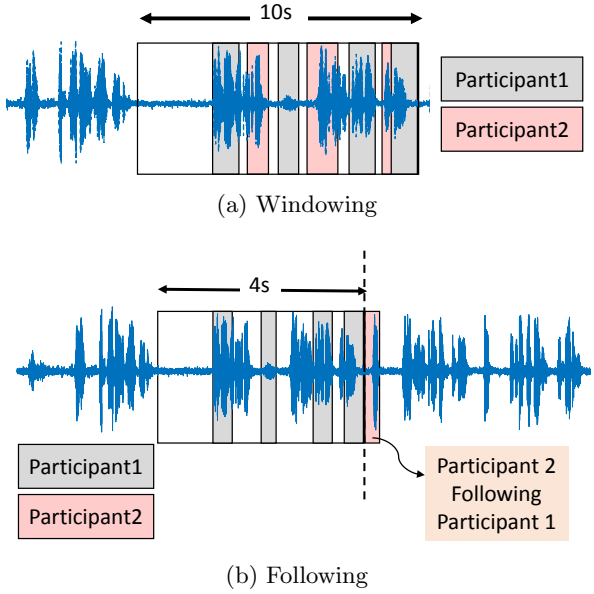
## 3.2 Data Set 2015

Although MM has been used and reported on across many data collections, the data set that forms the basis of the models described in Section 4 (hereafter, DS15) reflects play by 86 children (51.16% girls), ages four through nine, in 2015. Children in DS15 played MM first in child-child pairs, then one-on-one with Sammy, producing 43 CC games with a mean duration of 355 seconds, and 85 SC games (one of the children did not play with the robot) with a mean duration of 216 seconds. Data was recorded by two high resolution cameras and a high-precision, omni-directional microphone. Instances of "go," "jump," and non-keyword social speech were segmented and annotated by hand.

## 3.3 Analysis of Synchrony in DS15

**Windowing.** Following Chaspari et al., we examine the correlation between speech and prosodic features over 10-second window intervals (see Figure 2(a)). Using the human annotations for the keywords "go" and "jump," acoustic features for loudness and F0 are extracted using openSMILE [8], separately for each keyword. Because the system uses only one microphone, we exclude portions of the segment where the annotations indicate voice overlap in order to avoid cross-participant contamination of the values.[2] If a window contains multiple instances of the same keyword, the values of each feature are averaged across the 10-sec interval. We also derive an additional feature that encodes separately for each keyword the number of times that keyword was uttered in the segment (a point we will return to in Section 4). Pearson correlation ($r$) tests between the features extracted for each player in the same window of

[2]The custom word spotter avoids the problem of sound localization by explicitly modeling keyword overlap. Because children are assigned one keyword on each level and moving the mole requires cooperative use of both commands, it is almost always the case that different keywords belong to different voices. The small number of occasions when a child usurps the other player's role in the excitement do add some noise to data.




(a) Windowing

(b) Following

**Figure 2: Two strategies for analyzing the amount of synchrony between participants. In (a), average values for prosodic features are computed separately for each player's non-overlapping speech within a ten second window and correlation is computed for windowed pairs. In (b), the influence is assumed to be limited to the duration of the echoic buffer and the segment containing a non-overlapped instance of player 2's keyword is paired with all of player 1's non-overlapped speech in the previous four seconds.**

analysis showed higher correlation percentages in the CC sessions ($r_{F0}\,(11486) = 61.66, p < 0.001$, $r_{loudness}\,(11486) = 67.34, p < 0.001$, $r_{\#keywords}\,(11486) = 56.28, p < 0.001$)[3] than in the SC games ($r_{F0}\,(18177) = 5.69, p < 0.001$, $r_{loudness}\,(18177) = 13.40, p < 0.001$, $r_{\#keywords}\,(18177) = 14.26$, $p < 0.001$). Considering that Sammy's keywords were randomly selected from a small pool of pre-recorded keywords, the lower correlation values in the SC pairs are not surprising.

