Can an Attentive Robot Listener Better Engage Children as a Storyteller? *

Omitted for Blind Review

ABSTRACT

We improved the social interaction capability of a storytelling robot companion by developing a nonverbal backchanneling model of an attentive listener. By analyzing children's storytelling and backchanneling dyad interaction samples, we developed rules based on audio cues (a synergy between pitch, energy, and pauses) when a backchannel is the most effective. The proposed algorithm achieved high confidence level in identifying backchanneling opportunities, and the result from a user pilot study presents that children perceive the contingent backchanneling robot as a better listener and to appear more interested in their stories. This achievement and findings form the fundamental interaction basis of this research - that children want to tell stories to a social robot companion, not because it will improve their language, but because the robot itself is an engageable, likeable, and most importantly, a reciprocal peer.

CCS Concepts

•**Computing methodologies** → **Cognitive robotics;** •**Applied computing** → **Interactive learning environments;** •**Human-centered computing** → *Collaborative interaction;*

Keywords

Social Robotics, Nonverbal Communication, Artificial Listener

1. INTRODUCTION

Our research objective is to develop a novel social robot learning companion that can successfully foster the development of early language skills of preschoolers over long-term interaction in an educational storytelling context. Storytelling, or any other form of conversational activity, is a mutual act between a speaker and a listener. The two parties establish a mental model of each other's states based on their actions, and the model is continuously updated throughout the interaction. For instance, a storyteller can predict the engagement state of the listener through the listener's gaze locking, nodding, or a leaning forward behavior. For a robot companion to become a reciprocal peer, it not only needs to understand these social signals and create a mental model of the child, but it also needs a method to share its mental states at the right timing using the same signals.

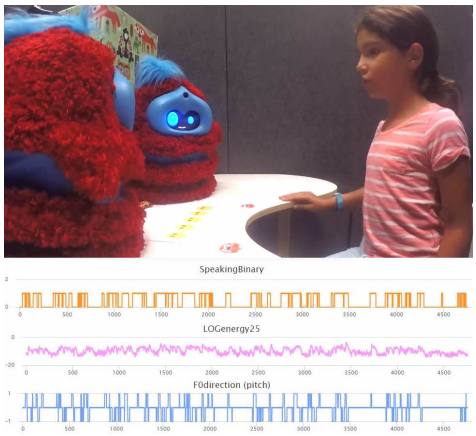


Figure 1: XXXXXXXXXXXXXXXXX

In this paper, we provide an extensive analysis of preschool children's nonverbal backchannel (BC) behavior and an approach to developing a BC model for a social robot companion to provide reciprocal behavior during a child's storytelling. From a corpus of children's dyad interaction telling and listening to a story, we analyzed speaker cues that prompt listener responses - a synergy between speaking binary, pause, pitch change, and energy, to which a BC response would have the most effect for the storyteller to stay engaged. We hypothesized that children would be more engaged with the contingent robot listener while telling a story, and attempted to validate the following hypotheses: *H1: Children will gaze more towards the contingent robot. H2: Children will direct storytelling more towards the contingent robot. H3: Children will perceive the contingent robot as more attentive and interested in their story.* We validated our hypotheses through a user study in which children told a story to two robots, one using the proposed BC model to produce contingent behavior and the other with a non-contingent behavior.

In the following sections, we present a quantitative analysis on preschool children's BC response behavior. Though there is an extensive research on human's listener response

*Video Link: `https://goo.gl/HMEfnt`

HRI '17 Vienna, Austria

DOI: 10.475/123_4

behavior based on adult's and adolescent's data, even including cultural differences [40], research on young children's BC behavior is seldom studied (Section 2). As a first step to developing an attentive robot listener for children, in Section 3 we present analyses of children's nonverbal behavior that indicates the engagement state of the listener (CB-1), speaker cues a child listener responds to (CB-2), and an analysis of mirror assumption (CB-3). We use the result of analysis CB-1 to develop a robot listener's actions and CB-2 to develop a BC opportunity prediction (BOP) algorithm tailored to child acoustic features (Section 4). In Sections 5 and 6, the experimental system architecture and study setups are introduced. Results and discussions of the study are presented in Section 7, and Section 8 concludes the paper.

