Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling [*]

Hae Won Park, Mirko Gelsomini[†], Jin Joo Lee, and Cynthia Breazeal
Personal Robots Group, MIT Media Lab
20 Ames Street, E15-468, Cambridge, MA 02139 USA
{haewon, gelso, jinjoo, cynthiab}@media.mit.edu

ABSTRACT

While there has been a growing body of work in child-robot interaction, we still have very little knowledge regarding young children's speaking and listening dynamics and how a robot companion should decode these behaviors and encode its own in a way children can understand. In developing a backchannel prediction model based on observed nonverbal behaviors of 4–6 year-old children, we investigate the effects of an attentive listening robot on a child's storytelling. We provide an extensive analysis of young children's nonverbal behavior with respect to how they encode and decode listener responses and speaker cues. Through a collected video corpus of peer-to-peer storytelling interactions, we identify attention-related listener behaviors as well as speaker cues that prompt opportunities for listener backchannels. Based on our findings, we developed a backchannel opportunity prediction (BOP) model that detects four main speaker cue events based on prosodic features in a child's speech. This rule-based model is capable of accurately predicting backchanneling opportunities in our corpora. We further evaluate this model in a human-subjects experiment where children told stories to an audience of two robots, each with a different backchanneling strategy. We find that our BOP model produces contingent backchannel responses that conveys an increased perception of an attentive listener, and children prefer telling stories to the BOP model robot.

1. INTRODUCTION

Social robots have great potential to support children's education as peer learning companions. As a physically embodied technology, social robots can leverage our social means of communication (e.g., speech, gestures, gaze, and facial expressions) to engage with us in more natural, intuitive, and interpersonal ways. They combine the general benefits of technology such as scalability, customization, easy content deployment with student-paced, adaptive software. Prior research has shown that young children will not only treat social robots as companions and guides [15, 29], but will also readily learn new information from them [19, 31] where a robot's perceived credibility as an educator is impacted by how socially contingent it behaves [3]. Beyond academic learning, peer-modeled robot companions have demonstrated to be social influencers that can positively impact a child's attitude by fostering curiosity or a growth mindset [24, 10] and teach positive interpersonal skills [23, 22]. In modeling children more holistically, researchers have found that incorporating the emotional experience of learners can personalize their engagement during educational activities [11] as well as lead to more accurate assessments of student performance [30].

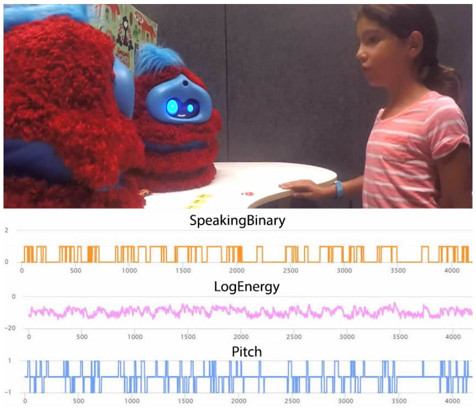


Figure 1: We investigate the effects of an attentive listening robot on a child's storytelling behavior. The robot detects nonverbal prosodic cues from the storyteller's speech to provide backchannel feedback to signal engagement.

Given this potential, we aim to study how a peer-like social robot can successfully foster the development of early language skills of preschoolers and kindergarteners. Storytelling and other forms of conversational activity are key to children's language development and are a mutually regulated activity between speaker and listener. In this paper, we specifically focus on one of the key dynamics in storytelling activities – the back-and-forth process of speaker cues and listener backchannels (BC) that encourages both parties to

[*] A teaser video of the research can be viewed from the following URL: `https://goo.gl/HMEfnt`

[†] Mirko Gelsomini was also affiliated with Politecnico di Milano, Italy at the time this research was conducted.

HRI '17, March 06-09, 2017, Vienna, Austria
© 2017 ACM. ISBN 978-1-4503-4336-7/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2909824.3020245

stay engaged in the activity and enables long-term interaction. Speakers elicit feedback from listeners through subtle nonverbal cues, and listeners respond nonverbally to communicate that the listener is paying attention and following along. We hypothesize that a social robot that contingently backchannels will be perceived as an attentive listener by child storytellers and will affect their storytelling behavior (Figure 1).

In this work, we offer the following contributions. First, we provide quantitative analyses of kindergarten children's (ages between 4–6) backchanneling response behavior. While adults' BC behavior has been widely studied including cultural diversities [37], it is seldom studied how children develop and produce (encode) speaker cues as a storyteller to request feedback from the listener and how they understand (decode) those signals as a listener. Here, we specifically present analyses on young children's nonverbal behavior in a storytelling context. Namely, we 1) identify backchanneling behaviors that indicate the engagement state of the listener, 2) identify speaker cues that child listeners acknowledge and respond to, and 3) characterize the bidirectionality of nonverbal behaviors (i.e., can children decode what they encode?) (Section 3). Secondly, we 1) define a robot listener's action space based on the identified child listener's backchanneling behaviors and 2) design a backchanneling opportunity prediction (BOP) model around the identified child speaker cues (Section 4). To the best of our knowledge, we pioneer the development of a computational backchanneling model based on children voices and behaviors. Finally, we evaluate the proposed BOP model on a social robot listener to respond appropriately and reciprocally to a child as he/she tells a story. We find that children do perceive the contingent backchanneling robot as an attentive listener and stay engaged by directing their storytelling to the contingent robot versus the non-contingent robot. (Section 6 & 7).