**Following.** An alternative way to analyze synchrony constrains player 1's influence on the prosodic features of player 2's keyword to the duration of player 2's auditory short-term memory [5]. Under this method, depicted in Figure 2(b), we produce paired values for each non-overlapping keyword with the average values for non-overlapping instances of the other keyword in the prior four seconds. Such pairs implicitly define two different "follow" relations: the subset of instances in which player 1 follows player 2 and the subset of instances in which player 2 follows player 1. Thus we can separately compute correlations when child follows child from the CC sessions and when child follows Sammy and Sammy follows child from the SC sessions of D15. We find that correlations between pairs of segments where Sammy follows the child ($r_{F0}\,(5656) = 0.019, p = 0.24$ and $r_{loudness}\,(5656) = 10.63, p < 0.001$) and where the child follows Sammy ($r_{F0}\,(6962) = 4.35, p < 0.001$ and $r_{loudness}$

[3]The Pearson correlations are reported as r(#instances), p-value.

$(6962) = 15.60, p < 0.001$) are lower than when the two children are following each other ($r_{F0}\,(6098) = 46.44, p < 0.001$ and $r_{loudness}\,(6098) = 52.90, p < 0.001$).

These results indicate that Sammy's vocal behavior was not synchronous with the child during the game, especially in terms of F0. Although the analysis shows a positive correlation of 10.63 when Sammy follows the child, this value is much smaller than the average correlation value when a child follows another child ($r = 46.44$), and may be the result of low variability in the loudness of Sammy's small set of pre-recorded keywords. Note that the correlations derived from the *following* method show the same overall trends as the correlations derived from *windowing* but give us a more cognitively-motivated approach to understanding the child's behavior in response to Sammy's.

# 4. SYNCHRONIZING SAMMY

Because greater engagement was seen in child-child pairs who were in sync, we consider their verbal behaviors to be a model of how Sammy should behave as a peer co-player with the goal of creating an enjoyable game experience. Thus, we take advantage of the child-child data in D15 to synthesize more synchronized verbal behaviors for Sammy. The process has two phases: building the entrainment model that captures the strong correlation of acoustic features in the CC sessions and deploying the model in real-time play.

## 4.1 Modeling Entrainment

We chose to model the children's coordination with a dynamic Bayesian network (DBN) learned from the child-child sessions in the DS15 corpus. DBNs are a type of statistical model that has been shown to be able to capture the strong correlations between time series. DBNs also learn the possible dependencies between consecutive frames as their transition probabilities, and have been used in the past to generate head movements synchronized with speech prosodic features [16, 23, 24].

The relationship Sammy's Bayesian graph must capture can be seen in Figure 3. The nodes $F0_{c1}$, $L_{c1}$, $F0_{c2}$, and $L_{c2}$ represent observed continuous variables for F0 and Loudness of two children playing together, which are modeled by Gaussian distributions. The two nodes affiliated with c2 are only observed during training, and are synthesized by the model during testing and play. The nodes $H_{c1,c2}$ are discrete variables representing the hidden states. The hidden states model the possible joint configurations between the prosodic features of the first and second child. For instance, if high F0 in one child is usually accompanied with high F0 in the other child, the hidden states learn this association. These discrete variables are trained to capture the correlation between the input modalities, and exploit that during synthesis.

For the model, we assume that the transition probabilities follow a Markov property of order one, i.e., they depend only on one previous time step. We choose a one second time step for responsiveness; this step size allows Sammy to adapt, close to real time, to changes that are happening with the features of the child. Although children may only be sensitive to prosodic features in the prior four seconds of speech, Sammy does not need to have that limitation. More importantly, by extending the window for computing features to 10 seconds, as was done in the windowing analysis, we increase the number of training instances available to

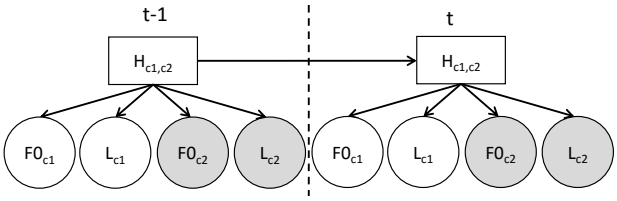



**Figure 3: The Bayesian graph of the model for entrainment, where L is loudness, and the subscripts c1 and c2 refer to the first and second child.**

the model from the DS15 corpus. For instances where there are no keywords from the other player even in the previous ten seconds, we use a fixed F0 and loudness based on the speaking child's averages in DS15.

