Contingent Backchannel Feedback Affects Children's Storytelling Behavior Towards Social Robots

Hae Won Park[1,*], Mirko Gelsomini[1,2,*], Jin Joo Lee[1], and Cynthia Breazeal[1]

***Abstract*— Inline with our goal of developing a personalized storytelling and listening companion for improving children's language skills, this paper investigates how a robot that can produce contingent listener response, i.e., backchannel, can deeply engage children as a storyteller. We trained a rule-based backchannel model from extracting backchannel-event labels and acoustic features such as voicing probability, pitch, energy, and formants from 58 episodes of children's dyad storytelling activity. A user study was conducted in which child participants (age range 4-8, $M = 6.13$, $SD = 1.36$) told stories to two social robots, one using the developed backchannel model and the other using an average frequency of backchannel response acquired from the dyad dataset with human coder labels. The results from analyzing children's gaze pattern, emotional features, and post survey questionnaire on perceived likeability, enjoyability, and attentiveness of the robot are reported along with the performance evaluation of our backchannel opportunity prediction model.**

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 resposne), joint attention, and shared gaze for language acquisition, thereby emphasizing the importance of a cooperative effort of the teacher and the learner [3].

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. However, 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. 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.

This work was supported by the National Science Foundation, under grant IIS-1523118.

[1]Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal are with MIT Media Lab, 20 Ames St., Cambridge, MA, 02141, USA. haewon,gelso,jinjoo,cynthiab@media.mit.edu

[2]Mirko Gelsomini is also with Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy. mirko.gelsomini@polimi.it

*These authors contributed equally to this work.

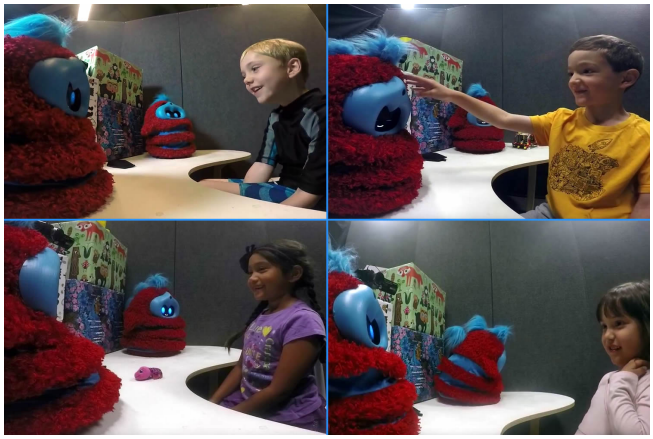


Fig. 1. Four children interacting with Tega

For this purpose, we developed a rule-based backchannel opportunity prediction model that will push the envelope of our understanding of childrens storytelling and eventually facilitate the development of early language skills in preschool aged children. Our model has been trained and tested on K2 children's dyad storytelling dataset collected from local Boston public schools. We then conducted a user study to answer the question of whether an attentive robot listener utilizing our model can positively affect a child storyteller's behavior.

II. RELATED WORK

A long-term 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]. The use of social robots as peers and tutors for children in educational settings has also been increasingly explored in recent years with promising results [8], [9]. Children have learned vocabulary from a tele-operated storytelling robot [10] and fostered curiosity-relevant behaviors [11]. Socially assistive robots have also been introduced in Kindergarten settings as teacher assistants to foster storytelling activities [12]. Taken together, findings such as these are highly suggestive of the ability of social robots to be perceived as engaging peers or to serve as instructors in learning. In other works, studies have shown that contingent behavior may help mitigate awkwardness during silences and can improve conveying sincerity of the robot in maintaining conversations towards the user [13] and affects children's preferred choice of informant, almost on-par with human collaborators [14].

