Backchannel Opportunity Prediction for Social Robot Listeners

Hae Won Park¹, Mirko Gelsomini¹², Jin Joo Lee¹, Tonghui Zhu¹, and Cynthia Breazeal¹

Abstract—This paper investigates how a robot that can produce contingent listener response, i.e., backchannel, can deeply engage children as a storyteller. We propose a backchannel opportunity prediction (BOP) model trained from a dataset of children’s dyad storytelling and listening activities. Using this dataset, we gain better understanding of what speaker cues children can decode to find backchannel timing, and what type of nonverbal behaviors they produce to indicate engagement status as a listener. Applying our BOP model, we conducted two studies, within- and between-subjects, using our social robot platform, Tega. Behavioral and self-reported analyses from the two studies consistently suggest that children are more engaged with a contingent backchanneling robot listener. Children perceived the contingent robot as more attentive and more interested in their story compared to a non-contingent robot. We find that children significantly gaze more at the contingent robot while storytelling and speak more with higher energy to a contingent robot.

I. INTRODUCTION

Early language ability (such as vocabulary skills and oral language knowledge during preschool) is one important predictor of children’s academic success throughout their school years [1], [2]. Extensive research in young children and infants has verified the importance of social cues like backchannel (listener response), joint attention, and shared gaze for language acquisition, thereby emphasizing the importance of a cooperative effort of the teacher and the learner [3]. The goal of our project is in developing a robot peer that can engage children in personalized storytelling experience. Before the robot can generate a personalized story content and tell it to the child, it needs to assess the child’s lexical and syntactic levels and storytelling skills from his/her storytelling. In the process, the robot should appear as an attentive listener, providing adequate backchannel feedback to keep children engaged as a storyteller.

Social robot learning companions offer unique opportunities of guided, personalized, and controlled social interaction and delivery of a desired curriculum. In contrast to other devices such as computers, tablets, and smartphones, robots can play, learn, and engage with children in the real world — physically, socially and emotionally. The development of personalized robot tutors has gained increased attention [4], [5], yet the application to long-term interaction with children, as well as the construction of a fully autonomous, cognitive, expressive, and responsive social companion has not been achieved. In order to serve as an effective long-term companion, social robots need to create models of the cognitive capacities and language and response behaviors of the child learners in order to provide autonomous, adaptive, and personalized interaction.

Here, we present our approach toward developing a rule-based backchannel opportunity prediction model, as a first step to engaging children as a storyteller. Our model was trained and tested on K-2 children’s dyad storytelling dataset collected from local Boston public schools. We then conducted a user study in which children told stories to two robots, one providing contingent and the other providing non-contingent backchannel feedback, to answer the question of whether an attentive robot listener utilizing our rule-based model can positively affect a child storyteller’s behavior (Fig. 1).

In the following section, we review prior works in robot tutor systems, human backchannel behavior research, and approaches to computationally modeling backchannel feedback. In Section III, an overview of the system including our robot platform and audio feature extraction modules are presented. Section IV then provides approaches to training backchannel opportunity rules from a dataset of children’s storytelling activity, as well as performance evaluation results. After describing the triad (child and two robots) storytelling and listening experimental protocol in Section V, we present our analysis methods and results with discussions in Section VI. In summary, children directed their storytelling significantly more to the contingent robot and perceived the...
contingent robot as more attentive and interested in their story.

II. RELATED WORK

An interaction with social robots has been shown to have a positive effect on learning and behavioral outcomes for both adults and children beyond mere “novelty effects” [6], [7]. Children have learned vocabulary from a tele-operated storytelling robot [8] and fostered curiosity-relevant behaviors [9]. Furthermore, studies have shown that a robot’s socially contingent behavior help mitigate awkwardness during silences and can improve conveying sincerity of the robot in maintaining conversations toward the user [10], and affects children’s preferred choice of informant, almost on-par with human collaborators [11].