The DBN comprises the prior probabilities of the hidden states, the transition probabilities between the states, and the observation probabilities given each hidden state. We optimize all these parameters using an Expectation Maximization algorithm (EM). To use the EM algorithm, we first derive the probabilities of the states given the observed sequences, by running the forward-backward algorithm [16,20] (E-step). Next, we maximize the likelihood of the model by updating all the parameters (M-step). The details of training a DBN are provided in [20].

Note that, during testing, we do not have access to the future data, and therefore, we run only the forward algorithm [20]. Given the observation vector until time $t$, the forward path gives us $\alpha_{i,t}$, which is the probability of the $i^{th}$ hidden state ($\alpha_{i,t} \triangleq P\left(H_{c1,c2_t} = i | y_{1:t}\right)$). Given a sequence of features for one of the players, the following equations calculate the expected values of the features for the other player (in our case, the robot), where $\mu_{F0_{i,2}}$, and $\mu_{L_{i,2}}$ are the mean of the F0 and loudness for the $i^{th}$ hidden state for the second player.

$$
\begin{aligned}
E\left[F0_{c2,t} | \, F0_{c1,1:t}, L_{c1,1:t}\right] &= \sum_{i=1}^{n} \alpha_{i,t} \times \mu_{F0_{i,2}} \\
E\left[L_{c2,t} | \, F0_{c1,1:t}, L_{c1,1:t}\right] &= \sum_{i=1}^{n} \alpha_{i,t} \times \mu_{L_{i,2}}
\end{aligned}
\tag{1}
$$

During testing with the entrainment model, we provide the DBN with F0 and loudness of the first child, and get the predicted values for F0 and loudness for the second child (which is Sammy), but there is no guarantee that Sammy's sound files have an instance of the keyword with values that exactly match the predictions.

### 4.1.1 Optimizing the model on DS15

Using the entrainment model, we aim to capture the strong correlation which is demonstrated in D15 recordings between children in the CC sessions. Utilizing this correlation will allow us to generate prosodic features for Sammy which are synchronous to the child. Overall, we used 11486 (10-s) windows for predicting child1 from child2, and because Sammy can be in either role, concatenated them with the same data reflecting child2 predicting child1. Therefore, we used 22972 samples for training in total.

Since the aim of this model is to predict the verbal features, we measure the R-squared between the predicted fea-