Backchanneling (BC) is a component of conversation and verbalization that is naturally embedded in our everyday interaction and is the part a listener plays in a conversation. There are both verbal and nonverbal BC signals. Throughout a conversation, the listener may nod their head (nonverbal) periodically to show that they are paying attention and/or verbally acknowledge using feedbacks, such as *yeah, ok, uh huh, mhmm*. BC has been studied as a form of feedback, acknowledgment, and turn-taking in both the psychology field as well as in human-robot interaction. In [15], authors report that BC serves four cognitive functions including indicating understanding, or lack thereof, repair or clarification of the message, and sentence completion. From this perspective, BCś main function is in establishing common ground by signaling that the receiver has understood the message [16]. Moreover, in [17], it is demonstrated how BC feedback provides the teacher to understand the robot learner's states while interacting with a child. Studies demonstrate that robots' positive feedbacks and rewards motivate children in the Autism Spectrum Disorder (ASD) [18], [19], [20].

A. Computational Backchannel Strategies

Since listener BC are generated rapidly and seem elicited by a variety of speaker verbal and nonverbal cues, generating appropriate BC is a difficult problem. There is evidence that people can generate such feedback without necessarily attending to the content of speech [21], and this has motivated diverse approaches that generate BC using different features that are available in real-time (e.g. VAD, energy, pitch).

We propose to generate autonomous nonverbal behaviors by using a broad range of real-time features. Based on the analysis of our existing corpus of child-robot tele-operated storytelling sessions from our prior study, we created a rule-based backchannel prediction model for children which, to our knowledge, does not exist at the moment. Commercial tools to extract prosody and timing in speech will be used to find differences in BC style, as study [22] proposes, and predict BC opportunities, which is an important milestone for building engaging [21] and natural [23] experiences. We will then integrate the generated autonomous nonverbal behaviors into the Tega robot and the responses will be variable, so as to increase engagement and believability.

III. THE SYSTEM

Tega is an Android smartphone-based robot with a fluffy exterior that is designed to animate squash-and-stretch-based motions with 5 degree of freedom (Fig. 8). Twenty facial expressions and gesture animations were developed for this study. These included BC motions like bowing and nodding, as well as expressions of joy, surprise, satisfaction, agreement, interest, and excitement in lip sync. While not backchanneling the robot breaths, blinks, gazes the user in the contingent BC condition and looks around in the non-contingent BC condition.

The system for the contingent robot, developed on ROS (Robot Operating System: a framework for robot software development providing operating system-like functionality on a heterogeneous computer cluster), includes the following nodes:

- OpenSmile: an audio feature extraction tool enables to extract large audio feature spaces in real time [22]
- SpeakingBinary: a voice activity detection (VAD) classifier that mixes live features from OpenSmile with our noise cancellation algorithm. Our implementation distinguishes voice signals from stationary and low-noise/external-voice signals based on energy variance and incremental noise reduction. Therefore, if the output exceeds a moving threshold value the signal is considered to be a voicing event.
- BackChannel: according to the SpeakingBinary node output, energy (Log and RMS), and pitch score and direction, the BackChannel node selects the right BC rule x and randomize the execution of a BC from a predefined set y of BC rule x. Prosodic features such as energy and pitch of a voice are indicators for detecting emotion, uncertainty, questions, and statements. The system uses these features as an input to our BC prediction model (details in Section IV-E) and detects BC opportunities in few milliseconds.
- RobotExecutor: receives the instructions from the BackChannel node, manages a small message queue, and send the action to TegaAction node. Thanks to its queue, the node is able to manage the turn-taking by: a) understanding if, at the same time of the BC opportunity, the speaker starts speaking b) not listening while it is actually speaking.
- TegaAction: running on the Tega Android phone, listens to Tega actions sent by RobotExecutor and executes the behavior (i.e. a list of motion, lookat, sound-speech).
- Affdex: using advanced facial analysis, it measures emotional responses from a video frame [24]

The system for the non-contingent robot is fairly simple: BackChannel, RobotExecutor, TegaAction work together to deliver randomize behavior from the same contingent behavior set every 5 +- 1.5 seconds (Section IV).

