frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 31 October 2017
doi: 10.3389/frobt.2017.00047

Role of Speaker Cues in Attention Inference

Jin Joo Lee[1], Cynthia Breazeal[1] and David DeSteno[2]*

[1] Personal Robots Group, Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States,
[2] Social Emotions Group, Department of Psychology, Northeastern University, Boston, MA, United States

Current state-of-the-art approaches to emotion recognition primarily focus on modeling the nonverbal expressions of the sole individual without reference to contextual elements such as the co-presence of the partner. In this paper, we demonstrate that the accurate inference of listeners' social-emotional state of attention depends on accounting for the nonverbal behaviors of their storytelling partner, namely their speaker cues. To gain a deeper understanding of the role of speaker cues in attention inference, we conduct investigations into real-world interactions of children (5–6 years old) storytelling with their peers. Through in-depth analysis of human–human interaction data, we first identify nonverbal speaker cues (i.e., backchannel-inviting cues) and listener responses (i.e., backchannel feedback). We then demonstrate how speaker cues can modify the interpretation of attention-related backchannels as well as serve as a means to regulate the responsiveness of listeners. We discuss the design implications of our findings toward our primary goal of developing attention recognition models for storytelling robots, and we argue that social robots can proactively use speaker cues to form more accurate inferences about the attentive state of their human partners.

Keywords: attention and engagement, nonverbal behaviors, speaker cues, listener backchannels, emotion recognition, children and storytelling, human-robot interaction

OPEN ACCESS

Edited by:
Hatice Gunes, University of Cambridge, United Kingdom

Reviewed by:
Gerard Bailly, UMR5216 Grenoble Images Parole Signal Automatique (GIPSA-lab), France
Amit Kumar Pandey, Aldebaran Robotics, France

****Correspondence:***
Jin Joo Lee
jinjoo@media.mit.edu

Specialty section:
This article was submitted to Humanoid Robotics, a section of the journal Frontiers in Robotics and AI

Received: *14 January 2017*
Accepted: *05 September 2017*
Published: *31 October 2017*

Citation:
Lee JJ, Breazeal C and DeSteno D (2017) Role of Speaker Cues in Attention Inference. Front. Robot. AI 4:47. doi: 10.3389/frobt.2017.00047

1. INTRODUCTION

Storytelling is an interaction form that is mutually regulated between storytellers and listeners where a key dynamic is the back-and-forth process of speaker cues and listener responses. Speaker cues, also called backchannel-*inviting* cues, are signaled nonverbally through changes in prosody, gaze patterns, and other behaviors. They serve as a mechanism for storytellers to elicit feedback from listeners (Ward and Tsukahara, 2000). Listeners contingently respond using backchannel feedback which is signaled linguistically (e.g., "I see"), para-linguistically (e.g., "mm-hmm"), and nonverbally (e.g., head nod).

To support human-robot interactions (HRI), prior approaches have typically treated speaker cues as timing mechanisms to predict upcoming backchannel opportunities. In contingently responding to a person's speaker cues, robot listeners are able to support more fluid interactions, engender feelings of rapport, and communicate attention (Gratch et al., 2007; Morency et al., 2010; Park et al., 2017). In this paper, we introduce additional functions speaker cues have in social interactions beyond this stimulus-response contingency. Our main contribution is demonstrating how:

1. speaker cues serve as a means to *regulate* the responsiveness of listeners.
2. speaker cues can modify the *interpretation* of backchannels when inferring listener's attention.

For our first claim, we begin by identifying backchannels that signal the attention and engagement of listeners as well as speaker cues capable of eliciting those backchannels. We examine multimodal speaker cues (prosody and gaze) and their emission as either singlets or combinations, and we find that compounded cues have a higher likelihood of eliciting a response from listeners.

We support our second claim through a two-part process. First, our video-based human-subjects experiment demonstrates that accurate inference about listeners' attentive state depends on observing not just the listeners but also their storytelling partner. Second, through a finer-grain analysis, we find that the interpretation of backchannels from a listener depends on the storyteller's cueing behaviors. This cue-response pair is necessary for an accurate understanding of listener's attention.

Our primary research goal is to develop contextually aware attention recognition models for social robots in storytelling applications. In this paper, we focus on the nonverbal behaviors of storytellers as key context in which we evaluate the attentive state of listeners. A storyteller's speaker cues play an important role in the attention inference about listeners. This social and interpersonal context to attention, or more broadly emotion, recognition is especially relevant for human–robot interactions. HRI researchers depend on emotion recognition technologies to better understand user experience. But a common approach in affective computing is to model only the expressions of the sole individual without reference to external context like the co-presence of a social agent. In using these technologies for storytelling robots, we miss out on the added value their cueing actions can bring to the inference process. In pursuit of our research goal, this paper's approach is to first deeply understand the interpersonal nature of attention inference from the human perspective. Based on our findings from human–human interaction studies, we extract design implications when developing attention recognition models for social robots.

Our paper is outlined as follows:

- **Section 2: Background:** We elaborate on how current emotion recognition technologies disagree with modern theories of human nonverbal communication. We review speaker cues and listener backchannels that have been studied among adult populations and highlight the limited findings surrounding young children in peer-to-peer interactions.
- **Section 3: Effect of Storyteller Context on Inferences about Listeners:** Through a video-based human-subjects experiment, we manipulate the presence, absence, or falseness of storytellers from original interactions with listeners. Although the listeners' nonverbal behaviors remain exactly the same, perceptions about their attentive state from a third-party observer are different across these contextual manipulations.
- **Section 4: Effect of Speaker Cues on Listener Response Interpretation and Regulation:** Through a data collection of peer-to-peer storytelling, we identify attention-related listener responses as well as speaker cues that children use amongst peers. We examine which speaker cues, taken singly or in combination, can elicit a contingent response from listeners, and we find that listeners are more likely to respond to stronger cueing contexts. Lastly, using a logistic regression model, we find that backchannels are interpreted differently if observed after a weak, moderate, or strong cue.
- **Section 5: General Discussion:** We summarize our findings based on our human–human interaction studies and draw implications when modeling attention recognition for HRI.