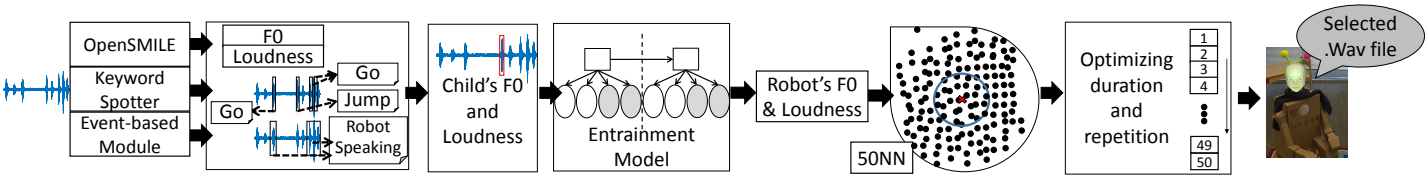



**Figure 4: Overview of the real-time system.**

tures and the original ones. We use the average of the R-squared for the two features as the metric to find the optimal number of states ($R^2_{o,p}$) by running a *Leave One Pair Out* (LOPO) cross validation – in every loop we use one fold for testing, one fold for validation, and the remaining 41 folds for training. The LOPO cross validation metrics are given in Table 1, where $r_{o,p}$ is the Pearson correlation between the original and predicted features, $r_{o_{c1},p_{c2}}$ is the Pearson correlation between the original feature for the first child, and the predicted feature for the second child. The result shows that our entrainment model is able to capture the high correlation displayed in the data. The average of the chosen number of states across folds is [15.6] = 16. Thus, the model trained with all the data, to be used by Sammy in real-time play, uses 16 states.

## 4.2 On-line adaptation for keywords

The entrainment model outputs, in real-time, an estimation of the F0 and loudness that the robot should approximate in its own speech in order to behave in synchrony with the child. This section describes the implementation details that enable the robot to perceive the child's prosodic features in real-time and make its own keyword selection based on the entrainment model. An overview of this pipeline is shown in Figure 4.

The first step of the process consists of segmenting children's speech and excluding the overlapping segments where the child and the robot are speaking in parallel. This is done by combining information from an event-based module that keeps track of when the robot begins and finishes speaking and a real-time keyword spotter that recognizes the keywords "go" and "jump" with 89% accuracy [10]. For the recognized keyword segments not belonging to Sammy, the system then extracts F0 and loudness in real time with the port-audio version of OpenSMILE [8]. Whenever Sammy receives a request from the game module to issue a game keyword, the system uses the most up to date F0 and loudness values extracted from the child's speech segments (recall that we use a 10-s window cache) as the input for the entrainment model, which in turn outputs the robot's desired F0 and loudness.

The next step is to select a sound file with the most similar features to the ones given by the entrainment model. The robot has available a pool of pre-recorded keyword samples, including 1923 samples for "jump" and 1600 samples for "go" with different volumes, F0 levels, and keyword durations (e.g., elongated and rapid keywords). From this pool, we select the 50 nearest neighbors of the features predicted by the entrainment model and create a list of candidate audio files.

We then sort the 50 candidate audio files according to how close their total duration is to the duration of the child's speech segment, because in the child-child data we observed that the children's total duration of keywords was correlated with each other ($r$ = 47.46% and their average distance = 1.062). Finally, we select from the ordered list the top file that has not been used in the past 20 seconds in order to ensure variability in the choices the child hears. In the unlikely situation that all the keyword files from the 50 NN have been played in the past 20 seconds, we ignore this rule and play the best fit.

**Table 1: Evaluation metrics on the predicted F0 and loudness with the entrainment model of 86 children.**

| Metric | F0 [%] | Loudness [%] |
|---|---|---|
| $R^2_{o,p}$ | 46.31 | 42.26 |
| $r_{o,p}$ | 69.53 | 66.24 |
| $r_{o_{c1},p_{c2}}$ | 95.33 | 89.30 |

# 5. EVALUATION

To evaluate the impact of the entrainment model during real-time game play, we invited a new group of children (hereafter, DS16) to play *Mole Madness* who had never played the game before. In a repeated-measures design, each child played with two different versions of Sammy: *synchronizing* and *non-synchronizing*. The two versions differ in the algorithm for selecting the sound file for the robot's speech when the game module issues a go/jump command. In the synchronizing version, the system chooses a sound file from the 50 nearest neighbors according to the process described in the previous section; in the non-synchronizing version, Sammy's sound file is randomly selected from all possible files **other than** the 50 nearest neighbors. We counterbalanced the order of conditions across participants: children who started by playing with the synchronizing version of the robot and then switched to the non-synchronizing version are labeled as being in the **SN** condition, while children who played with the non-synchronizing version of Sammy first are labeled as **NS**.