2. RELATED WORK

2.1 Backchannel (BC) behavior

In this work, we focus on the use of nonverbal listener response behaviors, typically referred to as backchannels (BC), including gestures such as gaze locking and nodding, and non-lexical utterances such as *yeah, ok, uh huh, mhmm.* Backchanneling is a listener response that serves cognitive functions indicating the state of engagement, understanding (or lack thereof), repair or clarification of the message, and sentence completion [7, 6]. There is surprisingly limited amount of work that reports on young children's BC behavior and the effects of BC on their language development. In [25], the positive effects of BC behaviors on children's language learning is reported. Twenty preschoolers (mean age 3.7) were randomly assigned to either an intervention or control group. During an year-long intervention, parents of the intervention group were asked to encourage their child's narratives through providing backchanneling responses, which resulted in longer narratives and improved vocabulary learning immediately after the session terminated. Children in the intervention group showed significant improvements in overall narrative skills. In particular, they produced more context-setting descriptions about where and especially when the described events took place. In [28], it is also reported that 4-year olds' narratives are influenced by adult listeners' behaviors. According to the authors' observations, children try to create more complex narratives when the adult listener appears more attentive. Also when the adult listener is more attentive, children tend to try harder to re-attract their attention when they are momentarily distracted.

While the above works provide some insights into how young children's narratives can be improved by adult listener's BC responses, it lacks important details of children's encoding (how children develop and produce speaker cues to request feedback from the listener) and decoding (how children understand and detect speaker cues as a listener) abilities. Learning what speaker cues and listener responses children in preschools and kindergartens can perceive and produce is necessary when developing an appropriate BC model. As such, we conducted a data collection of peer-to-peer storytelling interactions between kindergarten children and identified key speaker cues and prosodic features that prompt listener backchannels. This analysis is provided in Section /refsec:cb.

2.2 Computational models of BC

Backchannels are elicited by a variety of speaker verbal and nonverbal cues, and its contingent timing makes for a significant technical challenge. Without having to necessarily attend to the content of speech, approaches have been successful in detecting prosodic features (e.g. voice activity detection (VAD), energy, and pitch) in realtime to determine appropriate backchannel opportunities [18, 20, 21].

Ward and Tsukahara [34] suggest that an important prosodic cue is a region of low pitch located toward the end of an utterance. Their model was produced manually through an analysis of English and Japanese conversational data. Truong et al. [26] consider features from the speaker's speech along with gaze cues to determine the placement of BCs. They found that the number, timing, and type of BC has a significant effect on how human-like BC behavior is perceived. Morency et al. [18] present a realtime multimodal BC prediction system. They automatically select features from speech and gaze and train conditional random fields to model sequential probabilities. Other studies have used decision trees [20] based on pitch and power features, and others have used hidden Markov models [21] where state transitions correspond to changes in prosodic context.

All these prior approaches have design their computational models based on observed adult behaviors and trained on the adult voices. As a first work that attempts to develop a BC model specifically targeted at 4–6 year-old children, we propose a model comparable to Ward and Tsukahara's [34] and Truong et al.'s [26] with children storytelling and listening datasets.

3. NONVERBAL BEHAVIORS OF CHILDREN

3.1 Data Collection

Eighteen participants of typical development were recruited from a single kindergarten (K2) classroom in a local public elementary school. The average age was 5.22 years-old (SD=0.44) with a 61:39 male:female ratio. Each child participated in at least three rounds of storytelling with different partners and storybooks over a span of five weeks (58 episodes in total). In each session, a pair of students take turns narrating their story to the other; each turn generating a storytelling episode. Three time-synchronized cameras

					Linear Regression Model		
Behavior	N	Mean Freq	Mean Dur	%Pop	Main Effect	Freq Term	Dur Term
Gaze Partner	284	4.95	2.02	100	F(2, 223) = 23.66, p* = 4.80e^{-10}	B = 2.19, p* = 1.55e^{-06}	B = 1.48, p = 0.02
Lean Toward	124	2.14	7.97	100	F(2, 204) = 12.98, p* = 4.92e^{-06}	B = 0.91, p* = 8.55 e^{-05}	B = 0.29, p = 0.17
Brow Raise	103	1.78	2.32	100	F(2, 170) = 5.04, p* = 7.49e^{-03}	B = 0.70, p* = 0.02	B = -1.75, p* = 4.28e^{-03}
Smile	189	3.26	6.55	94	F(2, 202) = 7.14, p* = 1.01e^{-03}	B = 0.29, p = 0.31	B = 0.71, p* = 5.59e^{-03}
Nod	18	0.31	1.13	39	F(2, 43) = 2.24, p = 0.11	B = 1.37, p = 0.28	B = -0.40, p = 0.97
Utterance	18	0.31	0.88	50	F(2, 51) = 3.35, p* = 0.04	B = 1.33, p = 0.12	B = 1.51, p = 0.86

Table 1: Descriptive Statistics and the Linear Regression Model for Listener Responses. *N* is the total occurrences found in the dataset. The Mean Frequency is the average number of occurrences in a storytelling episode (i.e., N/58). The Mean Duration is the average duration of the emitted behavior in seconds. %Pop refers to the proportion of the population (18 participants) that demonstrated a single instance of the behavior across the repeated interactions. The linear regression model predicts a child's level of listening (coded as -1 or 1) based on the normalized duration and frequency rate of the observed behavior.

captured the frontal-view of each participant along with a bird's eye view.