Backchanneling (BC) is a component of conversation and verbalization that is naturally embedded in our role as a listener in a conversation. Throughout a conversation, the listener may nod their head periodically or use short utterances, such as yeah, ok, uh huh, mhmm, to show that they are engaged. In [12], authors suggest that BC serves four cognitive functions including indicating understanding, or lack thereof, repair or clarification of the message, and sentence completion. From this perspective, BC’s main function is in establishing common ground by signaling that the receiver has understood the message [13]. Since listener BC are generated rapidly and seem elicited by a variety of speaker verbal and nonverbal cues, detecting BC timings is a difficult problem. There is evidence that people can generate such feedback without necessarily attending to the content of speech [14], and this has motivated diverse approaches that detect BC opportunity using prosodic and prosodic features from a speaker’s voice [15], [16], [14]. In [17], the authors proposed a robot listener that detects backchannel timing using fundamental frequency (F0) and energy. However, this system’s evaluation is limited in that it relied solely on users’ subjective Likert-scale measure of system ratings. Rather than detecting BC opportunity from a speaker’s various prosodic cues during speech, [18] uses robot’s backchannel feedback to signal acknowledgement (or lack thereof) as part of a turn-taking policy. Here, we propose to predict backchannel opportunities by using a broad range of real-time prosodic features, such as fundamental frequency, voicing probability, energy, and pitch. We then integrate the algorithm to our social robot platform, Tega, and generate real-time listener BC responses while listening to a child’s storytelling. In addition to users’ subjective measures, we engage the users with two robots demonstrating contingent and non-contingent listeners and report the users’ behavioral differences toward the robots, such as gaze pattern and affect expressions.

III. SYSTEM DESCRIPTION

Tega is a social robot platform developed to support long-term deployment in homes and schools (Fig. 2). An Android smartphone mounted on the head graphically displays robot’s facial expressions and handles computational tasks such as sensor processing, data collection, and motor control and wireless communications. Tega is designed to animate squash-and-stretch-based motions with 4 rotational joints (Rx, Ry, Rz, Rxp) and one translational joint (Ty). Twenty facial expressions and gesture animations were developed for this study, each depicting the most frequently observed BC responses used by children, i.e., variations of gaze, lean forward, brow raise, smile, nod, and short utterances (discussed in Section IV-A).

We created Robot Operating System (ROS) modules to automatically subscribe to sensor topics and publish BC opportunities (Fig. 3). The topics are categorized into the following modules: sensors, prosodic feature extraction, facial affect feature extraction, BC opportunity prediction (BOP), and robot-action control. The sensor module publishes audio and vision data from the microphone and camera. The prosodic feature extraction module then processes the audio data and publishes fundamental frequency (F0), energy (voice loudness in log scale), and pitch direction (fall: -1, flat: 0, rise: 1) using OpenSmile [19]. The BC opportunity prediction (BOP) module subscribes to these prosodic feature topics and outputs backchannel opportunities using our proposed algorithm in Section IV. The robot-action control module subscribes to both the BOP and Tega state topics and makes decisions whether to backchannel or not given an opportunity depending on the current Tega states and previous BC decisions. The affect feature extraction module subscribes to the camera topic and extracts facial landmarks to detect head orientation, facial expressions, and affect states of the user.

IV. APPROACH

A. Children’s Backchanneling Behavior

BC behavior is one of the last communication skills acquired in a person’s developmental stage, and the frequency of backchannel responses increase significantly with age [20]. We collected a dataset of our target age (4–6) children telling stories to one another to analyze what types of BC response children produce to which speaker cues. Eighteen children from a single kindergarten (K2) classroom participated in the data collection (M = 5.22 years-old, SD = 0.44, 39% female), and in total, 58 individual storytelling episodes were collected.
Fig. 3. ROS computation graph depicts communication between the nodes and topics. The backchannel opportunity prediction (BOP) module detects response timing from low-level prosodic features extracted from the speaker’s voice. The speaker’s affect features are extracted from facial image frames. Finally, backchannel and facial information are combined to generate robot listener’s response behavior. The rosbridge websocket gathers and visualizes core data to the experimenter.