## 5.1 Participants

We recruited 40 new children (50% girls) via postings in physical and online community boards. The children's ages ranged from 4 to 10 years old ($M$ = 6.73 years, $SD$ = 1.72). Twenty-one of the children were assigned to the SN condition ($M$ = 7.07 years, $SD$ = 1.73) and the rest (19) to the NS condition ($M$ = 6.36 years, $SD$ = 1.69). Both conditions were gender-balanced.

## 5.2 Procedure

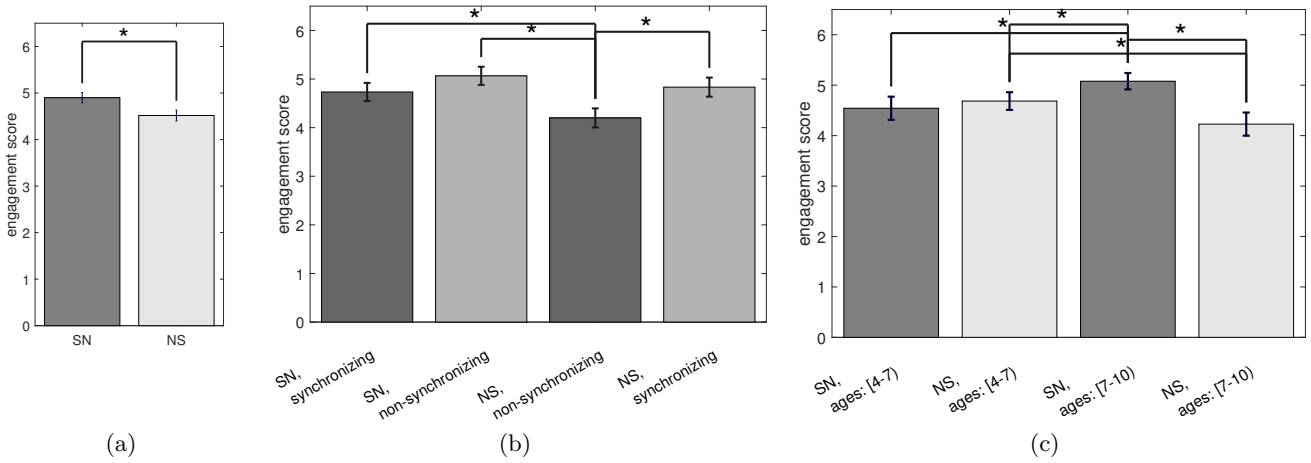
To put the children at ease, an experimenter began the session by introducing the child to Sammy and letting him/her

**Table 2: Percentage of correlation achieved using entrainment (Sammy follows child) in DS16. Values marked with asterisks are statistically different from zero ($p < 0.001$)**

| Condition | | SN | | NS | | – | |
|---|---|---|---|---|---|---|---|
| Robot behavior | | Synch. | Non-Synch. | Non-synch. | Synch. | All Synch. | All Non-synch. |
| #Samples | | 474 | 634 | 482 | 435 | 909 | 1116 |
| $r$ [%] | F0 | 66.25* | 6.31 | 3.14 | 68.08* | 66.99* | 5.06 |
| | Loudness | 36.39* | -0.00 | 5.69 | 33.90* | 32.17* | -2.43 |

**Table 3: Percentage of correlation for the child following Sammy in DS16. Values marked with asterisk are statistically different from zero ($p < 0.01$)**

| Condition | | SN | | NS | | – | |
|---|---|---|---|---|---|---|---|
| Robot behavior | | Synch. | Non-Synch. | Non-synch. | Synch. | All Synch. | All Non-synch. |
| #Samples | | 681 | 793 | 661 | 648 | 1329 | 1454 |
| $r$ [%] | F0 | 40.38* | 9.53* | 5.80 | 23.35* | 32.57* | 8.01* |
| | Loudness | 32.72* | 14.04* | 2.30 | 20.75* | 27.27* | 9.02* |




(a) (b) (c)

**Figure 5: Engagement scores of children in DS16 from multiple views. Asterisks denote significance ($p < 0.05$).**

personalize the robot's appearance using a variety of accessories. Next, the child watched a brief video tutorial about how to play the game, then was told by the experimenter that Sammy had two different ways to play the game and together they would play both.