For each storytelling episode, the nonverbal behaviors of both the listener and storyteller as well as the listeners' attentive state were manually coded using video-annotation software ELAN[1] with a custom developed template. Four coders marked the onset and offset times for the occurring nonverbal behaviors and achieved moderate levels of agreement (Fleiss' $\kappa = 0.55$). Three additional coders were recruited to simulate themselves being a listener and to mark the moments when they wanted to BC to the audio recording of the child storyteller. 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, a long contiguous utterance (or wordy), and other. We followed the Parasocial Consensus approach from [14] to build consensus of when backchannel opportunities occurred. More specifically, each of our three coders' registered backchannel times were added as a 'vote' on a consensus timeline with a duration of one second around the central moment. An area in the timeline with more than two total votes was counted as a valid backchannel moment.

3.2 Analysis of Listener Behavior

To identify attention-related nonverbal behaviors, we examined the ability of the frequency and duration of behaviors to predict whether a child is listening or not. Based on the annotations of the listener's mental state, the storytelling episodes were split into segments where the child was listening and not listening. For each nonverbal behavior, a linear regression analysis was performed to predict a child's level of listening based on the normalized duration and frequency rate of the behavior observed in each segment. As shown in Table 1, partner gazes, leaning toward, smiles, and utterances significantly predict a listener's mental state. Interestingly, the frequency of brow raises holds a significant positive relationship, while its duration holds a significant negative relationship. Nods have a positive relationship to the child listening, but their rare occurrences in this population make it difficult to evaluate as significant.

3.3 Analysis of Speaker Cues

To identify the speaker cues that child-listeners acknowledge and respond to, we first examined the ability of speaker cues to predict the likelihood of a positive response from the listener. From our storytelling episodes, we extract observational pairs of the type of cue generated by the speaker and whether a positive response was observed from the listener within 3 seconds. Based on our prior analysis on attention-related behaviors, the onset of a partner gaze, lean toward, smile, utterance, brow raise, and nod were considered to be positive responses to a cue. A logistic regression was performed to ascertain the effects of the individual speaker cues on the likelihood that a listener would respond. Based on the Wald Chi-square statistic, the overall logistic regression model was statistically significant, $\chi^2(6) = 40.69$, p = 3.33e^{-07}. As shown in Table 2, the speaker cues—gaze, pitch, and wordy—have the ability to elicit a positive re-

			Logistic Regression Model		
Variables	N	Mean Freq	B	tStat	Sig.
Intercept	—	—	-1.00	-5.01	*p=7.98e^{-07}
Gaze	218	3.76	1.08	5.62	*p=3.47e^{-08}
Pitch	175	3.02	0.43	2.53	*p=0.01
Pause	156	2.69	0.09	0.54	p=0.59
Energy	61	1.05	0.14	0.66	p=0.51
FilledP	37	0.64	0.23	0.81	p=0.42
Wordy	18	0.31	0.70	2.06	*p=0.04

Table 2: Descriptive Statistics and the Logistic Regression Model for Individual Speaker Cues. *N* is the total occurrences found in the dataset. The Mean Frequency is the average number of occurrences in a storytelling episode (i.e., N/58). The logistic regression model predicts the likelihood of a positive response from the listener based on the type of emitted speaker cue.

[1] https://tla.mpi.nl/tools/tla-tools/elan/

1 Cue	N	Rate	Sig.	2 Cues	N	Rate	Sig.	2+ Cues	N	Rate	Sig.
G.....	218	0.70	$p^*=2.74e^{-09}$	.CP...	80	0.55	p=0.22	GC.E..	17	0.94	$p^*=1.37e^{-04}$
.C....	175	0.57	$p^*=4.80e^{-02}$	.C.E..	50	0.66	$p^*=0.02$	GCP...	16	0.88	$p^*=2.09e^{-03}$
..P...	156	0.48	p=0.34	GC....	41	0.88	$p^*=3.92e^{-07}$	.CPE..	14	0.64	p=0.21
...E..	61	0.57	p=0.15	..P.F.	26	0.50	p=0.58	.CP..W	6	0.67	p=0.34
....F.	37	0.46	p=0.37	G.P...	21	0.86	$p^*=7.45e^{-04}$	GCPE..	5	1.00	$p^*=3.12e^{-02}$
.....W	18	0.72	$p^*=4.81e^{-02}$	..PE..	20	0.50	p=0.59				
				.C...W	12	0.75	p=0.07				
				.C..F.	10	0.20	p=0.05				

Table 3: Descriptive Statistics and the Binomial Test for Combinations of Speaker Cues. The most frequently observed cue combinations specified through the presence of the cue's symbolic letter. G:Gaze C:Pitch: P:Pause E:Energy F:Filled Pause W:Wordy. A dot represents the absence of that cue. N is the total occurrences of the cue combination found in the data set. The one-sided binomial tests whether the cue combination elicits a response rate that is greater than the expected chance rate of 0.5.

sponse from the young listeners. As expected, some of the speaker cues (energy, pause, and filled pause) taken alone did not offer significant predictive ability when examined in isolation. However, young children have been previously observed to respond more often in greater cue contexts where two or more cues are co-occurring [13].

Our next analysis examined the ability of cue combinations to predict the likelihood of a positive response from the listener. Speaker cues were considered to be co-occurring if they are within 1.3 seconds of each other (empirically determined). The aforementioned observational pairs were merged based on this criteria. The likelihood of observing a combination of cues is much smaller than individual cues, resulting in a sample size. Rather than performing a logistic regression, we use the binomial exact test to see whether the response rate of a cue combination is greater than an expected rate of 0.5. As shown in Table 3, the one-sided binomial test indicates that the response rate of the co-occurring cues Pitch-Energy, Gaze-Pitch, Gaze-Pause, Gaze-Pitch-Energy, Gaze-Pitch-Pause, and Gaze-Pitch-Pause-Energy is higher than chance.