A. Audio Feature Extraction

As stated in III, we decided to use OpenSmile [22]. It is written in C++ and is available open-source as both a standalone commandline executable as well as a dynamic library. The main features of openSMILE are its capability of on-line incremental processing and its modularity. Feature extractor components can be interconnected to create new and custom features, all via a configuration file. A ROS Sink (live publisher) and a CSV Sink (off-line saver) have been developed using openSMILE API to publish the following features (F):

- F0: smoothed fundamental frequency contour and voicing probability
- Energy: voice loudness in Log scale
- Pitch: voice frequency direction (fall: -1, flat: 0, rise: 1)

IV. APPROACH

A. Data Collection

Participants of typical development were recruited from a Boston public elementary school, whose curriculum already included an emphasis on storytelling. Eighteen children from a single kindergarten (K2) classroom participated in the study ($M = 5.22$ years-old, $SD = 0.44$, 39% female). Each student participated in at least three rounds of storytelling but with a different partner. In a dyad session, the pair of students take turns (T1, T2) narrating their story to the other. In sum, the data collection consists of 3 rounds (with a supplementary 4th round for redo opportunities) totaling 29 dyad sessions, which equates to 58 individual storytelling episodes. Three time-synchronized cameras captured the frontal-view of each participant along with a birds-eye view. For each storytelling episode, the nonverbal behaviors of both the listener (L) and storyteller (S) were manually coded using a video-annotation software. Four coders marked the onset and offset times for the occurring nonverbal behaviors: speaker turn for both S and L, type of short utterances for L only, and speaking times for S only. 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).

B. Observation of Children's Backchanneling Behavior

According to [25], backchannel response behavior is one of the last communication skill acquired. The frequency of back channel responses increased significantly with age. Upon analyzing the dyad dataset, we also found that children's backchannel frequency is significantly less than the adult coders' (child: $M = 2.18$ times per second, $SD = 1.93$, adult: $M = 5.10$, $SD = 1.47$; $t(114) = 9.17$, $p < 0.01$). Children used more gaze-oriented response behavior compared to utterances or nod gesture. The question of whether children can dictate adult's backchannel listener response is debatable, but many report that young children can perceive the signals [26], [25]. Hence, we developed a rule-based backchannel opportunity prediction model, detailed in Section IV-E, based on adult's backchannel behavior data.

We developed a visualization tool to easily observe a broad set of features while playing the video recordings. The tool enables user to select a video of the dyadic activity (Fig. 2 top center), a set of features (up to 5, Fig. 2 bottom center) and adjust different parameters such as Speed, Focus Window, etc.. (Fig. 2 top right).

In this way, we have been able to precisely query the value of different features at once at a particular time frame/window and develop our own model accordingly.

C. Analysis of Audio Feature Data

To evaluate the accuracy of VAD we developed a Testing Framework able to:

- analyze each episode with OpenSmile and export the features F with a sampling rate of 10 ms

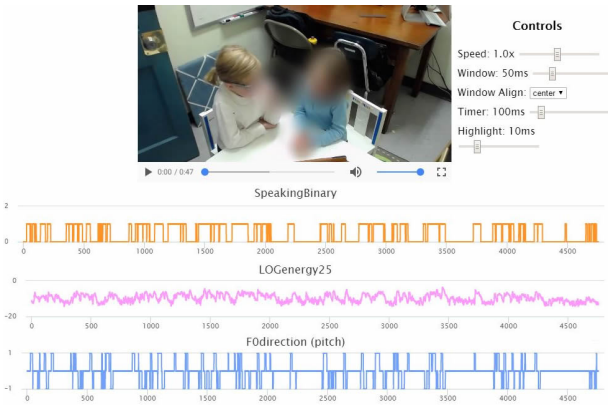


Fig. 2. Visualization tool

- read the list of exported features F for each episode, E where $E = [1...58]$
- import the list of exported labels from each coder
- merge the coders' tags to create 3 levels of consensus (L1, L2, L3). Constructed from the concordances and discordances among the tags' timestamps (hereinafter consensus), Level 1 (L1) represents when only 1 coder is considered, whereas Level 3 (L3), merges together the data of each coder for a specific storytelling episode E.
- compare ground truth data (consensus from video-coders) against the measurements (speaking binary, speaking cues) given by our system.
- generate all the combinations C between different feature values given as input, for all the episodes E[1...58].
- save the results for each episode E for each combination C (trail Tc), exporting Precision, Recall and FScore.
- save the global results averaging the precision, recall and fscore of all episodes E[1...58] in a given trail Tc

We then followed the above procedure to develop and test our VAD classifier and our rule-based BC Model.