2. BACKGROUND

2.1. Context in Emotion Recognition—Humans vs Machines

Emotion recognition systems typically discretize emotional states as a basic set of anger, surprise, happiness, disgust, sadness, and fear, while states such as boredom, confusion, frustration, engagement, and curiosity are considered to be non-basic (D'Mello and Kory, 2015). In our work, we focus on the social-emotional state of engagement which we interchangeably use with the word attention. Note, this should not be confused with joint attention, which is a different research problem of inferring what people are attending to in a physical environment (Scassellati, 1999). The nonverbal behaviors that support joint attention serve more as a mechanism to attend to objects and events rather than ones associated with communicating emotional states.

Emotion recognition systems have primarily focused on detecting prototypical facial expressions through facial muscle action units (FACS) (Sariyanidi et al., 2015). Based on a recent survey, facial expressions are still the main modality used for affect detection but have also extended to include gaze behaviors, body movements, voice features, spoken language, and biosignals such as electrodermal activity (D'Mello and Kory, 2015). Of the 90 systems reported, 93% of approaches focus on these within-person features and exclude extrinsic factors such as the environment or interaction partners.

This representation follows a classical theory in human nonverbal communication of nonverbal leakage where emotional states are direct influencers of exhibited nonverbal behaviors (Knapp and Hall, 2010). Traditional emotion understanding models such as those utilized by Ekman (1984) focus on the nonverbal expressions of single individuals without reference to any contextual elements such as setting, cultural orientation, or other people. By contrast, modern theories emphasize the contextual nature of nonverbal inference where greater accuracy comes from decoding expressions with reference to the social context (Barrett et al., 2011; Hassin et al., 2013).

Toward this, a growing amount of work has started to model the behaviors of both interactants to recognize social-emotional states, such as trust (Lee et al., 2013), rapport (Yu et al., 2013), and bonding (Jaques et al., 2016). Although the behaviors of both interactants are now being considered, they are fundamentally represented as a pair of independent events or captured as joint or dyadic features (like the number of conversational turns) for non-temporal models. As such, these approaches do not consider the added information that comes from the interpersonal call–response dynamic of social interactions. Although this is a foundation when modeling other domains such as turn-taking

(Thórisson, 2002) or conversational structure (Otsuka et al., 2007), emotion recognition models for dyadic interactions currently do not consider the causal properties between the behaviors of dyads and how they can influence each other.

2.2. Speaker Cues and Listener Responses—Children vs Adults

A well-known dynamic in face-to-face communication is the call–response contingencies between speaker cues and listener backchannels, which we will also refer to more broadly as "listener responses." The role of listener responses in conversations have been comprehensively characterized as carrying different functions such as signaling understanding, support, empathy, and agreement (Maynard, 1997) as well as facilitating conversational flow (Dittmann, 1972; Duncan and Fiske, 1977). However, in this paper, we specifically focus on the role of backchannels as evidence of continued attention, interest, and engagement of listeners (Kendon, 1967; Schegloff, 1982). It is important to note that we will consistently use the words *cues* and *responses* to differentiate the source of the emitted nonverbal behavior as either from a speaker or listener, respectively.

Although there is extensive research on adult listening and speaking behaviors, limited prior work exists in investigating younger populations especially in the context of peer-to-peer storytelling. In adult–child conversations, prior works have focused on demonstrating the effect of age on the backchanneling behaviors of children. More specifically, 11-year-olds were found to provide significantly more listener responses to adults than 7- or 9-year-olds and with a threefold increase between 7-year-olds and 11-year-olds (Hess and Johnston, 1988). In a separate study investigating 2- to 5-year-olds, older preschool children were found to use more head nods and spent more time smiling and gazing at adult speakers, suggesting that older children better understand a listener's role in providing collaborative feedback (Miller et al., 1985).

Both children and adult listeners were found to respond more frequently to joint cues (e.g., co-occurring speaker cues like simultaneous eye-contact with long speech pauses) over single cues. Joint cues were found to quadratically increase the likelihood of eliciting a backchannel response (Hess and Johnston, 1988; Gravano and Hirschberg, 2009). For an organized collection of prior research into speaker cues and listener responses of adults and children, see Tables S1 and S2 in the Supplementary Materials. We extend these prior works by pioneering the identification of attention-related listener responses and speaker cues that children employ amongst peers (not with adults) in storytelling interactions.

3. EFFECT OF STORYTELLER CONTEXT ON INFERENCES ABOUT LISTENERS

3.1. Overview

Although modern theories of human nonverbal communication emphasize the contextual nature of emotion understanding, current state-of-the-art approaches to emotion recognition primarily focus on the sole individual without reference to contextual elements such as the co-presence of interaction partners. The goal of this section is to demonstrate how a similar expectation placed on human observers results in them forming less accurate inferences about the emotions of others. Through a video-based experiment, we manipulate the presence, absence, or falseness of storytellers from original interactions with listeners. Although the listeners' behaviors remain exactly the same, we expect that the perception about their attentive state from third-party observers will be different across these conditions. We hypothesize the following:

Main Hypothesis: Inference performance about a listener's attentive state is best when observing both the storyteller's and listener's behaviors of a social interaction and worst when missing the storyteller context.

We quantify inference performance as a function of prediction speed and accuracy and aim to demonstrate that both measures improve when observing the true storyteller context to the listener's behaviors. We argue that accurate inference about listeners' attention depends on also observing the storyteller.