2. RELATED WORK

2.1 Backchannel behavior

Backchanneling (BC) 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 [9, 8]. BC has been studied as a form of feedback and acknowledgment in both the psychology field as well as in human-robot interaction. In studies with adults [9, 8], 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, BC's main function is in establishing common ground by signaling that the receiver has understood the message. Moreover, in [28], it is demonstrated how BC feedback provides the teacher to understand the robot learner's states while interacting with a child.

There is a lack in the literature on children backchanneling, but the following two paper provide great insights on the outcomes of BC on pre-schoolers. In [29] a study is reported on the positive effects of BC behaviors on children's language learning. Twenty preschoolers (mean age 3.7) were randomly assigned to an intervention or a control group. Mothers of intervention children were told to encourage lenghty child's narratives through back-channel responses. Children in the intervention group showed significant vocabulary improvement immediately after session terminated, and a year later they showed overall improvements in narrative skills. In particular, they produced more context-setting descriptions about where and especially when the described events took place. Furthermore in[33] three studies of the 4-year-old's ability to adjust to a listener are reported. Tape recordings were made of conversations with a 2-year-old, peers, and adults. Results revealed that the more attentive is the listener, the greater the tendency for the 4-year-old to create more complex utterances and to make efforts to attract and sustain attention while the same listener is momentarily distracted. Other studies ([5, 13, 3]) highlight that the BC decontextualized language and non-verbal feedbacks have been emphasized as important for language acquisition and to motivate children in the Autism Spectrum Disorder (ASD).

Overall, the development of dialogic behaviors for adults has gained increased attention ([6, 2]), yet the application to long-term interaction with young preschoolers, as well as the construction of a fully autonomous, cognitive, expressive and responsive social peers for communication and language improvement has not been achieved. In this work, we mainly focus on the use of nonverbal BC behaviors including gestures such as gaze locking and nodding, and non-linguistic utterances such as *yeah, ok, uh huh, mhmm.*

2.2 Computational models of BC

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 [25], and this has motivated diverse approaches that provoke BC using different features that are available in real-time (e.g. VAD, energy, pitch).

Ward and Tsukahara in [38] suggest that an important prosodic cue involved, in both English and Japanese, is a region of low pitch late in an utterance. With the best prediction from "utterance end and pitch region" accounting for an accuracy of 19% (for an English corpus), the authors conclude that the current rule must be deeply studied and that at least other additional types of factors need investigation. Truong et al. in [30] consider features from the speaker's speech and gaze to determine the placement of BCs and found the number, timing and type of BC had a significant effect on how human-like the BC behavior was perceived and their result accounted for 40% of precision. Furthermore they propose to use keyword spotting to respond immediately to acknowledgement questions such as âĂIJyou know?âĂİ and âĂIJright?âĂİ.

Instead of using rule-base methods, the easiness to mainly rely on audio features for BC, have motivated researchers to train machine learning models to automatically and in real-time predict the timing of BCs given the speaker's discourse. Morency in [25], using sequential probabilistic models (e.g., Hidden Markov Model and Conditional Random Fields), automatically learned from a corpus and show a statistically significant improvement (41%) over rule based approaches. Other studies used decision trees [26] based on pitch and power features and Hidden Markov Models [27] where state transitions corresponded to changes in prosodic context.

2.3 Artificial Listeners and SAR

Artificial Listeners or Virtual Agents systems look and act like people and can engage in conversation and collaborative tasks in different situations ranging from healthcare decision support systems [21] to teach negotiation strategies [23], from real-time navigation support to or team-work training [31]. Recent research has also established the potential for virtual characters to build rapport with adults through simple contingent nonverbal behaviors [16, 15]. As a well-known example, the Semaine system [32] is a fully autonomous integrated real-time system combining incremental analysis of user behaviour, dialogue management, and synthesis of speaker and listener behaviour of a SAL character displayed as a virtual agent. In [2] the system uses face and head movement analysis, as well as speech analysis and speech recognition and then triggers listening behaviors from the listener, using probabilistic rules based on the co-occurrence of the same input and output behaviors in the database.

On the other hand the robotics field has pursued the goal of creating robots capable of exhibiting natural-appearing social qualities. Beyond the basic capabilities of moving and acting autonomously, Socially Assistive Robotics (SAR) has focused on the use of the robot's physical embodiment

to communicate, interact and socialize with users in a social and engaging manner rather than physical interaction solely. SAR, in the last years, have been used successfully to attract children's attention, stimulate sustainable interactions and improve communication and socialization skills of young children [5, 22]. Social robot learning companions offer unique opportunities of guided, personalized, and controlled social interaction and delivery of a desired curriculum [34, 4, 11]. 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 [35, 14, 12].