D. VAD Classifier

Our VAD Classifier has been developed on top of OpenSmile input features and adjusted by continuously iterating on the ground truth data. The OpenSmile Voice Activity Detector (VAD), based on Line-Spectral-Frequencies, Mel spectra, and Energy, computes Fuzzy scores related to the deviation from the observed long-term mean values [27]. The VAD outputs 0 when no one is speaking, 1 when someone is speaking (hereinafter Speaking Binary, SB). We run the Testing Framework comparing SB from OpenSmile, OS (without any modification), and SB from ground truth, GT.

- Precision $P = TP/(TP + FP) = 92.5\%$
- Recall $R = TP/(TP + FN) = 75.5\%$
- Fscore $Fs = 2 * (P * R)/(P + R) = 82.8\%$

where TP is true-positive ($[GT, OS] = [1, 1]$), FP is false-positive ($[GT, OS] = [0, 1]$), and FN is false-negative ($[GT, OS] = [1, 0]$).

Then, we compared them using the visualization tool and noticed that, by using a combination of OpenSmile features (and not just the VAD output), we could have improved the Speaking Binary accuracy. We created a filter and normalization functions on top of OpenSmile features and run the Testing Framework iteratively. We implemented a low-pass filter as that:

Considering the current value of Speaking Binary at time t-1, SB(t-1), the next value at time t, SB(t) = SB(t-1)*SB_LAST+SB(t)*SB_CURRENT. Then we use a SB_CUTOFF value to decide whether the final SB value was 1 (SB greater than SB_CUTOFF) or 0 (SB less than SB_CUTOFF). Up to that we developed a normalization function to distinguish voice signals from stationary and low-noise/external-voice signals by normalizing the energy variance through an incremental noise reduction that started after SB_NORMSTART milliseconds.

Fig. 3. Visualization tool to compare Ground Truth vs our Speaking Binary

Precision and Recall saturated at 96.8% and 87.5%. We further investigated the reason for the saturation as follows, which was mainly due to the human coder limitations:

- at SB falling edge (when speaker pauses), there is a perceptual delay of the coder with a mean of 41ms (4 frames) which influences Recall.
- at SB rising edge (when speaker starts speaking), there is a perceptual delay of the coder with a mean of 28ms (3 frames) which influences Precision.

Given that the human reaction time for a fast-click activity is 276 ms in average [28] and that the video-coding software we used eases eight times the coding task (e.g. reducing speed, zooming), we can conclude that our system is fully reliable on measuring the Speaking Binary by using the following values:

- SB_LAST = 0.8
- SB_CURRENT = 1 - SB_LAST
- SB_CUTOFF = 0.9
- SB_NORMSTART = 700 ms

E. Reference BC Prediction Models

- Wordy Model (Fig. 4)

The rule-based algorithm introduced by [29] predicts BC based on IPU length (W). An Inter Pausal Unit (IPU) is a maximal sequence of words surrounded by silence longer than 50 ms. A turn then is defined as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs there is no speech from the interlocutor.

A BC opportunity comes upon detection of:

```
P1 a pause of W_PAUSE length,
P2 preceded by at least W_SPEAK of speech,
P3 provided that no BC has been output
    within the preceding BC_RATE
```

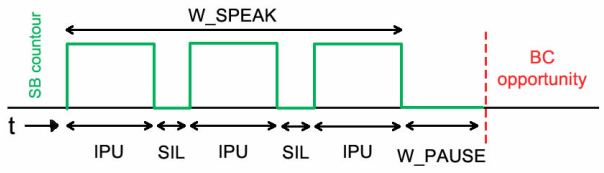


Fig. 4. Wordy model

- Long Pause Model (Fig. 5)

The rule-based algorithm introduced by [30] computes the BC feedback upon detection of a long pause LP that is preceded by a speech $LP_SPEAK less than W_SPEAK$.