3.2. Method

Through a video-based human-subjects experiment, we study how the perception of listeners changes when observing their original behaviors in different storyteller contexts.

3.2.1. Participants

Participants were recruited online through Amazon Mechanical Turk. Turk Workers were from the United States to ensure cultural relevance. To limit the participation pool to high-quality workers, their qualification requirements met the following:

- Number of approved HITs (Human Intelligence Tasks) greater than 5000,
- Approval rating from former requesters greater than 98%.

From the 542 Turk workers that submitted to the HIT task, 36 individuals were rejected for not fully completing all parts of the task or for not properly following the task's instructions. The average age of the remaining 506 participants was 38-years-old (SD = 11). Nearly half (56%) were parents and gender was close to balanced (53% female). Below we detail two exclusion principles applied in removing participants from our analysis.

3.2.2. Study Procedure

The online survey-based experiment took an average 19 minutes (SD = 12) to complete the following three parts: Affect Recognition Assessment, Training Exercise, and Inference Task.

3.2.2.1. Affect Recognition Assessment

The Diagnostic Analysis of Nonverbal Behavior (DANVA2) is an assessment to measure an individual's nonverbal affect recognition ability (Nowicki and Duke, 1994). The evaluation consists of viewing a series of facial expressions as well as listening to paralinguistic expressions of children to identify the expressed emotion: happiness, sadness, anger, or fear. Individuals are scored based on the number of items incorrectly identified from 24

different pictures of children's faces and 24 different recordings of children's voices. Participants took this assessment through a web-based flash program that would present the stimuli and record their multiple choice response.

Overall, participants scored a mean error of 2.9 (SD = 2.0) in recognizing children's facial expression and 4.8 (SD = 2.6) in recognizing children's paralinguistic expressions. To ensure that our population consisted of individuals of average affect recognition ability, 23 participants that scored an error greater than two standard deviations from the population's average on either the DANVA face or voice subtests were excluded from our analysis below.[1]

3.2.2.2. Training Exercise

To familiarize participants with the procedure of the primary task, they first experienced the task procedure on a simple example video as a training exercise. Participants were asked to carefully watch the video and immediately pause it when they heard the word "bat." Then they were instructed to report the number in the upper-left hand corner of the video, which represented the video frame corresponding to the paused scene.

Overall, participants were on average 41 frames, or 1.4 seconds, away from the exact moment of the target event (SD = 147 frames or 4.9 seconds). Participants that were within two standard deviations from the population's average response frame passed this training exercise. As a measure of task adherence to filter out low-quality Turk workers, the 22 participants who failed to meet these criteria were excluded from our analysis below.

3.2.2.3. Inference Task

Participants were asked to watch a series of videos (each around 30 seconds in duration) of different children listening to a storyteller. Participants were told that in all the videos the listener is at first paying attention to the story, but we want to know when/if the listener stops being attentive to the narrator's story. Following the same procedure introduced in the training exercise, participants reported their paused frame, which represented the moment they perceived the listener transitioning from attentiveness to inattentiveness. They also had the option of reporting if they believed that the listener was paying attention the entire time.

3.2.3. Experiment Design

From an original interaction between a listener and their storytelling partner, we manipulate the presence, absence, or falseness of the storyteller through a video-based experiment. Although the listener's behaviors remain the same, we investigate how an observer's perception about the listener's attentive state changes across the different contextualizations. As a within-subject study design, a participant viewed a video from each of the three conditions but of three different listeners in a random order. In using three different listeners, we can generalize our results to be beyond a listener-specific phenomenon. Our three conditions are defined as the following:

1. **TRUE (control):** Participants viewed the original interaction between a storyteller and listener. With access to both the storyteller's and the listener's behaviors, they made an inference about the listener's attentive state.
2. **ABSENT:** Participants only viewed the listener. They made their inference based solely on the listener's nonverbal behaviors.
3. **FALSE:** Participants viewed an unmatched interaction where the original storyteller is replaced with one from a different storytelling episode.

From three different storytelling interaction videos collected in *Section 4.2*, we created a set for the TRUE condition with the audio and video (AV) of the original storyteller, a set for the ABSENT condition with the storyteller's AV removed, and a set for the FALSE condition with the AV of a different storyteller (see **Figure 1A**). It is important to note that although the audio recordings captured both of the storyteller's and listener's voices, in general only the storyteller is speaking and the listener is quiet. To preserve the illusion that the FALSE condition was showing real interactions, we avoided moments containing any dialog-related coordination. For example, we carefully selected video snippets that did not include when a storyteller asked a direct question or was interrupted by the listener.

All the videos were composed and edited to allow a viewer to easily see the facial expressions of both the storyteller and listener. We also preserved their gaze cues by arranging the images to mimic the original interaction geometry. As shown in **Figure 1B**, we ensured that a listener's behavior between each condition remained exactly the same. Please see Videos S1–S3 in the Supplementary Materials to watch an example set of videos used for this experiment.

3.2.4. Dependent Measures

The video snippets contain a single point where the listener transitions from attentiveness to inattentiveness as illustrated in **Figure 2**. This transition point is based on the hand-annotated attention labels from trained experts (see *Section 4.2.3*). From a participant's report on where he/she believed the transition point to be, we defined two dependent measures for inference performance.

1. **Accuracy:** A response frame after the transition point is marked as correct and elsewhere as incorrect, including the option of reporting the listener as attentive for the entire time. Accuracy is a dichotomous variable, where a value of 0 means incorrect and 1 means correct.
2. **Latency:** Latency is measured as the distance between the response frame from the target frame. This difference represents the participant's delay and is only calculated for correct inferences.