Our research objective is to develop an autonomous social robot learning companion that, through contingent backchannel feedback, can successfully foster the development of early language skills of preschoolers over long-term interaction in an educational storytelling context.

3. CHILD BACKCHANNEL BEHAVIOR

3.1 Analysis of Listener Behavior

To identify attention-related nonverbal behaviors, we examined the ability of frequency and duration of behaviors to predict whether a child is listening or 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 the frequency of leaning toward and brow raises as well as the duration of a smile hold a significant positive relationship to child listening. Both the frequency and duration of partner gazes hold a significant positive relationship to child listening. Both nods and utterances in their frequency have a positive relationship to the child listening, but their rare occurrences in this population make it difficult to evaluate as significant.

3.2 Analysis of Speaker Cues

To identify 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 generate 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 lean toward, partner gaze, nod, brow raise, and utterances as well as a continued smile 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. As shown in Table XX, the speaker cues—gaze, pitch, and wordy—in isolation have the ability to elicit a positive response from the young listeners. As expected, some of the speaker cues (energy, pause, and filled pause) taken singly 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 [18].

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 an empirically found 1.3 seconds of each other. 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 chance. As shown in Table XX, the one-sided binomial test indicated that the response rate of cue combinations of two cues: pitch+energy, gaze+pause, gaze+pitch and of three cues: gaze+pause+pitch, gaze+pitch+energy and of four cues energy+gaze+pause+pitch was higher than the expected rate of 0.5.

3.3 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 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 in 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 children listeners were found to respond in our prior analysis, 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 the child demonstrated a response to the same cue.

3.4 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 accounting for a total of 58 episodes. In a dyad session, the pair of students take turns narrating their story to the other; each turn generating a storytelling episode. Three time-synchronized cameras 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 (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).

[2] 18 participants x 9 cue contexts - 2DF = 160

4. BACKCHANNEL OPPORTUNITY PREDICTION (BOP) MODEL

In the following sections, we present our rule-based method to predicting BC opportunities developed using approximately 70% of dataset 1 and two evaluations conducted on i) approx 30% of dataset 1 ii) the entire dataset 2. Our work combines previous works in the field (Wordy [17], Long Pause [7], Pitch & Pause [39, 37, 38], Energy [36]) and extends them so to reach the best results in BC with the children's datasets given.

4.1 BOP Models

Wordy Model (Fig. 2(a))

The algorithm introduced by [17] predicts BC based on inter pausal unit IPU. IPU is a maximal sequence of words surrounded by a pause of duration W_PAUSE. A turn of duration W_SPEAK is then defined as a maximal sequence of IPUs from one speaker, such that between any two adjacent IPUs the silence in no longer than SIL. The BC opportunity is predicted when the following three conditions are met:

```
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 $BC_RATE$
```

Long Pause Model (Fig. 2(b))

The rule-based algorithm introduced by [7] computes the BC feedback upon detection of a long pause LP that is preceded by a speech $LP\_SPEAK lessthan W\_SPEAK$. W_SPEAK is composed by one or more IPUs separated by silences of SIL duration. 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. 2(c))

The rule-based algorithm introduced by [39] and enhanced by [37] predicts BC opportunity based on Pitch and Pause (P&P) features in English audio-only settings. 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. 2(d))