Fig. 4. Nonverbal listener responses and speaker cues were annotated and analyzed from a dataset consisting of 58 episodes of K-2 students telling and listening to a story. This dataset served as a training set for our backchannel opportunity prediction model as well as a guide to implementing Tega’s BC behaviors.

For each storytelling episode, the nonverbal behaviors of both the listener and storyteller were manually coded using a video-annotation software by four coders (Fig. 4). Three additional coders were recruited to simulate themselves being a listener and mark the moments when they wanted to BC. After this simulation, coders reviewed the audio snippets surrounding these moments to further categorize the type of speaker cues perceived (pitch, energy, pause, filled pause, long utterance, clause ending, other).

From the analysis of this dataset, the most observed backchannel responses were gaze at partner, lean toward, brow raise, smile, nod, and short utterances. Our robot listener implements these nonverbal behaviors to signal its engagement to the child speaker. Speaker cues that the child listeners most frequently responded to were gaze at partner, pitch change, long pause, energy change, filled pause (such as umm, uh, .), and long utterance. Using this analysis, we introduce prosodic feature based speaker-cue detection algorithm to predict backchannel opportunities. Detailed statistics for child nonverbal listener response and speaker cues are published in [21].

The coders’ labels were compared to create three levels of consensus dataset. Within a time frame (1000ms), if all three coders tagged a same label, then that label in that time frame is assigned level 3, and so on. In the following, we compare the results of our speaking binary classifier and BC opportunity prediction algorithm to the labels in the consensus dataset for performance evaluations.

B. Speaking Binary Classifier

Accurately knowing when there is a speaking event is key to any prosodic cue based analyses. Our speaking binary (SB) classifier improves OpenSmile’s voice activity detector (VAD) by implementing a low-pass filter and an energy-based comparison that continuously iterated on ground-truth labels. The OpenSmile’s VAD, based on line-spectral-frequencies, mel spectra, and energy, computes fuzzy scores related to the deviation from the observed long-term mean values [22]. The VAD outputs 0 when no one is speaking and 1 when someone is speaking. First, we evaluated OpenSmile VAD’s performance compared to consensus label: Precision = 92.5%; Recall = 75.5%; F-score = 82.8%.

To improve the recall rate, we applied a low-pass filter, such that

$$VAD(t) = \alpha \cdot VAD(t - 1) + (1 - \alpha) \cdot VAD(t).$$

Then we use a cut-off value $\beta$ to determine $SB(t)$, such that

$$SB(t) = 1 \text{ if } (VAD(t) > \beta), \text{ or } 0 \text{ otherwise.}$$

Next, voice signals are differentiated from stationary and low-noise signals by using an energy-variance check, an incremental noise reduction method that filters out background noise.

The speaking binary classifier’s precision and recall saturated at 96.8% and 87.5% using the following values acquired through iteration: $\alpha = 0.8$ and $\beta = 0.9$.

C. BC Opportunity Prediction Models

In this section, we present four rule-based BC opportunity prediction (BOP) models based on the most frequently observed child speaker prosodic cues (pitch, energy, long pause, and long utterance) and their patterns that prompt listener response. In the following, the values for the model parameters were iterated at steps of 100ms to minimize least square errors based on human coder labels.

Long Utterance Model (Fig. 5): An inter pausal unit (IPU) is a maximal sequence of words surrounded by silence
(SIL) longer than 50 ms. A sequence of consecutive IPU and SIL longer than W_SPEAK followed by W_PAUSE is a backchannel opportunity. Similar observation was made in [23].

P1 a pause of W_PAUSE (800ms) length,
P2 preceded by at least W_SPEAK (1.5s) of speech,
P3 provided that no BC has been output within the preceding BC_RATE (1.3s)

**Long Pause Model** (Fig. 6): This opportunity is triggered when a long pause, LP_PAUSE, is detected after a speech LP_SPEAK.