In accordance with our hypotheses, we expect an increasing trend (TRUE > FALSE > ABSENT) where participants achieve their best inference performance in both accuracy and latency

[1] There is a bit of irony in using a standard contextless test to exclude participants from a study that is investigating the influence of context on affect recognition. It is possible to make an inference (of lesser accuracy) in contextless situations, but we are investigating the added value of context. This exclusion is to ensure a population of typical development.

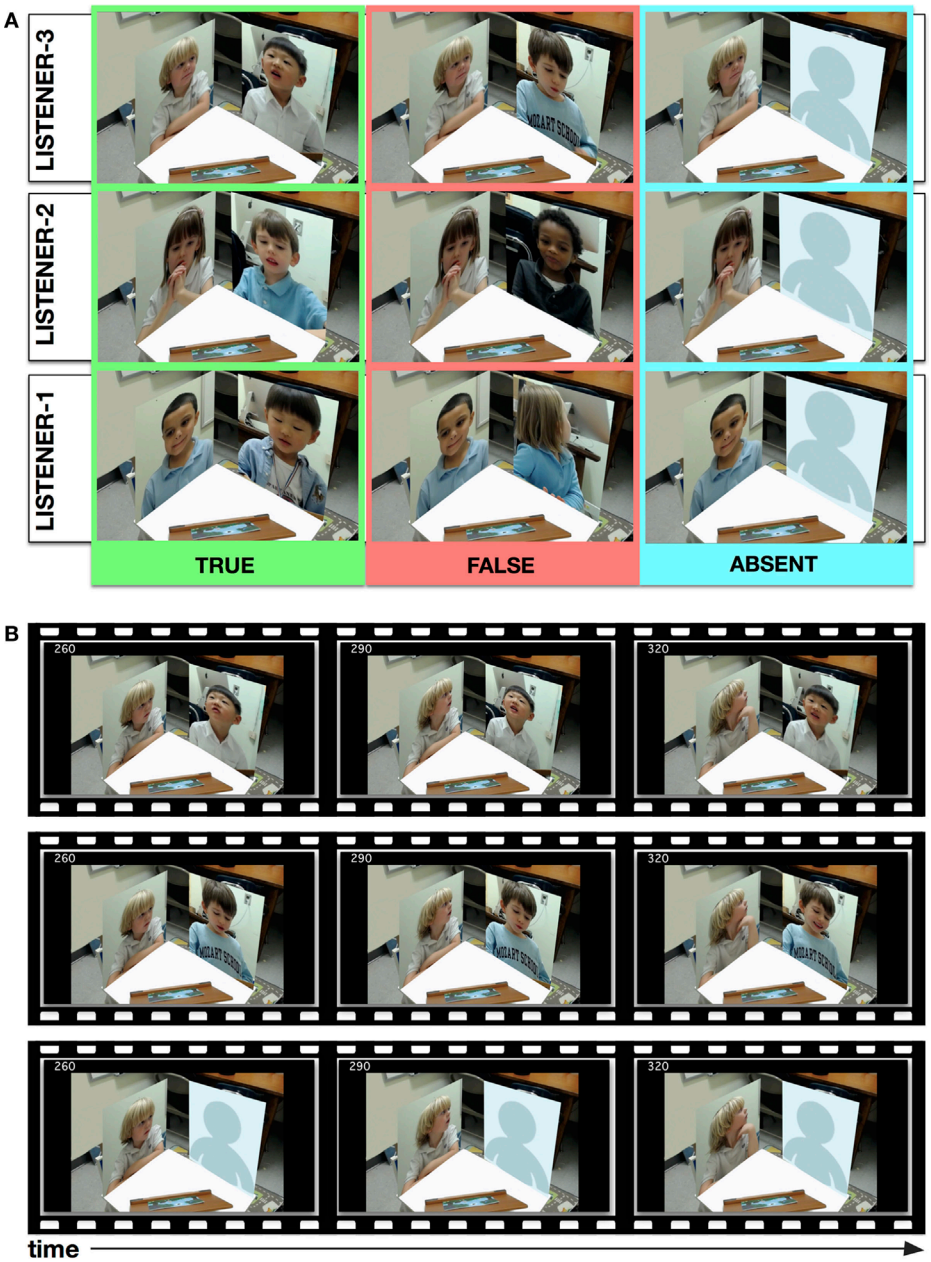

FIGURE 1 | Video-based human-subjects experiment. **(A)** From TRUE interactions between a storyteller and listener, we manipulate the absence and falseness of the storyteller context. For the FALSE condition, we replace the original storyteller with the audio and video of a different storyteller. The ABSENT condition removes all storyteller context (both audio and video). **(B)** We illustrate how for Listener-3, at frames 260, 290, and 320, we retain his exact behavior across the three conditions: TRUE (*top row*), FALSE (*middle row*), ABSENT (*bottom row*).

with the TRUE condition and their worst inference performance with the ABSENT condition since it lacks any storyteller context. We anticipate that participants will have a difficult time with the FALSE condition since the disjointed set of storyteller's cues to listener's responses will either delay or confuse their inference process. Although participants were informed, through a brief description, of the storytelling context of their upcoming videos, the FALSE condition at least visually presents the listeners' behaviors in an interpersonal context. As such, we hypothesize that a false/unmatched context is better than having no context.