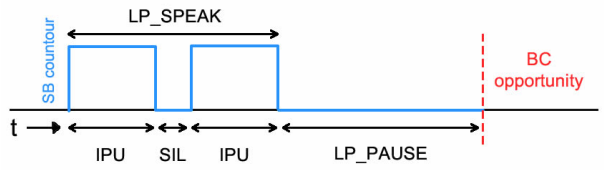


Fig. 5. Long Pause model

The feedback is executed upon detection of:

```
P1 a pause of LP_PAUSE length (900ms),
P2 preceded by at least LP_SPEAK of speech,
P3 provided that no BC has been output
    within the preceding BC_RATE
```

- Pitch Model (Fig. 6)

The rule-based algorithm introduced by [31] and enhanced by [32] predicts BC based on Pitch and Pause (P&P) features in English audio-only settings.

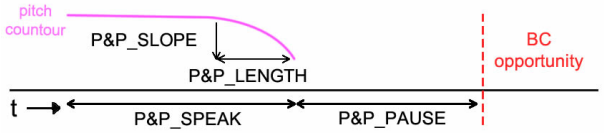


Fig. 6. Pitch & Pause model

The algorithm executes the feedback, upon detection of:

```
P1 a pause of P&P_PAUSE (400ms),
P2 preceded by at least P&P_SPEAK (1000ms)
    of speech,
P3 where the last P&P_LENGTH (100ms),
P4 contain a rising/falling pitch of at
    least P&P_SLOPE rise/drop (30Hz).
P5 provided that no BC has been output
    within the preceding BC_RATE (1400ms).
```

In short, the P&P algorithm computes the BC feedback upon detection of a pause, and a falling or rising pitch slope, that is preceded by speech.

- Energy Model (Fig. 7)
Based on [33], the BC feedback activates, upon detection of:

```
P1 a pause of E_PAUSE,
P2 preceded by at least E_SLOPE_LENGTH of
    speech,
P3 contain a rising/falling energy of at
    least E_SLOPE rise/drop.
P4 provided that no BC has been output
    within the preceding BC_RATE.
```

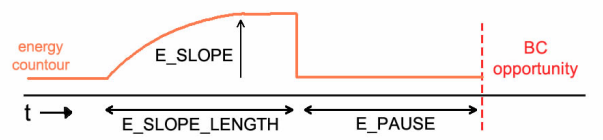


Fig. 7. Energy model

F. Analysis of Backchannel Consensus Data

We applied the same logic of IV-C to define the right times for each model. As explained above we utilized available rule-based models and changed according to our data collected with K2 children. Since With few training data, our parameter estimates will have greater variance, whereas with few test data, our performance statistic will have greater variance. What is the compromise

The 59 episodes were partitioned into training and test sets in a ratio of 42:17. A classification model was trained on the training set and was applied to the test set. The test set (28%) has never been seen by the model (72%) so to have a good balance between parameter estimates and statistics performance. In order to ensure that the agreement or consent of all participants is valued, we choose to test our system against a ground truth with a consensus level 3. Level 3 consensus compares the labels of the 3 video-coders; merges them only if, in a given time frame (1000ms), the labels tagged are the same and removes those labels that were not in common. During the evaluation we considered Precision (other than recall) as performance measure of the estimates. Precision, in the case of matching both labels and times, represents the percentage of accuracy of finding many label matches (hit) and few misses (ground truth says to BC, our model says no). By focusing on precision we are then explicitly not considering the miss error rate when our model says to BC and the ground truth does not. Taking into account that BC_RATE, the minimum rate at which the robot can BC, should be high and that the robot cannot BC while the speaker is speaking, the recall can be unstudied. The evaluation of the test set returned the following precisions:

- Wordy: 89.5%
- Long Pause: 78.3%
- Pitch & Pause: 61.1%
- Energy: 67.3%

using the following parameters:

- BC_RATE = 3000 ms
- W_PAUSE = 800 ms
- W_SPEAK = 1500 ms
- LP_PAUSE = 1700 ms
- LP_SPEAK = 1000 ms
- P&P_SLOPE = 25 %
- P&P_LENGTH = 300 ms
- P&P_PAUSE = 400 ms
- E_SLOPE = 30 %
- E_SLOPE_LENGTH = 500 ms
- E_PAUSE = 300 ms