P1 a pause of LP_PAUSE (1.7s) length,
P2 preceded by at least LP_SPEAK (1.0s) of speech,
P3 provided that no BC has been output within the preceding BC_RATE (1.3s)

**Pitch Model** (Fig. 7): This model detects rising and falling change in a pitch (P&P_SLOPE) after a speech P&P_LENGTH followed by a pause (P&P_PAUSE).

P1 a pause of P&P_PAUSE (400ms),
P2 preceded by at least P&P_SPEAK (1.0s) of speech,
P3 where the last P&P_LENGTH (300ms),
P4 contain a rising/falling pitch of at least P&P_SLOPE (25%) rise/drop.
P5 provided that no BC has been output within the preceding BC_RATE (1.3s).

**Energy Model** (Fig. 8): This model detects rising and falling change in energy (E_SLOPE) after a speech E_SLOPE_LENGTH followed by a pause (E_PAUSE).

P1 a pause of E_PAUSE (300ms),
P2 preceded by at least E_SLOPE_LENGTH (500ms) of speech,
P3 contain a rising/falling energy of at least E_SLOPE (30%) rise/drop.
P4 provided that no BC has been output within the preceding BC_RATE (1.3s).

**D. BOP Evaluation**

The four BOP models were tested with 28% of the level-3 consensus dataset that were not used for training. We prioritize avoiding inappropriate BC response generation (false-positives) over missing BC opportunities (false-negatives). Precision measure emphasizes false-positives while recall emphasizes false-negatives, and therefore we measured the performance using precision while comparing cue labels and BC timestamps.

The evaluation of the test set returned the following precisions:

- Long Utterance: 89.5 ± 7.4%,
- Pitch: 61.1 ± 13.8%,
- Long Pause: 78.3 ± 8.3%,
- Energy: 67.3 ± 13.2%.

**V. EXPERIMENTAL SETUP**

We hypothesized that a social robot providing contingent BC feedback would be perceived as more attentive which in turn will encourage children to attend to it more while telling a story compared to a non-contingent robot. To evaluate our hypothesis, we designed two studies. The first study was a within-subjects study in which children told stories to two contingent and non-contingent BC Tega robots at the same time. In the between-subjects study, children were divided into two groups and only told stories to either a contingent or a non-contingent robot.
two identically looking Tega robots each demonstrating contingent and non-contingent BC behavior. The second study was a between-subjects study in which participants were divided into two groups, and each child told stories to either a contingent of a non-contingent robot. The experimental room was setup as depicted in Fig. 9.

The behavior of the contingent robot closely mirrored that of a child listener’s based on our previous analyses (Section IV-A). It used our BOP models to predict backchannel timing from the participant’s prosodic cues and produced one of the backchannel response animations. The non-contingent robot also used the same animations, but was played at random every 5.53 ± 1.5 seconds.

To prevent any bias in the study results, the appearance and behavior of the robots were carefully controlled. The robots looked identical, used the same name, and the number of BC animations played was matched between conditions. In the within-subjects study, the placement of the robots was counter-balanced: left (45%) and right (55%).

A. Participants

In the within-subjects study, 23 children (age \( M = 6.13, SD = 1.36; 43.5\% \text{ female} \)) between the age of 4–8 years old were recruited through a local parents’ mailing list. Different children were recruited for the between-subjects study, in which 54 children (age \( M = 5.92, SD = 1.40; 50\% \text{ female} \)) between the age of 4–8 were recruited through a local parents’ mailing list and referrals. Children were assigned to either contingent or non-contingent condition. The two groups were balanced in terms of age and gender (contingent: age \( M = 5.96, SD = 1.32, \text{female:male}=15:12 \); non-contingent: age \( M = 5.89, SD = 1.53, \text{female:male}=12:15 \)).