3.3. Analysis of Inference Accuracy

We examine the ability of storyteller context to predict an increasing trend of inference accuracy using generalized linear models (GLM). A multilevel (i.e., mixed-model) logistic regression was performed to determine the effect of storyteller context on the likelihood of participants making a correct

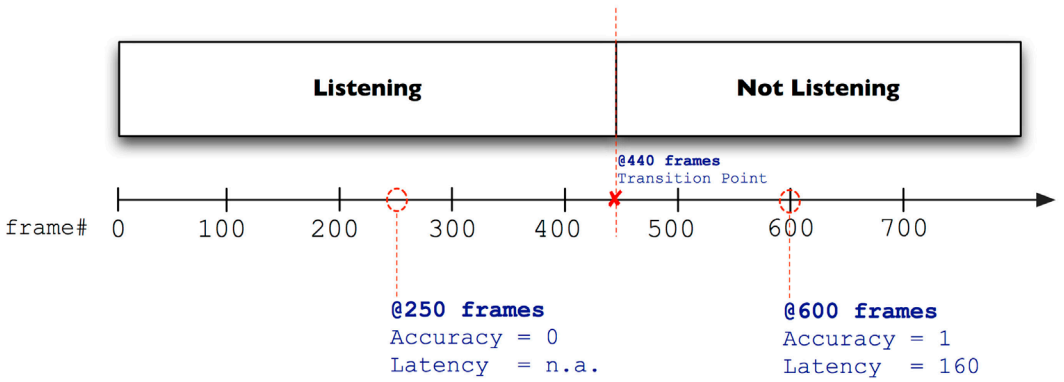

FIGURE 2 | Example of scoring accuracy and latency. At frame 440, a listener is annotated by experts as transitioning from an attentive to inattentive state. As such, a participant who predicted and reported the transition occurring at 250 frames is marked as being incorrect (accuracy of 0). A participant who made a prediction at 600 frames is accurate with a latency of 160 frames.

TABLE 1 | Effect of storyteller context on inference accuracy and latency.

	Conditions					
	True		False		Absent	
	Measures	N	Measures	N	Measures	N
Accuracy	58.4% correct	461	57.7% correct	461	51.0% correct	461
Latency	96 frames	269	107 frames	266	99 frames	235

Accuracy is reported as the percentage of participants correctly inferring attention transitions of a condition. Latency is reported as the median time-to-respond when a correct inference is made. N is the number of samples per condition (542 participants − 81 exclusions = 461 samples). Note, latency's N varies per condition since it only includes the samples with correct inferences.

inference about listeners' attentive state while controlling for within-subject dependencies from repeated measures (see Table S3 in Supplementary Material for model details). Based on our expectation that inference accuracy increases across treatment groups, the predictor variable is contrast coded as ordered values [−1, 0, 1] to model a linear trend, where having access to the true context yields the highest likelihood of accurate inference while having access to no context yields the lowest.

Based on the Wald Chi-square statistic, the logistic regression model was statistically significant, [$\chi^2(1) = 4.15$, $p = 0.04$], which indicates a linear relationship between our expected order of storyteller-context treatment and likelihood of correct inference. As shown in **Table 1**, an ascending trend in inference accuracy is observed with the TRUE condition obtaining the highest percentage of participants that correctly predict the attentive state of listeners and the ABSENT condition obtaining the lowest.

3.4. Analysis of Inference Latency

Similar to the trend analysis described for accuracy, we examine the ability of storyteller context to predict an increasing trend of inference latency values. The latency observations are positive whole numbers and have a skewed distribution since the highest density of observations are found closest to the target frame and then drop-off over time. Given the nature of the data, we used a gamma GLM (versus the typical normal distribution assumption) with storyteller context as the primary predictor of the log latency while again controlling for within-subject dependencies. We expected an increasing trend where participants experienced the greatest delays in the ABSENT condition, followed by the FALSE condition, and with the TRUE condition obtaining the lowest latencies, but no significant trend was found [$\chi^2(1) = 0.01$, $p = 0.94$].

However, rather than looking for a trend, we instead looked for any differences between the conditions. By treating the predictor as a categorical variable, a statistically significant gamma GLM was found, [$\chi^2(2) = 6.35$, $p = 0.04$][2] (see Table S4 in Supplementary Material for model details). More specifically, there is a significant difference between the TRUE and FALSE conditions, $t(767) = 2.21$, $p = 0.03$, with the TRUE condition obtaining lower latencies ($\tilde{x} = 96$ frames) than the FALSE condition ($\tilde{x} = 107$ frames). No significant difference was found between the TRUE and ABSENT conditions.

3.5. Discussion

Our main hypothesis was upheld regarding inference accuracy and partially upheld regarding inference latency. When

[2]There are two degrees-of-freedom since the three conditions are dummy coded as two categorical predictor variables.

predicting the attentive state of listeners, participants are most accurate when able to observe the true storyteller, less accurate with a false storyteller, and worst with no storyteller.

In regard to inference latency, we found that participants were faster in forming correct inferences with true storytellers over false ones. We had anticipated participants to be slowest in forming their predictions when missing the storyteller context, but they actually achieved similar speeds as when having it. If we view latency as an operationalization of confidence, we can interpret this result to mean that they felt similarly confident about their appraisals.

In sum, by changing the storyteller context in which listener behaviors are observed, we can delay or even cause incorrect inferences to be formed about the listener's attentive state. Participants are most accurate when observing *both* the storyteller's and listener's behaviors of a social interaction. They are least accurate when missing the interpersonal context of the storyteller. When presented with a false storyteller context, participants are again less accurate but also slower. This demonstrates the extent to which we can degrade an observer's perceptions about the social-emotional state of listeners.

4. EFFECT OF SPEAKER CUES ON LISTENER RESPONSE INTERPRETATION AND REGULATION

4.1. Overview

Our video-based human-subjects experiment demonstrated that the accurate interpretation of listeners' attentive state depends on also observing the storyteller. But what is it about the partner's behaviors that lead human observers to form more accurate inferences? In this section, our goal is to better understand the relationship between the storyteller's speaker cues and listener's backchannels as well as how their joint meaning impacts perceptions about listener's attention.