V. EXPERIMENTAL SETUP

Fig. 8. Setting: a contingent and a non-contingent Tega (unordered)

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. In order to evaluate our hypothesis, the experimental room was setup as depicted in Fig. /reffig:setting that had identically looking Tegas to the left and the right side of the child sitting in the center. After a short introduction of the robots, the child was brought to the experimental room and was asked to tell stories to the robots. One of the robot was providing contingent BC feedback with audio and gaze using our algorithm detailed in Section IV, and the other robot was providing random non-contingent BC feedback every 5 ± 1.5 seconds based on our research on the average interval of backchanneling from human coders. The rest of the features (appearance, expressivity, and name) of the robots remained identical in order not to bias participants' preference of the robot. The robots were separated far enough so that it was obvious which robot the child was gazing at at a given time. We used an external microphone to capture audio signals of the child during storytelling and video recorded the interaction from the front (child face close up) and the side (full study view). The frontal view was used post study to analyze children's gaze pattern and affect states.

A. Participants

Twenty three children (age $M = 6.13$, $SD = 1.36$; 43.5% female) between the age of 4-8 years old were recruited

to participated in the study. All children were in a single condition in which they interacted with the contingent and non-contingent robots at the same time. The study protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) review board, and a consent was collected from the parent as well as a verbal assent from the child.

B. Protocol

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

1) Story brainstorming: At the time of study enrollment, parents were asked to provide information on what story their child likes to tell. Using this information, the experimenter engaged the child in a story brainstorming session. The experimenter used graphic books to help children who had difficult time creating a story of their own or asked about they experience. Afterwards, the experimenter provided the following backstory of Tega to help the child immerse in the interaction:

> *We have a problem. The two Tegas you were supposed to meet today are baby Tegas and they fell asleep and I can't wake them up. But their favorite activity is listening to children's stories, though they are still learning language and can't speak yet. May be 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.

2) Storytelling interaction with backchanneling robots: The child participant was brought to the experimental room with two Tegas fast asleep. The child was asked to sit on a chair in the center and the parent was invited to observe the session from a chair two feet behind the child. Children gently rubbed, greeted, and told Tegas they were here to tell stories. The Tegas then woke up yawning at random intervals. Among the session data we analyzed, 45% of the sessions had the contingent robot placed on the left side and 55% of the sessions on the right. The robots started backchanneling as the participant began telling stories. When the child indicated he/she was done, the robots fell back asleep.

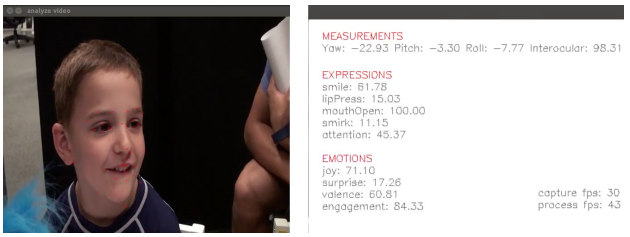
3) Post survey: The post survey consisted of questionnaires asking a likeability of the robots (how much did you like Tegas?), enjoyability of the storytelling task (how much did you like telling stories to Tegas?), and the level of interest each robot showed towards the story it heard (how much do you think this Tega enjoyed your story?) in 5-point smiley Likert scale. Children were then given stickers, some to keep for themselves and some to distribute to the robots. The experimenter asked the participant to give a sticker to the robot who was a better listener, then another sticker to the robot who they want to tell another story to. The experimenter asked and documented the reason of each answer.

C. Measurements

During the storytelling phase, we collected the total length of the interaction, acoustic features from participant's speech, and head orientation and facial affect features from the camera (Table I). We also recorded when the robots provided BC feedback and the expressivity intensity (small and large) of the animated motion they used to express response. All data was time synchronized.