Based on [36], 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.
```

Table 1 shows the list of the pairs parameter-value used after we iteratively developed and tested our model with 71% of data form dataset 1 of Section 4.4. The combinations of these parameters were tested by incrementing/decrementing their values using a fitting step of 100ms and were chosen to maximize correspondence to corpus data.

Table 1: BOP Model Parameters

BOP Model	Parameter	Value
Wordy	W_PAUSE	800ms
	W_SPEAK	1500ms
Long Pause	LP_PAUSE	1700ms
	LP_SPEAK	1000ms
Pitch & Pause	P&P_SLOPE	25%
	P&P_LENGTH	300ms
	P&P_PAUSE	400ms
Energy	E_SLOPE	30%
	E_SLOPE_LENGTH	500ms
	E_PAUSE	300ms
BC_RATE		1300ms

4.2 Procedures of evaluation

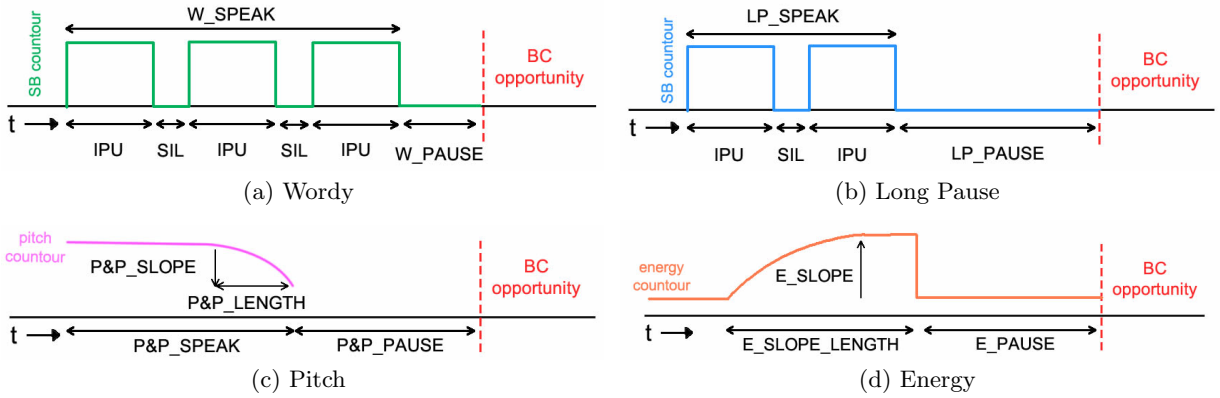
To evaluate the accuracy of the following tests we developed a Testing Framework able to:

- analyze each episode with OpenSmile and export the features F with a sampling rate of 10 ms
- 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), intersect the data of each coder for a specific storytelling episode E. The intersection A âĹl' B âĹl' C of three video-coders A, B and C is the set that contains all elements of A that belong to B and also belong to C but no other elements.
- 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 Speaking Binary classifier and our rule-based BC Model.

4.3 Speaking Binary Classifier

Our Speaking Binary Classifier has been developed on top of OpenSmile input features and adjusted by continuously iterating on 71% of the ground truth data from Dataset 1. 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 [10]. The tool outputs 0 when no one is speaking, 1 when someone is speaking. We run the Testing Framework comparing VAD from OpenSmile (OS) without any modification, and SB from ground truth (GT).



(a) Wordy (b) Long Pause

(c) Pitch (d) Energy

Figure 2: Based on the analysis of speaker cues that children respond with a backchannel behavior, the four models of backchannel opportunity detection (BOP) were developed.

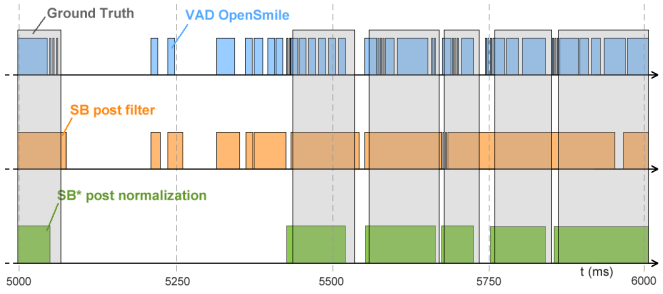


Figure 3: Comparison between VAD OpenSmile and our Speaking Binary (SB*)

- 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 ($[GT, OS] = [0, 1]$)
- 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 applied a filter and a normalization functions on top of OpenSmile features and run the Testing Framework iteratively.

We implemented a low-pass filter as that: considering the previous value of Speaking Binary at time $t-1$, $SB(t-1)$, the current value at time t, $SB(t)$, is,

$$SB(t) = 0.8 \cdot SB(t-1) + 0.2 \cdot SB(t), \quad (1)$$

then,

$$SB(t)^* = \begin{cases} 1, & \text{if } SB(t) > SB\_CUTOFF. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

explain why SB_CUTOFF is 0.9.

Afterwards, we applied an energy-based normalization function to distinguish voice signals from background noise by normalizing the energy variance through an incremental noise reduction that started after $SB\_NORMSTART = 7000ms$. Figure 3 highlights 10 seconds of a child's conversation and represents in its rows:

- VAD output from OpenSmile
- SB output from low-pass filter
- $SB*$ final output of our system with normalization function
- GT across the rows, the Speaking Binary from the Ground Truth

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

- 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 [1], that the video-coding software we used eases eight times the coding task (e.g. reducing speed, zooming) and that most of the errors are happening in imperceptible moments of speech (e.g. between IPUs as Figure3 shows in the right half), we can conclude that our system is fully reliable on detecting a speech from the human's voice.

4.4 Analysis of BC Consensus Data

Approximately 71% of the dataset was used for training the classification model, and the rest was used for testing. To generate a consensus dataset, we compared the labels of the 3 video-coders and only considered the identical labels if they were only annotated in 1000ms as [20] suggests.

During the evaluation we considered Precision (other than recall) as a 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 using the parameters shown in Section 4.1:

Table 2: Evaluation Against Consensus Dataset 1

	Mean	STD
Wordy	89.5%	7.4%
Long Pause	78.3%	8.3%
Pitch & Pause	61.1%	13.8%
Energy	67.3%	13.2%

Dataset 2: A secondary testing was conducted on a dataset collected from two local public preschools. We recruited 17 participants of typical development with an average age of 4.88 years-old (SD=0.49) and a 41:59 male-female ratio. Each child participated in eight rounds of storytelling with a robot partner except for 2 children (1 male, 1 female) who completed only the first 4 sessions accounting for a total of 128 episodes. In a dyad session, a student and a robot took turns narrating their story to the other. During the child's turn, a tele-operator made the robot's behavior as socially contingent as possible reacting to the child as closely to as a human would in the same circumstance. The robot, during its turn, told stories based on past stories told by other children. Audio and video of children's interactions with the robot were recorded with a microphone situated near the robot, a camera behind the robot facing the child and a camera behind the child facing the robot accounting for 27 hours of recordings for each device (mean: 13 minutes per session). Two video-coders, forming the ground truth for this set, were recruited to simulate themselves being a listener for each child's turn and mark the moments when they wanted to BC.

Our model was tested against a consensus level 2 ground truth (i.e. Level 2 represents the intersection A âĹl' B of two BC timings of video-coders A and B. The result is a set that contains all BC timings of A that have also been coded by B within 500ms [-250ms,+250ms] but no other BC timings. Coders were not asked to review their audio snippet and categorize the type of speaker cues perceived since we were only interested on the evaluation of the whole system. Conversely, we studied both precision and recall and calculated the F-score for each episode. Precision, in the case of matching BC timings, represents the percentage of accuracy of finding many BC timings matches (hit) and few misses (ground truth says to BC, our model says no). Recall, represents the percentage of accuracy of finding many BC timings matches (hit) and few misses (our model says to BC, ground truth says no). F-score (also called F1 score) value combines precision and recall in their harmonic mean.

Table 3: Evaluation Against Consensus Dataset 2

	Precision	Recall	F Score
AVG	84.85%	76.32%	79.31%
STD	9.35%	15.27%	10.30%
MIN	64.30%	44.40%	52.53%
MAX	100%	100%	96.80%

4.5 Discussion

5. SYSTEM

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. It uses 5 degrees of freedom to generate varying degrees of reciprocal behaviors including listener BC responses studied and discussed in Section **??**. Twenty additional backchannel response behaviors were developed on top of 18 existing nonverbal behaviors.

5.2 System Modules

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) and shown in Figure 4, includes the following elements:

- OpenSmile (os): an audio feature extraction tool enables to extract large audio feature spaces in real time [19] and to output them in a packet called *msg_os*
- SpeakingBinary (sb): 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. SpeakingBinary's messages are stored in *msg_sb*.
- BackChannel (bc): according to the SpeakingBinary output (*msg_sb*), energy (Log and RMS), pitch score and direction, the BackChannel node selects the right BC rule and randomize the execution of a BC from a predefined set of BC rule category. 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 4.1), detects BC opportunities in few milliseconds and output them in *msg_bc*
- RobotExecutor (re): receives the instructions from the BackChannel node, manages a small message queue by controlling the state of the robot (TegaState), and send the action *msg_re* 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 (ta): 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).
- TegaState (ts): running on the Tega Android phone, outputs the result in *msg_ts* of the required TegaAction sent by RobotExecutor.