We conduct a series of analyses of human–human interactions. We begin by detailing our method for data collection and annotation of peer-to-peer storytelling interactions of young children in *Section 4.2: Data Collection*. Before we can start to investigate the relationship between cues and responses, we first identify the relevant nonverbal behaviors of our particular young population. As such, in *Section 4.3: Analysis of Listener Behavior*, we find backchanneling behaviors that communicate attention.

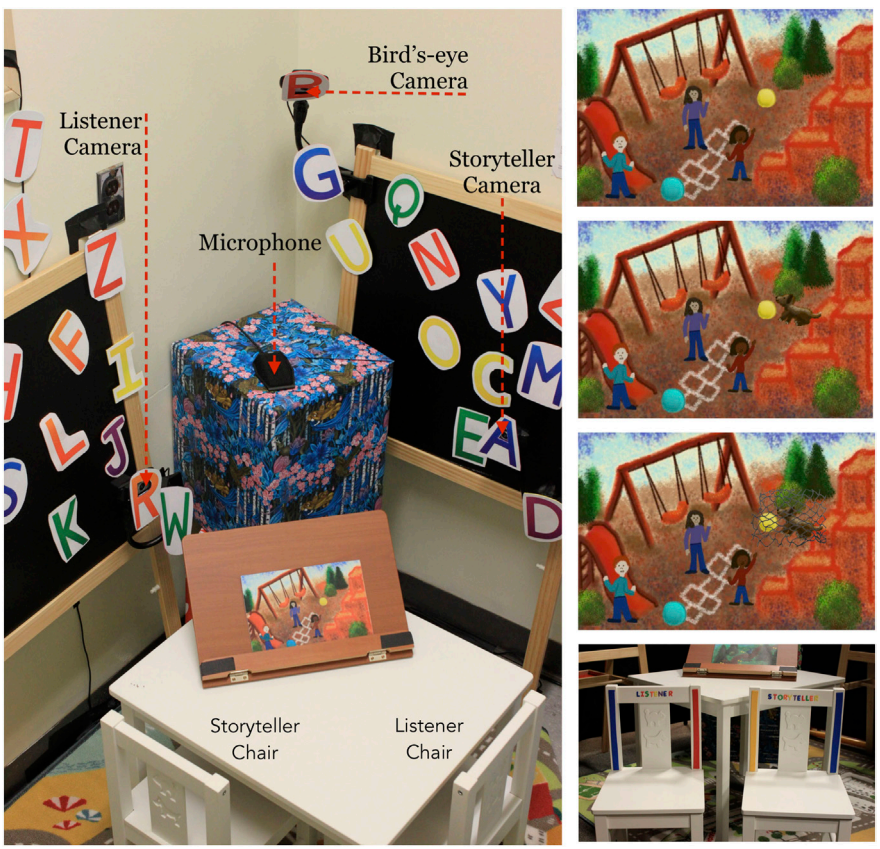

FIGURE 3 | Story space setup. Setup included three different camera angles, a high-quality microphone, listener and storyteller chairs, and a storybook with compounding story elements per page. The bottom-right photo shows how we labeled each chair to further emphasize to a child his/her role as either the storyteller or listener.

Furthermore, in *Section 4.4: Analysis of Speaker Cues*, we examine which of the coded multimodal speaker cues are observed to elicit contingent backchannels from listeners. Finally, in *Section 4.5: Analysis of Cues and Responses to Predict State*, we model the relationship between cues and responses and their effects on the perceived attentiveness of listeners.

4.2. Data Collection

4.2.1. Participants

Children of typical development were recruited from a Boston public elementary school whose curriculum already included an emphasis on storytelling. A total of 18 students from a single kindergarten (K2) classroom participated in the study. The average age was 5.22 years old (SD = 0.44) and 61% were male. Overall, 10 participants identified as White, 3 as Black or African American, 2 as Hispanic or Latino, 1 as Asian, 1 as Mixed, and 1 not specified.

4.2.2. Storytelling Task

Over a span of 5 weeks, each child completed at least three rounds of storytelling with different partners and storybooks. The storybooks were a series of colored pictures with illustrated characters and scenes that the children used to craft their own narratives (see **Figure 3** for an example storybook). In a dyad session, the pair of students took turns narrating a story to their partner with each turn generating a storytelling episode. Importantly, for each child participant, we had multiple examples of them being a storyteller and a listener. In sum, our data collection consisted of 58 storytelling episodes. The average length of a child's story was 1 minute and 17 seconds.

4.2.3. Video-Coded Annotations and Data Extraction

For each storytelling episode, the behaviors of both the listener and storyteller were manually annotated by multiple independent coders. We achieved moderate levels of agreement (Fleiss' $\kappa = 0.55$). For storytellers, we coded for gaze- and prosodic-based speaker cues. For listeners, we annotated for gaze direction, posture shifts, nods, eyebrow movement, smiles and frowns, short utterances, and perceived attentiveness. From the video recordings of the three time-synchronized cameras shown in **Figure 4**, coders used a video-annotation software called ELAN (Wittenburg et al., 2006) to mark the start and stop times for all the behaviors listed in **Table 2** except for the prosodic cues. For the attentive state annotation, a "listening" label meant that the participant was paying attention to the storyteller's story. It is important to note that our state annotation included when a listener took a "speaking-turn" as a mutually exclusive event. This enabled us to filter observations regarding conversational behaviors or turn-yielding cues, which has been demonstrated to be different from backchannel-inviting cues (Gravano and Hirschberg, 2009). Based on the "Task" annotation, we further excluded moments from our analyses when both children participants were off-task from the storytelling activity.