TABLE I
DATA COLLECTED DURING STORYTELLING

Type	Features
Acoustic features	voicing probability, energy, pitch, formants
Head orientation	roll, pitch, yaw
Facial expressions	attention, brow furrow, brow raise, cheek raise, chin raise, dimpler, eye closure, eye widen, inner brow raise, jaw drop, lid tighten, lip corner depressor, lip press, lip pucker, lip stretch, lip suck, mouth open, nose wrinkle, smile, smirk, upper lip raise
Emotions	anger, contempt, disgust, fear, joy, sadness, surprise
Hidden affect features	valence, expressivity



(a) (b)

Fig. 9. Affect analysis using facial features. (a) facial features extracted from participant's face (b) head orientation and affect analysis (dominant expressions, and corresponding emotion probabilities)

VI. 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 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. Children gazed more at the contingent robot while storytelling

We analyzed children's gaze pattern using the yaw information of the head orientation (Fig. 9). At the moment each robot woke up from sleep, we detected a gaze-locking pattern and used it as a baseline to compute which robot the child was gazing at at a given time. We also correlated this data with speaking binary, i.e., when the child was speaking, in order to differentiate nonverbal affective reaction to a robot's motion versus attending to a robot 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 while telling a story (SB=1) (contingent: $M = 0.185$, $SD = 0.076$, non-contingent: $M = 0.146$, $SD = 0.040$; $t(38) = 2.031$, $p = 0.049$). The nonverbal (SB=0) reaction pattern also revealed significant difference between the two robots (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 interrupted the child's speech causing an affective reaction.

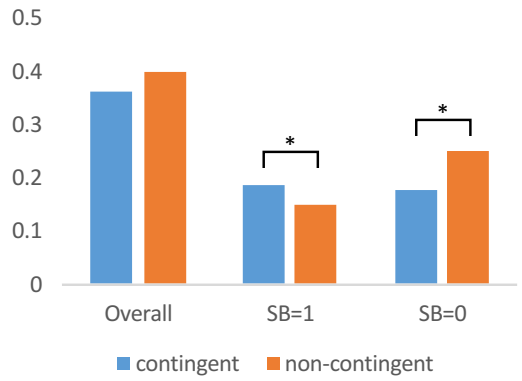


Fig. 10.

B. Children were more calm towards the contingent robot.

Affective signals hold information about a person's emotional and engagement state. From facial features (Table I), Affdex extracts 21 physical 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) (Fig. 9).

Analysis of expressiveness (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 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 affectively to the non-contingent robot making random feedback.

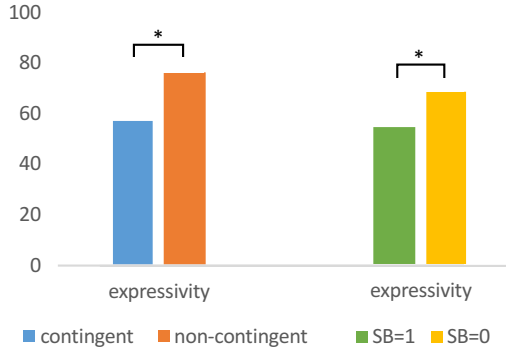


Fig. 11.

C. Children perceived the contingent robot more attentive.

After children finished telling stories, the two robots went back to sleep, and the experimenter conducted the post survey. A five-point Likert scale revealed high perceived 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 perspective of the robots, 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: $M = 4.53$, $SD = 0.61$). 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.

Among 20 children, 15 responded that the contingent robot was more attentive than the non-contingent robot (75%). Children who chose the non-contingent robot answered that the robot "made large motions" ($N = 1$), "seemed very happy/excited" (2), and "made less 'mmm' sound" (4). We particularly found the last reason interesting, since as discussed in IV-B, young children use significantly less filled pause feedback compared to adults, and thus could have been the reason why they perceived the contingent robot which often utilized filled pauses, e.g., uh-huh, mhmm, as a distraction.

VII. CONCLUSIONS

We presented a method to developing an attentive robot listener. Our observation revealed that children tend to attend to a contingent backchanneling robot more while telling

a story. As future work, we will conduct a 6-month longitudinal study at multiple preschool sites to evaluate the impact of long-term interactions with the storytelling robot on childrens engagement and language skill development. The future work will increase our understanding of the impact of longitudinal interactions with social robot companions on childrens 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. Parenting groups and educators shall be engaged to facilitate the learning activities as well as to provide input and feedback. Project-participating students will be trained in the multidisciplinary aspects of story generation, automatic assessment tools, robotics and developmental psychology.