Each child played a total of four levels of the game, alternating between the "go" and "jump" roles. Depending on the experimental condition to which the child was assigned, Sammy played the first two levels using the synchronizing or non-synchronizing version, then after a short break, switched to the other mode for the third and fourth levels. To ensure the same difficulty in both conditions, the first two levels were exactly the same as the last two but had different backgrounds to make the repetition of the levels less conspicuous to the child. Data from these sessions were recorded by two high-resolution cameras and a high-definition microphone for future analysis.

## 5.3 Results and Discussion

We analyzed the recordings in DS16 to explore possible effects of the robot's synchronizing and non-synchronizing behaviors. We first calculated the amount of synchrony that we actually created in the robot when it played with the children. Next, we measured the amount of synchrony elicited from the children by different versions of Sammy. Finally, we investigated the relationships among the children's engagement scores, the versions of Sammy (synchronizing vs. non-synchronizing), the order of the games (first two rounds vs. second two rounds), and the age of the children.

### 5.3.1 Analysis of Synchrony in DS16

We replicated the *Following* methodology described in Section 3.3 to analyze the audio data collected in the D1S6 experiment and compared it with the DS15 results. F0 and loudness were extracted for all of the non-overlapping keywords from both Sammy and the child and paired in a similar manner as in Section 3.3. We then sorted the keyword pairs into two groups: Sammy follows Child and Child follows Sammy. The Pearson correlation ($r$) percentages, organized by experimental condition and robot's behavior, are displayed in Tables 2 and 3[4].

The results for the instances in Table 2, where Sammy follows the child, serve as a basic assessment of the online entrainment method in real play. While the synchronizing games show significant high correlations for both F0

[4]The results for the 10-s *Windowing* method revealed exactly the same trends as the ones we report for *Following*.

and loudness, the correlations are not significant for either prosodic feature in the non-synchronizing games. As mentioned previously, we cannot guarantee "perfect synchrony" unless we synthesize the voice or have available a large enough number of sound files that the robot can always find a perfect match for every combination of features. Nevertheless, these results indicate that our method was successful.

The most interesting results, however, are the degree of correlation for the instances of the child following Sammy, shown in Table 3. When children played with the synchronizing version of the robot, our features of interest were significantly positively correlated, regardless of the order that children were assigned (SN or NS). However, the patterns for the non-synchronizing games are more complex: when children began by playing with the non-synchronizing version of Sammy (NS condition), the prosodic features of the child and the robot are not significantly correlated in those non-synchronizing levels, but when they started by playing with the synchronizing version of Sammy (SN condition) and then switched to the non-synchronizing version, significant positive correlations in those non-synchronizing levels remained. These results suggest that children maintain their prosodic synchrony with the robot for some time even after the robot stops synchronizing its prosody with them.

### 5.3.2 Analysis of Engagement

We asked three annotators with extensive experience in behavioral analysis (including coding more than 100 children in MM in previous years) to judge the children's levels of engagement during the game. The annotators were blind to our research questions. They were asked to watch videos of the DS16 MM sessions with only the child visible and segment the video according to a seven-point scale describing the child's willingness to continue to play. Four of the scale's ratings were labeled; 1 as *ready to do something else*, 3 as *could take it or leave it*, 5 as *very much into the game* and 7 as *can't drag him/her away*, while ratings 2, 4, and 6 were unlabeled. One level could have multiple annotations with different ratings; thus, to derive the level's engagement score, we performed a weighted average of the ratings given by an annotator based on the duration of the ratings in the entire game. Because each child played four levels of the game, annotation of DS16 produced 4 engagement scores$\times$40 children$\times$3 annotators.