- EmotionEngine (ee): using advanced facial analysis, it measures emotional responses from a video frame [24] and exports the results in *msg_ee*
- ControlPanel (cp): manages the flow of the conversation during a dialogue scenario publishing from and

subscribing to messages *msg* from other elements.

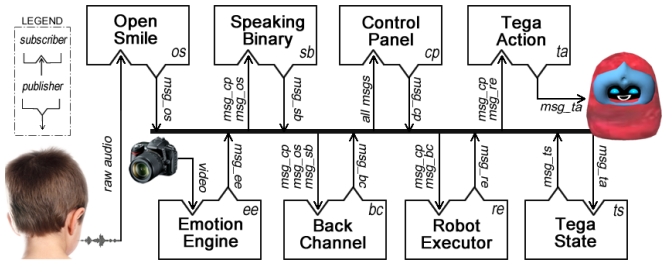


Figure 4: Architecture of the system

The system for the non-contingent robot is fairly simple: BackChannel, RobotExecutor, TegaAction work together to deliver randomize behaviors from the same contingent behavior set every 5 ± 1.5 seconds (Section **??**). A mean of 5 seconds and STD of 1.5 seconds have been calculated using the Ground Truth BC timings of Dataset 1.

5.3 Audio Feature Extraction

As stated in 5, we decided to use OpenSmile [19], an open-source standalone commandline executable and dynamic library, due to its capability of on-line incremental processing and 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)

6. EXPERIMENTAL SETUP

We hypothesized that a social robot providing contingent BC feedback would be perceived as more attentive which in turn will encourage children to attend to it more while telling a story compared to a non-contingent robot. In order to evaluate our hypothesis, the experimental room was setup as depicted in Fig. 5 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 **??**, 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.

Figure 5: Setting

6.1 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 [omitted] 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.

6.2 Conditions

6.3 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.

6.3.1 *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.

6.3.2 *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

Figure 6: Children playing with Tega Robot

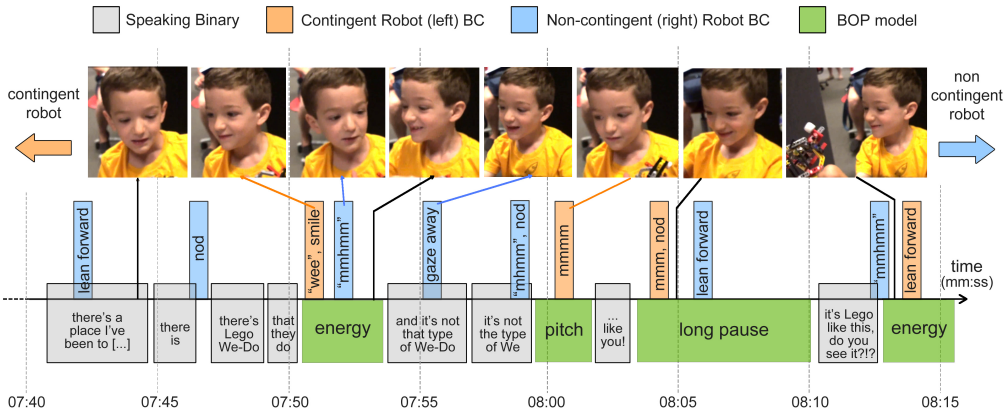


Figure 7: Children playing with Tega Robot

sessions had the contingent robot placed on the left side and 55% of the sessions on the right. The autonomous robots started backchanneling as the participant began telling stories. When the child indicated he/she was done, the robots fell back asleep.

6.3.3 *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 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.

6.4 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 4). 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.

7. RESULTS AND DISCUSSIONS

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.

Table 4: Data collected during Storytelling

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

7.1 Children gazed more at the contingent robot while storytelling

We analyzed children's gaze pattern using the yaw infor-

mation of the head orientation (Fig. **??**). 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. 8).

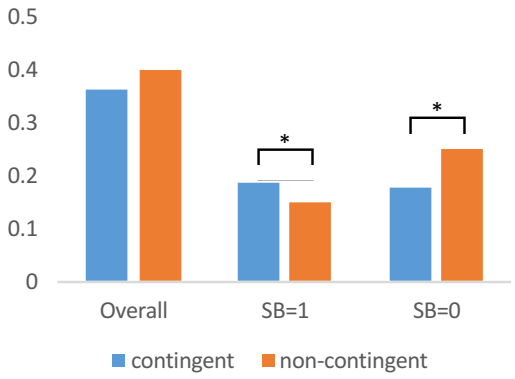


Figure 8: Average fraction of gaze length over the length of each session. Overall, not much difference is observed which robot children gazed at more. However, children gazed significantly more at the contingent robot ($p < 0.05$) when telling a story ($SB = 1$), while silent gazing behavior was more dominant ($p < 0.01$) towards the non-contingent robot. Detailed analysis reveals that a robot's non-contingent behavior tends to distract the child's attention and causes pause in storytelling.

7.2 Children were more calm towards the contingent robot.

Affective signals hold information about a person's emotional and engagement state. From facial features (Table 4), Affdex extracts 15 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. **??**).

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. 9).

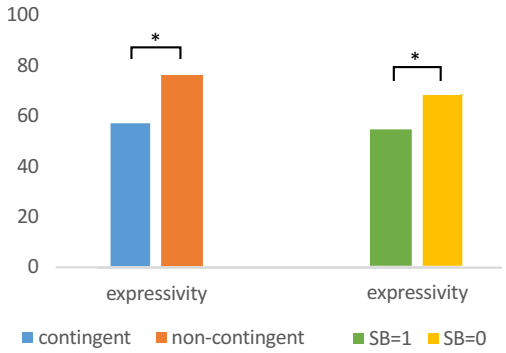


Figure 9: Expressivity in scale [0,100]. Left bar graphs show significant difference ($p < 0.01$) in children's expressivity towards the contingent and non-contingent robot. Right bar graphs show significant difference ($p < 0.05$) difference in children's expressivity when they were telling a story ($SB = 1$) to the robot and when just gazing ($SB = 0$) at the robot.

7.3 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 **??**, 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.

8. CONCLUSION