REFERENCES

[1] B. Hart and T. R. Risley, *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, 1995.

[2] M. M. Páez, P. O. Tabors, and L. M. López, "Dual language and literacy development of spanish-speaking preschool children," *Journal of applied developmental psychology*, vol. 28, no. 2, pp. 85–102, 2007.

[3] L. J. Hess and J. R. Johnston, "Acquisition of back channel listener responses to adequate messages," *Discourse Processes*, vol. 11, no. 3, pp. 319–335, 1988.

[4] M. K. Lee, J. Forlizzi, S. Kiesler, P. Rybski, J. Antanitis, and S. Savetsila, "Personalization in hri: A longitudinal field experiment," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 2012, pp. 319–326.

[5] D. Leyzberg, S. Spaulding, and B. Scassellati, "Personalizing robot tutors to individuals' learning differences," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 423–430.

[6] E. Short, K. Swift-Spong, J. Greczek, A. Ramachandran, A. Litoiu, E. C. Grigore, D. Feil-Seifer, S. Shuster, J. J. Lee, S. Huang *et al.*, "How to train your dragonbot: Socially assistive robots for teaching children about nutrition through play," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 924–929.

[7] C. D. Kidd and C. Breazeal, "A robotic weight loss coach," in *Proceedings of the national conference on artificial intelligence*, vol. 22, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007, p. 1985.

[8] 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*. ACM, 2009, pp. 307–308.

[9] T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu *et al.*, "Multimodal child-robot interaction: Building social bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[10] J. J. M. Kory, "Storytelling with robots: Effects of robot language level on children's language learning," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[11] 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*. ACM, 2015, pp. 91–98.

[12] M. Fridin, "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education," *Computers & education*, vol. 70, pp. 53–64, 2014.

[13] N. Ohshima, K. Kimijima, J. Yamato, and N. Mukawa, "A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings," in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 2015, pp. 325–330.

[14] 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.

[15] 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*, vol. 9, no. 3, pp. 256–274, 1998.

[16] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on socially shared cognition*, vol. 13, no. 1991, pp. 127–149, 1991.

[17] H. W. Park, R. A. Coogle, and A. Howard, "Using a shared tablet workspace for interactive demonstrations during human-robot learning scenarios," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 2713–2719.

[18] A. Bonarini, F. Clasadonte, F. Garzotto, and M. Gelsomini, "Blending robots and full-body interaction with large screens for children with intellectual disability," in *Proceedings of the 14th International Conference on Interaction Design and Children*. ACM, 2015, pp. 351–354.

[19] F. Garzotto and M. Gelsomini, "Integrating virtual worlds and mobile robots in game-based treatment for children with intellectual disability," *Virtual Reality Enhanced Robotic Systems for Disability Rehabilitation*, p. 69, 2016.

[20] L. Bartoli, F. Garzotto, M. Gelsomini, L. Oliveto, and M. Valoriani, "Designing and evaluating touchless playful interaction for asd children," in *Proceedings of the 2014 conference on Interaction design and children*. ACM, 2014, pp. 17–26.

[21] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[22] L. Huang and J. Gratch, "Crowdsourcing backchannel feedback: understanding the individual variability from the crowds," in *The Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

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

[24] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2016, pp. 3723–3726.

[25] L. J. Hess and J. R. Johnston, "Acquisition of back channel listener responses to adequate messages," *Discourse Processes*, vol. 11, no. 3, pp. 319–335, 1988.

[26] W. A. Corsaro, "The clarification request as a feature of adult interactive styles with young children," *Language in Society*, vol. 6, no. 02, pp. 183–207, 1977.

[27] 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*. ACM, 2013, pp. 835–838.

[28] Human Benchmark. [Online]. Available: http://www.humanbenchmark.com/tests/reactiontime/statistics

[29] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue." in *INTERSPEECH*, 2009, pp. 1019–1022.

[30] 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*. Association for Computational Linguistics, 2003, pp. 51–58.

[31] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

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

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