3.4 Analysis of Mirrored Encoding-Decoding

Traditionally, nonverbal communication research has been divided into the encoding and decoding of nonverbal behaviors with little research on the differences or similarity between the two processes. With one notable exception, in the field of developmental psychology, 12-month-old infants were found to more likely succeed in producing communicative pointing gestures if they also demonstrated their comprehension of an adult's pointing intention [2]. With some evidence of the bidirectional understanding of communicative nonverbal behavior, we pose the question of whether children understand the function of cues both in the role of the communicator (the storyteller) and recipient (the listener). To support our assumption of a mirrored process when decoding and encoding nonverbal behaviors, we examine the correlation between the frequency a child produces a certain speaker cue as the storyteller and the frequency the child responds to that particular cue when in the role of a listener. Looking only at the set of cue contexts, a strong positive correlation exists between the frequency in which a child exhibits and responds to a particular cue, $r(160)^2 = 0.56$, $p = 5.26e^{-15}$. The more frequent a child expressed a speaker cue, the more frequent that child demonstrated a response to the same cue.

4. BACKCHANNEL OPPORTUNITY PREDICTION (BOP) MODEL

In the following sections, we present a rule-based method to predicting BC opportunities. We first formulate these rules by consulting children's speaker-cue analyses from Section 3.3. The four combinations of features we model are: wordy & pause, long pause, pitch & pause, and energy & pause. We used 71% of Dataset 1 (i.e., from Section 3.1) to develop these rules, and tested the BC models on the remaining 29% of Dataset 1 and another dataset collected from a different kindergarten class (Dataset 2). Our work combines prior approaches to BC modeling, in detecting cues such as wordy utterances [12], long pauses [5], pitch [35, 33, 34], changes in energy [32], and extends them to better correspondence to our corpus of children's prosodic features.

Wordy Model (Figure 2(a)): This model predicts BC opportunity based on inter-pausal units, $IPUs$. An IPU is a maximal sequence of words surrounded by a pause of duration W_PAUSE. A turn of duration W_SPEAK is defined as a maximal sequence of IPUs from a speaker, such that between any two adjacent IPUs, the silence is not greater than the value of SIL. The BC opportunity is predicted when the following conditions are met:

```
P1 a pause of W_PAUSE (800ms) length,
P2 preceded by at least W_SPEAK (1.5s) of speech,
P3 given that no BC triggered within BC_RATE (1.3s).
```

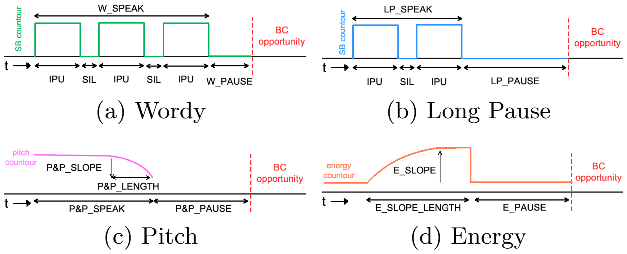
Long Pause Model (Figure 2(b)): This model predicts BC opportunities based on detecting a long pause LP_PAUSE that is preceded by a speech LP_SPEAK.

```
P1 a pause of LP_PAUSE (1.7s) length,
P2 preceded by at least LP_SPEAK (1.0s) of speech,
P3 given that no BC triggered within BC_RATE (1.3s).
```

Pitch Model (Figure 2(c)): This model predicts BC opportunities by detecting, after a certain amount of speech, a falling or rising pitch change that is followed by a 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 rise/drop (25%).
P5 given that no BC triggered within BC_RATE (1.3s).
```

Energy Model (Figure 2(d)): This model predicts BC



(a) Wordy (b) Long Pause (c) Pitch (d) Energy

Figure 2: Four rule-based models were developed to detect prosodic speaker cues that collectively make up our backchannel opportunity prediction (BOP) model. Model parameters were trained and tested against children storytelling dataset.

opportunities similar to the Pitch model but detects for energy changes.