The Cronbach's alpha between the ratings given by the three annotators is 0.79. Note that randomly selecting one of the coders when their agreement is high or averaging the ratings of multiple coders are both valid approaches for analyzing behavioral data. Despite the high agreement between DS16 coders, our past experience with engagement coding revealed that often coders use the full range of the coding scale differently. Therefore, similar to the previous studies with MM [6, 7], we analyzed the results for each annotator separately and found the same trends for all three coders. Hence, we present the results for one coder in the remainder of the section to avoid redundancy.

Overall, children in the SN group had significantly higher engagement scores ($M = 4.90, SD = 0.73$) than children in the NS group ($M = 4.52, SD = 0.67$), $t(158) = -3.4316, p < 0.001$ (see Figure 5(a)). Given that the only difference between the two conditions was the order in which children played with the two versions of the robot, we further investigated these results by conducting a two-way ANOVA considering the version of Sammy (synchronizing vs. non-synchronizing) collapsed across conditions, and the order of the games (first two rounds vs. second two rounds) as within-subjects factor. The analysis of variance showed a significant main effect for order, $F(1, 156) = 21.086, p < 0.001$, but no significant main effect for the version of the robot (synchronizing vs. non-synchronizing), $F(1, 156) = 2.06, p = 0.51$. We also found a significant interaction effect between robot version and the order of the rounds, $F(1, 156) = 13.307, p < 0.001$. As depicted in Figure 5(b), while children who began playing with the synchronizing version of Sammy sustained their engagement levels when switching to the non-synchronizing games, children who began playing with the non-synchronizing version of Sammy showed lower engagement when playing in this mode and ended their session in a less engaged state. This result strengthens the findings obtained in the prosodic feature correlation showing that when first exposed to a synchronizing robot, children maintain their behavior for some time even when they then interact with a non-adaptive robot.

Finally, since previous research on child developmental theory suggests that language interactions can be affected by age differences [21], we looked at potential age effects in our data. We divided children into two age groups based on their developmental stage [21]: less than 7 years old, and greater than or equal to 7 years old. We have 19 children in the younger age group and 21 children in the older age group. The average engagement scores in these four conditions is given in Figure 5(c). A two-way ANOVA revealed significant differences in engagement between the NS and SN experimental groups, $F(1, 156) = 10.209, p = 0.0017$, and a significant interaction effect between condition and age, $F(1, 156) = 17.766, p < 0.0001$. Pairwise comparisons revealed a significant age difference between SN and NS conditions among the older age group ($p < 0.0001$), but no significant differences for the younger children ($p = 0.892$).

## 6. CONCLUSION

In this paper, we proposed a framework to create synchronous verbal behavior for a social robot, Sammy, playing a fast-paced, speech-based game with a child. Analysis of the gameplay of 86 children in 2015 showed a strong correlation between the acoustic characteristics of children playing in pairs. Using the data from those children, we built a DBN model that learns the joint representation of prosodic speech features in those child-child pairs. In real-time play of the game with Sammy as co-player, we measure the prosodic speech features and keyword duration from the child, use the DBN to predict synchronous prosodic values for Sammy, and select the nearest match for the prosodic features and keyword duration from a large but fixed space of possible utterances. To test the performance of the method, we recorded data from 40 new children, each of whom played with both a synchronizing and non-synchronizing version of Sammy, balanced for order. We analyzed their gameplay with subjective and objective metrics, both of which show that the the order of conditions matters. Objectively, children who started with the synchronizing version of Sammy showed more synchronous behavior, even in the non-synchronizing levels, compared to the children who started with the non-synchronizing version. Subjectively, children who started with the synchronizing version ended the session with higher engagement levels compared with those that started with

the non-synchronizing version of the robot. Moreover, the results showed an age effect, demonstrating that most of the engagement result was due to differences in the older children under the two order conditions. It remains for future work to see whether entrainment of the same paralinguistic features in other interaction scenarios will produce the same patterns of behavior and enjoyment.

# 7. REFERENCES


[1] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. Furhat: A Back-Projected Human-Like robot head for multiparty Human-Machine interaction. In *Cognitive Behavioural Systems*, Lecture Notes in Computer Science, pages 114–130. Springer Berlin Heidelberg, 2012.