We developed a custom program to help coders easily annotate when and what type of prosodic cue was detected in speech. The program played back the audio recording of a storytelling episode, and coders were asked to simulate in real time being a listener and mark the moments when they wanted to backchannel by simply tapping the space bar. After this simulation, coders reviewed the audio snippets surrounding these moments to reflect on what

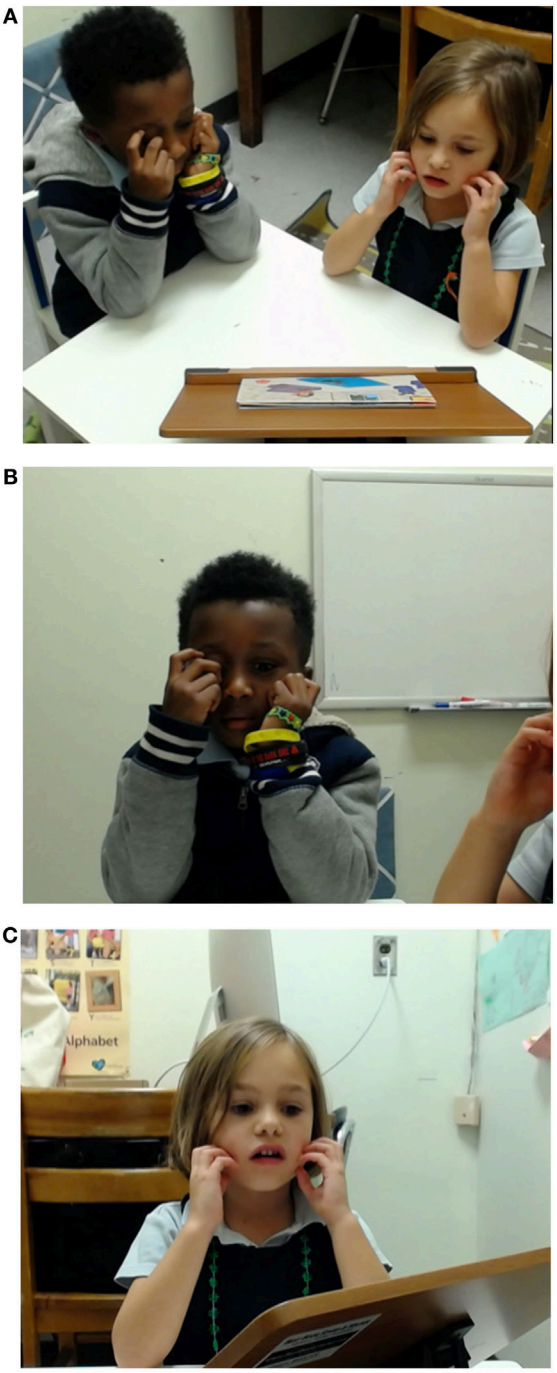

FIGURE 4 | Video-recordings of storytelling interactions. Three time-synchronized cameras captured the frontal view of each participant along with a bird's-eye-view **(A)**. Image **(B)** is the listener view. Image **(C)** is the storyteller view.

TABLE 2 | List of all annotated behaviors.

Category	Labels	S	L
Gaze	*book*, partner, away	X	X
Posture	*upright*, toward, away, other		X
Nod	*none*, nod		X
Eyebrows	*neutral*, raise, furrow		X
Mouth	*neutral*, smile, frown, other		X
Utterance	*none*, "ok," "oh," "so," "then," "yeah," "uh-huh," "ok then," "and then," "and they"		X
Voicing	*silence*, storyteller's voice, listener's voice, both	joint	
Task	*on-task*, off-task	joint	
Attentive State	listening, not listening, speaking-turn		X
Prosodic Cue	*none*, pitch, energy, pause, filled pause, long utterance, other	X	

The selected set of nonverbal behaviors were either found in prior works (see Tables S1 and S2 in Supplementary Material) or commonly observed in the storytelling interactions. Each annotation category has a set of mutually exclusive labels and was coded for storytellers (S) and/or listeners (L) or jointly evaluated (joint). An italicized label is the default behavior of an annotation category.

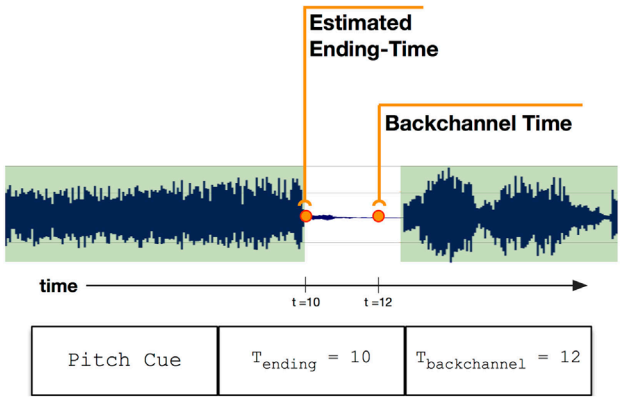

FIGURE 5 | Estimating the ending-time of a prosodic cue. Based on the backchannel time, we extract the last speaking-turn of the storyteller and estimate that its terminating edge is the ending-time of the prosodic cue. This is calculated for all prosodic-based cues except for the pause cue, which is roughly estimated to be halfway between the backchannel time and the terminating edge.

prompted their backchannel and categorize their reasoning into one or more of the following prosodic cues:

- pitch (intonation in voice, change in tone)
- energy (volume of voice, softness/loudness)
- pause (pause in speech, long silence)
- filled pause (e.g., "um," "uh," "so," "and")
- long utterance or wordy (a long contiguous speech segment)
- other

This stimulus-based coding was a method for annotators to identify *when* they wanted to backchannel (i.e., backchannel moment) in addition to categorizing their *why* (i.e., speaker cue(s) event). The null-space that was not marked had an implied default label of "none."