```
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 rise/drop (30%).
P4 given that no BC triggered within BC_RATE (1.3s).
```

Model thresholds were selected by incrementing/decrementing values at steps of 100ms and trying various combinations to determine best fit using Dataset 1.

4.1 Speaking Binary Classifier

Much of the BOP model depends on accurately knowing the start and stop of speech events. A speaking binary (SB) classifier detects voicing activity from the speaker's acoustic features. We used openSMILE [8], an open-source audio-feature-extraction software[2], and applied a low-pass filter and an energy-based check to their voice activity detector (VAD). We implemented a low-pass filter that weighs the previous value of VAD, $VAD(t-1)$, higher than the current value $VAD(t)$ and applied a step function with a cut-off value of 0.9 for an SB decision. More specifically:

$$VAD(t) = 0.8 \cdot VAD(t-1) + 0.2 \cdot VAD(t), \qquad (1)$$

$$\text{then,} \quad SB(t) = \begin{cases} 1, & \text{if } VAD(t) > 0.9 \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

We reduced false-positives from the SB signal due to background noise using a simple energy-based check. During moments when SB reports no voicing activity, we sample the energy signal to establish a rolling average of the environment's noise level. So when a voice is believed to be detected, we also check to see if the current energy level is above this established minimum. We do not apply this discounting until at least 700ms into the detection process to form a reasonable baseline. Figure 3 illustrates how with each step we improve on openSMILE's VAD detector and get closer to the ground truth. Using Dataset 1, we found that our SB classifier achieves a precision of 96.8% and recall of 87.5%.

4.2 BOP Model Evaluation

We evaluated the performance of our BOP model on the remaining 29% of Dataset 1 and on a new dataset. The second dataset consists of 128 episode of 17 children (age M = 4.88, SD = 0.49, 59% female) telling a story to a robot rather than a peer [16]. We again had coders simulate being a listener to annotate BC opportunities. Using Dataset 1 and Dataset 2, we measured the BOP model's performance by comparing the cue label and BC timestamp to the coders' ground truth. We measured the performance as a function of precision which emphasizes false-positives. We prioritized avoiding inappropriate BC responses over missing a BC opportunity, which would be emphasized through false-negatives if we used a recall measure. The evaluation results are presented in Table 4.

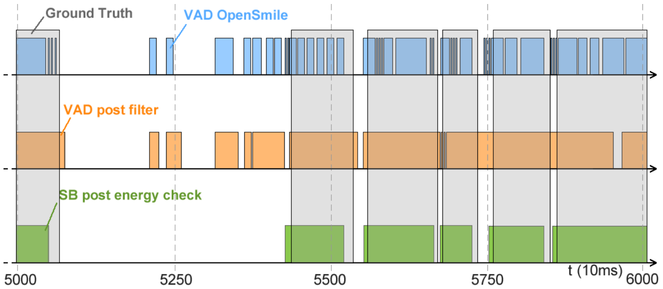


Figure 3: Identifying voice activities. The ground truth (transparent black) from human coders is compared to voice activity detection from openSMILE (VAD, blue), VAD after applying a low-pass filter (VAD post-filter, orange), and a speaking binary classifier after performing a energy-based check to filter background noise (SB, green).

Table 4: Performance of the BOP model to detect speaker cues in two different datasets. Precision is measured against the ground truth of coders.

	Dataset 1				Dataset 2
	Wordy	Long Pause	Pitch & Pause	Energy	Overall
MEAN	89.5%	78.3%	61.1%	67.3%	84.9%
STD	7.4%	8.3%	13.8%	13.2%	9.4%

5. SYSTEM DESCRIPTION

5.1 Robot Platform

Tega is an expressive social robot designed for long-term deployment in homes and schools to support children's early education [36]. An Android smart-phone mounted in the head is used to graphically display the robot's animated face as well as perform computational tasks such as sensor processing, data collection, wireless communications, decision making, and motor control. Tega expresses full-bodied animations using five degrees-of-freedom: head tilt up/down, waist tilt left/right, waist lean forward/back, body extension up/down, and body twist left/right.

5.2 System Architecture for Autonomous Interaction

For our human-subject experiment, we developed a system architecture to support a fully autonomous human-robot interaction (see Figure 4). Each module publishes its own states or computed results over the Robot Operating System (ROS) [27] network and subscribes to messages that

[2] https://github.com/naxingyu/opensmile

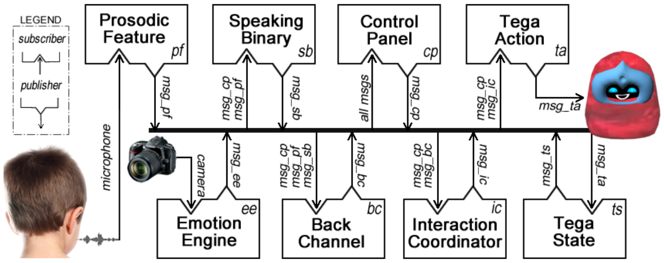


Figure 4: Realtime system architecture to support a backchanneling robot listener.

are input to their process. Using a high-quality microphone, we extracted speakers' prosodic features in realtime using openSMILE (pf) to determine speaking binaries(sb) and predict backchannel (bc) opportunities with our BOP model. These modules publish messages to maintain our interaction controller (ic) which coordinates robot behavior based on past and present events (i.e., Tega action (ta) and state(ts)). Camera images are fed into our emotion engine (ee) to capture the participant's head orientation and facial affect features for post-study analyses. Information such as this as well as all messages published are recorded in our logging module (control panel (cp)).

6. HUMAN-SUBJECTS EXPERIMENT

Through a human-subjects experiment in which children told stories to an audience of two robots with contingent and non-contingent backchanneling strategies, we investigate the effects of an attentive listening robot on a child's storytelling.

6.1 Hypotheses

We hypothesize that children will prefer to engage with the more attentive and contingent BOP robot. We demonstrate this preference through both behavior-based and subjective measures:

- **H1**: Children will direct their storytelling more toward the contingent BOP robot.
- **H2**: Non-contingent BC responses from a robot will interrupt children's storytelling.
- **H3**: Children will perceive the contingent BOP robot as more attentive and interested in their story.

6.2 Participants