B. Protocol

- **Pre-activity:** In the within-subjects study, the child, parent(s), and the experimenter brainstormed about a story the child likes to tell. In the between-subjects study, all children watched a story on a tablet and were asked to retell that story to the robot.
- **Storytelling activity:** In both studies, the experimenter invited the participant to the study room and introduced the robot(s). When the participant told the robot(s) that he/she is here to tell them stories, the Tega robot(s) woke up and started backchanneling to the child’s speech. When the child indicated an end of a story, the robot(s) went back to sleep, and the experimenter conducted a post survey.
- **Post survey:** Using a 5-point smileyometer Likert scale, likeability of the robots (how much did you like Tega?) and enjoyability of the storytelling task (how much did you like telling stories to Tega?) were first asked in both studies. Afterwards in the within-subjects study, children were asked about the level of interest each robot showed toward their story (how much do you think this Tega enjoyed your story?). Participants were then asked to give a sticker to the robot who they thought was a better listener was was more interested in their story. In the between-subjects study, perceived animacy (Do you think Tega is alive? How much?), perceived intelligence (Do you think Tega understood your story? How well did it understand?), perceived emotional support (Does Tega care about you? How much?), and the participant’s self comprehension level of Tega’s facial expressions (How well did you understand Tega facial expressions?) and backchannel behaviors (When you told a story to Tega, how well did you understand what Tega was doing?) were asked. For each question, the experimenter further asked and documented the reason of children’s answers.

C. Measurements

During children’s storytelling, we collected the total length of the interaction, \( F_0 \), speaking binary, energy, and pitch from an audio data stream, and head orientation (pitch, yaw, roll) and facial affect features from the camera. We used Affdex, a commercially available automated affect recognition software [24]. As depicted in Fig. 10, Affdex extracts 15 facial expressions that are used as predictors to calculate the likelihood of emotions or to estimate a point in a continuous space defined by valence (a degree of positive and negative emotion) and expressiveness (intensity of an expression). The continuous value ranges are as follows: 15 facial expressions ([0,100]; attention, brow furrow, brow raise, chin raise, eye closure, inner brow raise, lip corner depressor, lip press, lip puckter, lip suck, mouth open, nose wrinkle, smile, smirk, upper lip raise), and underlying 7 emotions ([0,100]; anger, contempt, disgust, fear, joy, sadness, surprise), valence ([−100,100]), and expressiveness ([0,100]). We also recorded the total number and frequency of the robots’ BC feedbacks. All data was time synchronized.

VI. RESULTS AND DISCUSSION

Among 23 participants, we were able to analyze data from 20 children (age \( M = 6.25, SD = 1.33; 45\% \text{ female} \)). One 4 yr-old did not want to tell a story and withdrew from the study. We excluded two participants’ data because the frontal view camera was out of focus and we couldn’t extract facial features from the videos. The average length of children’s storytelling was 10.77±4.12 minutes. We found no statistical significance in the number of backchannel feedback provided and the intensity of backchannel motions (categorized as...
small or large) between the contingent and non-contingent robots, thereby we can safely assume that the expressivity of both robots was similar.

In the following, we report our major findings as subsections. We first analyzed the gaze pattern of the child in correlation to the speaking binary. Then we evaluated the child’s affective reaction to each robot condition, and lastly we summarized the post-survey result.

A. Gaze and Speech

In the within-subjects study, we analyzed children’s gaze pattern using the yaw and pitch information of the head orientation and correlated this data with speaking binary. In the between-subjects study, we analyzed the amount of storytelling children in each group did to the robot normalized to the overall duration of the session.

In the within-subjects study, the overall gaze direction didn’t show significant difference toward the contingent and non-contingent robots (contingent: \(M = 0.359, SD = 0.070\), non-contingent: \(M = 0.396, SD = 0.076\); \(t(38) = 1.598, p = 0.118\)). While telling a story (speaking binary (SB) = 1), however, children significantly gazed more at the contingent robot (contingent: \(M = 0.185, SD = 0.076\), non-contingent: \(M = 0.146, SD = 0.040\); \(t(38) = 2.031, p = 0.049\)). When children were silent (SB=0), they were significantly gazing more at the non-contingent robot (contingent: \(M = 0.174, SD = 0.031\), non-contingent: \(M = 0.250, SD = 0.053\); \(t(38) = 5.523, p < 0.01\)). In the between-subjects study, the overall duration of the sessions were not much different between the two conditions, but children in the contingent group spoke significantly more than children in the non-contingent group (contingent: \(M = 0.474, SD = 0.118\), non-contingent: \(M = 0.405, SD = 0.102\); \(t(49) = 2.232, p = 0.031\)). An inspection of the videos suggests that the non-contingent robot’s off-timing feedback interrupts the child’s speech causing a short to long silent gaze and affect reaction in both studies (Fig. 11).