[2] E. Avrunin, J. Hart, A. Douglas, and B. Scassellati. Effects related to synchrony and repertoire in perceptions of robot dance. In *Proceedings of the 6th international conference on Human-robot interaction (HRI)*, pages 93–100. ACM, 2011.

[3] W. S. Condon and W. D. Ogston. Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, 143(4):338–347, 1966.

[4] C. Crick, M. Munz, and B. Scassellati. Synchronization in social tasks: Robotic drumming. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 97–102. IEEE, 2006.

[5] C. J. Darwin, M. T. Turvey, and R. G. Crowder. An auditory analogue of the sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology*, 3(2):255–267, April 1972.

[6] T. C. et al. Exploring children's verbal and acoustic synchrony: Towards promoting engagement in speech-controlled robot-companion games. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 21–24, Seattle, November 2015.

[7] T. C. et al. An acoustic analysis of child-child and child-robot interactions for understanding engagement during speech-controlled computergames. In *Proceedings of INTERSPEECH. 2016*, pages –, San Francisco, September 2016.

[8] F. Eyben, M. Wöllmer, and B. Schuller. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *ACM International conference on Multimedia (MM 2010)*, pages 1459–1462, Florence, Italy, October 2010.

[9] R. Feldman. Parent–infant synchrony biological foundations and developmental outcomes. *Current directions in psychological science*, 16(6):340–345, 2007.

[10] A. for blind review. pages –, September 2016.

[11] G. Hoffman and K. Vanunu. Effects of robotic companionship on music enjoyment and agent perception. In *The 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–324. ACM/IEEE, 2013.

[12] T. Iqbal, M. J. Gonzales, and L. D. Riek. Joint action perception to enable fluent human-robot teamwork. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, pages 400–406, Japan, September 2015.

[13] A. Jakkam and C. Busso. A multimodal analysis of synchrony during dyadic interaction using a metric based on sequential pattern mining. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, pages 6085–6089, Shanghai, China, March 2016.

[14] J. F. Lehman and S. Al Moubayed. Mole madness–a multi-child, fast-paced, speech-controlled game. In *Proceedings of AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction. Stanford, CA*, 2015.

[15] C. D. Looze, C. Oertel, S. Rauzy, and N. Campbell. Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *International Conference on Phonetic Sciences (ICPhS). Hong Kong*, pages 1294–1297, 2011.

[16] S. Mariooryad and C. Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2329–2340, October 2012.

[17] A. N. Meltzoff and W. Prinz. *The imitative mind: Development, evolution and brain bases*, volume 6. Cambridge University Press, 2002.

[18] M. P. Michalowski, S. Sabanovic, and H. Kozima. A dancing robot for rhythmic social interaction. In *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*, pages 89–96, Washington DC, USA, March 2007.

[19] M. P. Michalowski, R. Simmons, and H. Kozima. Rhythmic attention in child-robot dance play. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 816–821, 2009.

[20] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California Berkely, Fall 2002.

[21] J. Piaget. *The moral judgement of the child*. Simon and Schuster, 1997.

[22] K. Prepin and P. Gaussier. How an agent can detect and use synchrony parameter of its own interaction with a human? In *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 50–65. 2010.

[23] N. Sadoughi and C. Busso. Retrieving target gestures toward speech driven animation with meaningful behaviors. In *International conference on Multimodal interaction (ICMI 2015)*, pages 115–122, Seattle, WA, USA, November 2015.

[24] N. Sadoughi, Y. Liu, and C. Busso. Speech-driven animation constrained by appropriate discourse functions. In *International conference on multimodal interaction (ICMI 2014)*, pages 148–155, Istanbul, Turkey, November 2014.

[25] N. Suzuki and Y. Katagiri. Prosodic alignment in human-computer interaction. *Connection Science*, 19(2):131–141, June 2007.