Twenty-three children (age $M = 6.13$, $SD = 1.36$; 43.5% female) were recruited through a local parents' mailing-list. All children interacted with both the contingent and non-contingent robots at the same time. Such study design, a comparison study that presents two stimuli concurrently, is a popular method utilized in studying child language acquisition through social interaction [17] as well as children's vocabulary learning with a robot companion [4] and studying users' perception of a robotic agent's verbal and nonverbal feedback in a dialogue interaction [9].

6.3 Experimental Conditions

After a short introduction to the robots, participants were brought to the experimental room and was asked to tell stories to the two robots (Figure 5). One of the robots provided contingent BC feedback using the BOP model. The other robot provided random non-contingent BC feedback with a set frequency.

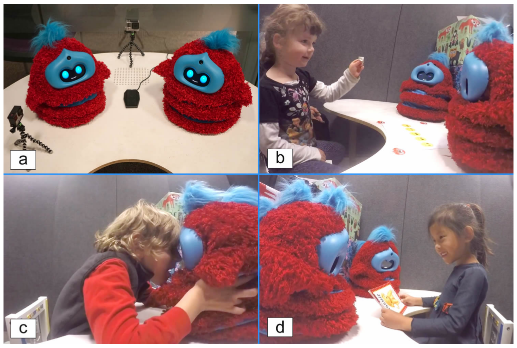


Figure 5: (a) Experimental room setup with two identically looking Tegas each representing the contingent and non-contingent condition. (b-c) Participants are engaged in storytelling with the robots. (d) a child is presenting a sticker to the more attentive robot.

- **Contingent Robot:** The behavior of the contingent robot closely mirrored that of the child listener's. Using the analyses on child listener reponses in Table 1, we developed twenty combinations of facial and body expressions that represent BC responses — gazing, leaning forward, nodding, smiling, eye-widening to emulate eyebrow movement, and short utterances [3]. When a BC opportunity is detected, the robot adapted its level of expressiveness depending on the speaker's energy level. For instance, when high energy level is detected at an energy event, the robot plays large excited motion.
- **Non-contingent Robot:** The non-contingent robot used the same set of animated nonverbal behaviors as the contingent robot but gazed at the speaker randomly and did not adapt its expressiveness to the speaker's energy level. However, the overall expressiveness of the behavior was matched in realtime to the contingent robot. This robot triggered animations and gaze behavior every 5.53 ± 1.5 seconds, which is the average frequency children produces BC response in our storytelling dataset.

To prevent any bias in the study results, we carefully controlled the robots' appearance and behavior. The robots looked identical, used the same name, and the expressivity level of the behaviors was matched between conditions. The placement of the robots was counter-balanced — in 45% of the sessions, the contingent robot was placed on the left side and 55% of the sessions on the right.

6.4 Study Protocol

The study procedure had three phases: a story brainstorming, a storytelling session with robots, and a post survey. The story brainstorming session took place in the waiting area while other phases were conducted in the study room with the robots.

6.4.1 Story brainstorming

At the time of study enrollment, parents were asked to provide information on what story their child likes to tell.

[3] Please refer to the trailer video for examples of these behaviors.

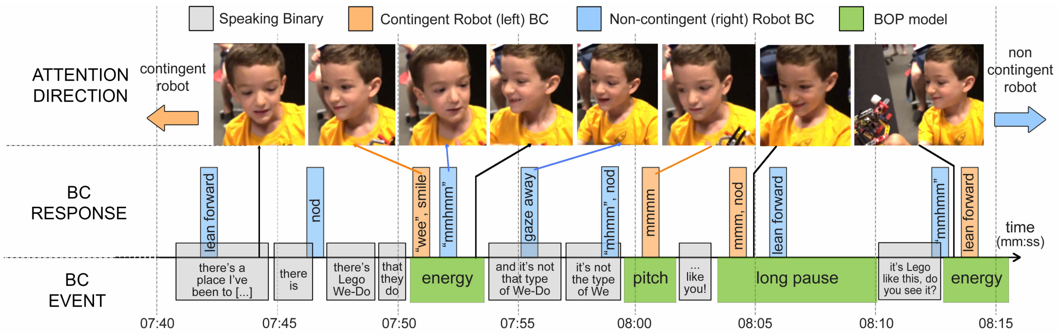


Figure 6: A snippet of a study session depicting the *backchannel events*, two robots' *backchannel responses*, and a child's *attention* towards the robots. The orange bar shows the contingent behavior of the BOP robot, and the blue bar shows a static frequency BC response of the non-contingent robot. The child significantly gazed more towards the contingent robot (on his left side) when storytelling.

Using this information, the experimenter engaged the child in a story brainstorming session. The experimenter asked about the child's experience over the breaks or used graphic books to help children who had difficulty in creating a story of their own. Afterwards, the experimenter provided the following backstory:

> *We have a problem. The two Tegas you were supposed to meet today are baby Tegas. They fell asleep and I can't wake them up. But their favorite activity is listening to children's stories! Maybe if you tell them you're here to tell them stories, they might wake up! Would you like to come try?*

This session successfully prepared children for the following phase, and only one child refused to tell stories to the robots.

6.4.2 Storytelling interaction with robots

Participants were brought to the study room with two sleeping Tegas on a table. The child was asked to sit on a chair in the center of the table, and the parents were invited to observe the session from a chair three feet behind the child. Children initiated the interaction by either gently rubbing, greeting, or telling Tegas that they were here to tell stories. Tegas "woke up" yawning at random intervals and started backchanneling as the participant told his/her story. When the child indicated he/she was done, the robots fell back asleep before the post survey began in order to prevent the child from feeling bad about making comments on Tegas' behaviors.

6.4.3 Post survey

The post survey consisted of questionnaires asking the likeability of the robots (how much did you like Tegas?), enjoyability of the storytelling task (how much did you like telling stories to Tegas?), and the level of interest each robot showed toward the story (how much do you think this Tega enjoyed your story?) in a 5-point Likert scale based on smileyometer. Participants were then asked to give a sticker to the robot they thought to be the better listener and was more interested in their story. The experimenter further asked the reasons behind their answers.

7. RESULTS AND DISCUSSION

Among 23 participants, we were able to analyze data from 20 children (age $M = 6.25$, $SD = 1.33$; 45% female). One 4-year-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 could not extract gaze-orientation data from the videos. The average length of children's stories was 10.77 ± 4.12 minutes. We found no statistical significance in the number of BC feedback provided and the level of expressiveness of the BC motions (categorized as small or large) between the contingent and non-contingent robots, thereby we can safely assume that the expressiveness of both robots was similar. Figure 6 presents a snippet of one of the study sessions.