These results suggest that a contingent backchanneling robot better engaged children as a storyteller, validating our main hypothesis.

B. Facial and Prosodic Affect Cues

In the within-subjects study, an analysis of facial expressiveness (continuous scale of [0,100]) showed that children were more expressive toward the non-contingent robot (contingent: \(M = 56.42, SD = 19.23\), non-contingent: \(M = 76.34, SD = 24.35\); \(t(38) = 2.871, p < 0.01\)). Children expressed emotions with higher valence and laughed more toward the non-contingent robot, which children described the robot as "funny", "made me laugh", and "shy". Analysis revealed high correlation between affect expressiveness and pause from storytelling (SB=0) (SB=0: \(M = 67.83, SD = 19.21\), SB=1: \(M = 54.25, SD = 12.38\); \(t(38) = 2.658, p = 0.012\)), consistently suggesting that children paused from storytelling and reacted to the non-contingent robot when it made off-timing responses. In the between-subjects, we observed mean prosodic energy difference between conditions. Children who told stories to the contingent robot had higher log energy in their speech (contingent: \(M = 0.059, SD = 0.031\), non-contingent: \(M = 0.042, SD = 0.023\); \(t(47) = 2.14, p = 0.037\)).

The contingent robot not only encouraged children to speak more, it also prompted children to speak with more energy which often correlates to excitement. On the other hand, the non-contingent robot’s behavior caused distraction which prompted high valence and expressiveness reaction in children’s facial expressions but suppressed storytelling (Fig. 12).

C. Post survey

After children finished telling stories, all robots went back to sleep, and the experimenter conducted a post survey. In regards to the overall interaction experience, a five-point Likert scale revealed high perceived likeability toward Tegas (\(M = 4.70, SD = 0.66\)) and enjoyability of telling a story to Tegas (\(M = 4.50, SD = 0.69\)).

In the within-subject study, when asked about the perspective of the robots, most children answered both Tegas enjoyed their story, and no difference was observed between the two robots (contingent: \(M = 4.63, SD = 0.60\), non-contingent: \(M = 4.53, SD = 0.61\)). Fischer’s exact test revealed that there was no statistical significance between
which side the contingent robot was placed versus the robot child indicated as a better listener. In the sticker test, 15 out of 20 children responded that the contingent robot was more attentive than the non-contingent robot (75%).

In the between-subjects study, children’s self-reported level of Tega’s backchanneling response comprehension was significantly correlated to their enjoyment and perception of Tega across all children. More specifically, children who reported higher level of comprehension of Tega’s backchanneling response also said they enjoyed telling stories to Tega more \((r = 0.519, p = 0.0001)\) and perceived Tega as more intelligent \((r=0.297, p=0.027)\) and more caring \((r = 0.286, p = 0.044)\).

VII. CONCLUSIONS

We presented a method to developing a robot listener that autonomously and actively responds to child speaker’s prosodic cues. The two within- and between-subjects studies consistently revealed that children are more engaged with a contingent backchanneling robot while telling a story and perceived it as more attentive and interested in their stories. The approach and findings of this work will enable our social robot to deeper engage in exchanged storytelling activities with children while providing personalized early literacy education in the near future. A long-term deployment of these robots will increase our understanding of the impact of interactions with social robot companions on children’s language development. This could inspire new tools and practices for early pre-literacy and language education (as well as other domains such as STEM) in the home, classroom, and beyond.

REFERENCES