We hypothesized that the contingency, not just the frequency of positive feedback is crucial when it comes to cre-

ating rapport. The primary goal in this study was evaluative: can an agent generate BC that engenders feelings of rapport with pre-schoolers? A secondary goal was to answer the question: Is contingency (as opposed to frequency) of agent feedback crucial when it comes to creating feelings of rapport? Results suggest that contingency matters when it comes to creating rapport and that our contingent robot generated more positive behaviors under different points of view.

9. ACKNOWLEDGMENTS

[Acknowledgments omitted for blind review]

10. REFERENCES

[1] Human Benchmark.

[2] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, A. Mahdhaoui, J.-C. Martin, S. Ondas, C. Pelachaud, J. Urbain, and M. Yilmaz. Generating robot/agent backchannels during a storytelling experiment. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3749–3754. IEEE, 2009.

[3] 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*, pages 17–26. ACM, 2014.

[4] 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*, 1(2):33–53, 2012.

[5] 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*, pages 351–354. ACM, 2015.

[6] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 708–713. IEEE, 2005.

[7] 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.

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

[9] 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.

[10] 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.

[11] N. A. Freed. *" This is the fluffy robot that only speaks french": language use between preschoolers, their families, and a social robot while sharing virtual toys.* PhD thesis, Massachusetts Institute of Technology, 2012.

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

[13] 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*, page 69, 2016.

[14] 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.

[15] J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler. Creating interactive virtual humans: Some assembly required. Technical report, DTIC Document, 2002.

[16] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*, pages 125–138. Springer, 2007.

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

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

[19] 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.

[20] 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.

[21] P. Kenny, T. Parsons, J. Gratch, and A. Rizzo. Virtual humans for assisted health care. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, page 6. ACM, 2008.

[22] J. J. M. Kory. *Storytelling with robots: Effects of robot language level on children's language learning.* PhD thesis, Massachusetts Institute of Technology, 2014.

[23] R. Krovi, A. C. Graesser, and W. E. Pracht. Agent behaviors in virtual negotiation environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 29(1):15–25, 1999.

[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, pages 3723–3726. ACM, 2016.

[25] 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.

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

[27] 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.

[28] 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)*, pages 2713–2719. IEEE, 2014.

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

[30] 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.

[31] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

[32] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Ter Maat, G. McKeown, S. Pammi, M. Pantic, et al. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183, 2012.

[33] 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.

[34] 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*, pages 924–929. IEEE, 2014.

[35] F. Tanaka, A. Cicourel, and J. R. Movellan. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, 104(46):17954–17958, 2007.

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

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

[38] 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.

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

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