We evaluated whether the contingency of a robot's BC response affects children's storytelling behavior by analyzing children's gaze patterns, affective reactions toward each robot, and post-survey answers.

7.1 Children direct their storytelling more toward the contingent robot

We analyzed children's gaze pattern using the yaw data of the head orientation. We also correlated this data with speaking binary to understand which robot the child attends to while telling a story.

The overall gaze direction during the entire interaction showed insignificance between the two robots measured as a fraction of each session length (contingent: $M = 0.359$, $SD = 0.070$, non-contingent: $M = 0.396$, $SD = 0.076$; $t(38) = 1.598$, $p = 0.118$). However, children significantly gazed more at the contingent robot when telling a story (SB=1) (contingent: $M = 0.185$, $SD = 0.076$, non-contingent: $M = 0.146$, $SD = 0.040$; $t(38) = 2.031$, $p = 0.049$). This confirms our main hypothesis **H1**. Also, when no speech event was detected (SB=0) children gazed more at the non-contingent robot (contingent: $M = 0.174$, $SD = 0.031$, non-contingent: $M = 0.250$, $SD = 0.053$; $t(38) = 5.523$, $p < 0.01$). An inspection of the videos suggests that the non-contingent robot's random feedback interrupts the child's speech prompting a momentary gazing reaction (Figure 7).

7.2 Children are distracted by the non-contingent robot

To better understand the affective content of participants' facial expressions, we used Affdex, a commercially available automated affect recognition software [1]. It extracts 15 physical expressions from facial features that are used as predictors to calculate the likelihood of an exhibited emotion as well as to estimate the degree of valence (positive and negative emotion) and expressiveness (intensity of an expression).

Although no significant difference was found in positive affect between the two conditions, an analysis of expres-

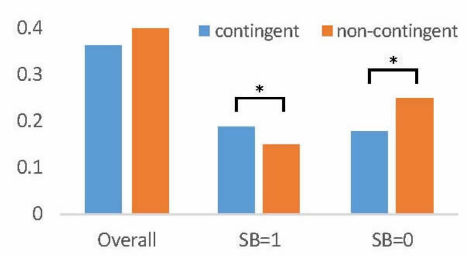


Figure 7: An average fraction of gaze length over the entire length of a session. Children gazed significantly more at the contingent robot ($p < 0.05$) when telling a story ($SB = 1$).

siveness (continuous scale of [0,100]) showed that children were more calm towards the 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 towards the non-contingent robot, which children described the robot as "'funny", "made me laugh", and "shy." Analysis revealed high correlation between expressiveness and pauses from storytelling (SB=0) ($\big[$SB=0: $M = 67.83$, $SD = 19.21$ $\big]$;$\big[$SB=1: $M = 54.25$, $SD = 12.38$$\big]$; $t(38) = 2.658$, $p = 0.012$), consistently suggesting that children paused from storytelling and reacted affectively to the non-contingent robot making random feedback (Figure 8).

The analyses of the participants' gaze behavior and affect response consistently suggest that the non-contingent robot distracts children from telling a story, supporting our second hypothesis, **H2**.

7.3 Children perceive the contingent robot more attentive

From the 5-point Likert-scale survey, our analysis revealed a high degree of likeability towards Tegas ($M = 4.70$, $SD = 0.66$) and enjoyability of telling a story to Tegas ($M = 4.50$, $SD = 0.69$). When asked about the robots' perspective, most children answered both Tegas enjoyed their story, and no difference was observed between the two conditions (contingent: $M = 4.63$, $SD = 0.60$; non-contingent: $M = 4.53$, $SD = 0.61$). Fisher's exact test revealed that there is no statistical significance between which side the contingent robot was placed versus the robot child indicated as a better listener.

Based on the sticker test, 15 out of 20 children selected the contingent robot as the more attentive listener than the non-contingent robot (p=0.0038). Children who chose the non-contingent robot as the more attentive listener reported that the robot "made large motions" ($N = 1$), "seemed very happy/excited" ($N = 2$), and "made less 'mmm' sound" ($N = 4$).

We particularly find the last reason interesting, since as observed in Section 3.2, only 50% of child listeners use short utterances like 'mmm' as backchannel feedback. This observation in conjunction with how children decode what they encode from Section 3.4, we can assume that a possible reason behind why these children did not choose the contingent robot is because they perceived the backchannel utterances as interruptive and distracting rather than as a signal of an attentive listener (since they themselves do not use the signal in a similar way).

In summary, subjective analysis on children's questionnaire responses supports our last hypothesis **H3**, that children would perceive the contingent BOP robot as more attentive and interested in their story.

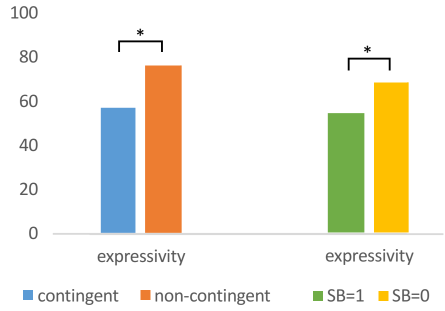


Figure 8: Level of expressivity is shown in scale [0,100]. Children's facial expressivity significantly increases toward the non-contingent robot. Correlation analysis reveals that children tend to be more calm when telling a story ($SB = 1$).

8. CONCLUSION

Our work has addressed the challenge of developing a backchannel prediction model based on observed child nonverbal behaviors as well as evaluating its effects when utilized as a listening robot. We began by first understanding the demonstrated nonverbal behaviors of children in a storytelling context. We identified backchannel behaviors—partner gazes, leaning toward, smiles, utterances, brow raises, and nods—that indicate a positive engagement state of the listener. We identified speaker cues—gaze, pitch, word, energy, pauses taken singly and in combinations—that children listeners acknowledge and respond to. We characterized the bidirectionality of nonverbal behaviors in that children understand the function of nonverbal behaviors both in the role of the communicator and recipient. Children can decode nonverbal behaviors they encode. Based on our findings, we developed a rule-based backchannel opportunity prediction (BOP) model capable of accurately predicting backchanneling opportunities.

Our aim was to evaluate two crucial questions that motivate the design and development of social robot learning companions that can successfully foster early language skills of preschoolers and kindergarteners. Can a social robot generate backchannel behaviors that children perceive as attentiveness? Is contingency of agent feedback crucial toward creating these perceptions? Our study suggests that contingency matters as it leads children to appropriately direct their storytelling to their audience and also contributes to the perception of an attentive audience.

9. ACKNOWLEDGMENTS

We thank professors Stephanie Gottwald, David DeSteno, and Paul Harris for their feedback in this research. We also thank Aradhana Adhikari, Katharine Brush, Raina Hall, Emily Holding, Sean Ko, Yiyu Li, Madelyn Payne, Jose Salas, Matthew Sears, and Michal Shlapentokh-Rothman for annotating videos and helping conduct the user studies. This work was supported by the National Science Foundation (grant IIS-1523118), Dolores Zohrab Liebmann Fellowship, and William Asbjornsen Albert Memorial Fellowship.

10. REFERENCES

[1] Affectiva. Emotion recognition software and analysis, 2016.

[2] T. Behne, U. Liszkowski, M. Carpenter, and M. Tomasello. Twelve-month-olds' comprehension and production of pointing. *British Journal of Developmental Psychology*, 30(3):359–375, 2012.

[3] C. Breazeal, K. Dautenhahn, and T. Kanda. Social robotics. In *Springer Handbook of Robotics*, pages 1935–1972. Springer, 2016.

[4] C. Breazeal, P. L. Harris, D. DeSteno, K. Westlund, M. Jacqueline, L. Dickens, and S. Jeong. Young children treat robots as informants. *Topics in cognitive science*, 2016.

[5] N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 51–58. Association for Computational Linguistics, 2003.

[6] H. H. Clark and S. E. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149, 1991.

[7] A. R. Dennis and S. T. Kinney. Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information systems research*, 9(3):256–274, 1998.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[9] S. Fujie, K. Fukushima, and T. Kobayashi. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *Proc. ICARA Int. Conference on Autonomous Robots and Agents*, pages 379–384, 2004.

[10] G. Gordon, C. Breazeal, and S. Engel. Can children catch curiosity from a social robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 91–98. ACM, 2015.

[11] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[12] A. Gravano and J. Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *INTERSPEECH*, pages 1019–1022, 2009.

[13] L. J. Hess and J. R. Johnston. Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, 11(3):319–335, 1988.

[14] L. Huang, L.-P. Morency, and J. Gratch. Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 1265–1272. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[15] P. H. Kahn Jr, T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, and S. Shen. "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2):303, 2012.

[16] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 643–648. IEEE, 2014.

[17] P. K. Kuhl. Is speech learning 'gated' by the social brain? *Developmental science*, 10(1):110–120, 2007.

[18] L.-P. Morency, I. de Kok, and J. Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2010.

[19] J. Movellan, M. Eckhardt, M. Virnes, and A. Rodriguez. Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 307–308, 2009.

[20] H. Noguchi and Y. Den. Prosody-based detection of the context of backchannel responses. In *ICSLP*, 1998.

[21] Y. Okato, K. Kato, M. Kamamoto, and S. Itahashi. Insertion of interjectory response based on prosodic information. In *Interactive Voice Technology for Telecommunications Applications, 1996. Proceedings., Third IEEE Workshop on*, pages 85–88. IEEE, 1996.

[22] H. W. Park and A. M. Howard. Understanding a child's play for robot interaction by sequencing play primitives using hidden markov models. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 170–177. IEEE, 2010.

[23] H. W. Park and A. M. Howard. Retrieving experience: Interactive instance-based learning methods for building robot companions. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6140–6145. IEEE, 2015.

[24] H. W. Park, R. Rosenberg-Kima, M. Rosenberg, G. Gordon, and C. Breazeal. Growing growth mindset with a social robot peer. In *Proceedings of the 12th ACM/IEEE international conference on Human robot interaction*. ACM, 2017.

[25] C. Peterson, B. Jesso, and A. McCabe. Encouraging narratives in preschoolers: An intervention study. *Journal of child language*, 26(01):49–67, 1999.

[26] R. Poppe, K. P. Truong, D. Reidsma, and D. Heylen. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*, pages 146–158. Springer, 2010.

[27] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[28] M. Shatz and R. Gelman. The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the society for research in child development*, pages 1–38, 1973.

[29] M. Shiomi, T. Kanda, I. Howley, K. Hayashi, and N. Hagita. Can a social robot stimulate science curiosity in classrooms? *International Journal of Social Robotics*, 7(5):641–652, 2015.

[30] S. Spaulding, G. Gordon, and C. Breazeal. Affect-Aware Student Models for Robot Tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[31] F. Tanaka and S. Matsuzoe. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1), 2012.

[32] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. 2011.

[33] K. P. Truong, R. Poppe, and D. Heylen. A rule-based backchannel prediction model using pitch and pause information. 2010.

[34] N. Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1728–1731. IEEE, 1996.

[35] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207, 2000.

[36] J. K. Westlund, J. J. Lee, L. Plummer, F. Faridi, J. Gray, M. Berlin, H. Quintus-Bosz, R. Hartmann, M. Hess, S. Dyer, and others. Tega: A Social Robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 561–561. IEEE, 2016.

[37] S. White. Backchannels across cultures: A study of americans and japanese. *Language in society*, 18(01):59–76, 1989.