Three coders underwent this simulation, and we followed the Parasocial Consensus approach from Huang et al. (2010) to build consensus of when backchannel opportunities occurred. More specifically, each of our three coders' registered backchannel times were added as a "vote" on a consensus timeline with a duration of one second around the central moment. An area in the timeline with more than two total votes was counted as a valid backchannel moment.

From these backchannel moments and the voicing annotations, we estimated the emission time of prosodic cues (see **Figure 5** for more detail). To capture the complete cue context embodied by storytellers, we combined the prosodic cues with physical gaze cues to gain sets of multimodal cues.[3] In sum, for the proceeding set of analyses, we know when and which multimodal cues occurred throughout the storytelling episodes.

4.3. Analysis of Listener Behavior

A logistic regression analysis finds the best model to describe the relationship between the outcome and explanatory variables. Based on the fitted coefficients (and its significance levels), we can determine how much the explanatory variables can predict the outcome. Our goal is to identify nonverbal behaviors (explanatory variables) that can predict a listener's perceived attentive state (outcome). More specifically, for each annotated listener behavior listed in **Table 2**, a logistic regression analysis was performed to predict attention (0/1) based on the behavior's normalized duration and frequency rate. Normalized duration and frequency rates of behaviors were observed during a block period of either attentiveness or inattentiveness. Note, multiple block periods can exist in a single storytelling episode. For nonverbal behaviors that are quickly expressed (i.e., an average duration less than 90 seconds), the frequency rate was the only predictor.

Shown in **Table 3**, gazes, leans, brow-raises, smiles, nods, and utterances are nonverbal behaviors that significantly predict listeners' attention. Based on the sign of the coefficients (b) and significance (p) of the explanatory variables, we determine that frequent partner-gazes, frequent forward-leans, frequent brow-raises, prolonged smiles, frequent nods, and frequent utterances are positively associated with an attentive listener. By contrast, prolonged away-gazes from the partner, frequent away-leans, and prolonged brow-raises are negatively associated. Interestingly, brow-raises can hold opposite associations depending on their form of emission.

4.4. Analysis of Speaker Cues

In *Section 4.2.3*, adult-coders annotated when and what type of speaker cue was detected in the storytelling interactions. Therefore, the annotated speaker cues are based on adult

[3]Based on the prosodic cue ending-time and gaze onset-time, events were considered to be co-occurring and merged if they are within an empirically found 1.3 seconds of each other. This merging averages the times and reflects a collective moment of emission. When looking at the period between back-to-back gaze cues, we found a minimum time of separation of 1.5 seconds between cues. This establishes an upper bound of a merge window when trying to collect co-occurring cues. Beyond this window, we start encroaching on cues that could be a part of the next cueing instance.

TABLE 3 | Descriptive statistics and logistic regression models to estimate attention from listener behaviors.

					Logistic Regression Models		
Behavior	**Total**	**Mean Freq**	**Mean Dur**	**% Pop**	**Overall**	**Freq Term**	**Dur Term**
Gaze Partner	270	4.66	2.19	100	$\chi^2(2, 192) = 62.06$ $p^* = 3.34e^{-14}$	$b = 10.23$ $p^* = 8.25e^{-08}$	$b = 3.06$ $p = 0.22$
Gaze Away	698	12.03	4.43	100	$\chi^2(2, 192) = 152.34$ $p^* = 8.33e^{-34}$	$b = 0.10$ $p = 0.95$	$b = -8.56$ $p^* = 3.27e^{-13}$
Lean Toward	110	1.90	8.98	100	$\chi^2(2, 173) = 22.25$ $p^* = 1.48e^{-05}$	$b = 2.97$ $p^* = 7.57e^{-04}$	$b = 0.79$ $p = 0.18$
Lean Away	78	1.34	5.81	94	$\chi^2(2, 173) = 11.60$ $p^* = 3.02e^{-03}$	$b = -1.98$ $p^* = 4.57e^{-03}$	$b = 0.04$ $p = 0.96$
Brow–Raise	102	1.76	2.33	100	$\chi^2(2, 141) = 11.88$ $p^* = 2.63e^{-03}$	$b = 2.47$ $p^* = 0.01$	$b = -5.28$ $p^* = 0.02$
Brow-Furrow	17	0.29	3.23	44	$\chi^2(2, 141) = 2.06$ $p = 0.36$	$b = 1.96$ $p = 0.38$	$b = -3.43$ $p = 0.33$
Smile	173	2.98	7.23	94	$\chi^2(2, 173) = 12.35$ $p^* = 2.08e^{-03}$	$b = 0.88$ $p = 0.26$	$b = 1.69$ $p^* = 0.04$
Frown	9	0.16	2.55	28	$\chi^2(2, 173) = 1.50$ $p = 0.47$	$b = -3.83$ $p = 0.27$	$b = 2.17$ $p = 0.58$
Nod	18	0.31	1.13	39	$\chi^2(1, 34) = 7.61$ $p^* = 5.80e^{-03}$	$b = 5.28$ $p^* = 0.03$	–
Utter	18	0.31	0.94	50	$\chi^2(1, 42) = 4.24$ $p^* = 0.04$	$b = 6.72$ $p^* = 1.27e^{-03}$	–

Total is the collective frequency counts found in the dataset. The Mean Frequency is the average number of occurrences in a storytelling episode (i.e., Total/58). The Mean Duration is the average duration of an emitted behavior in seconds. % Pop refers to the proportion of the population (of the 18 participants) that demonstrated a single instance of the behavior across the repeated interactions. The logistic regression models predict the listener's attention based on the normalized duration and/or frequency rate of the nonverbal behavior. Note, the number of observations N for the chi-squared tests (i.e., χ^2 (DF, N)) are different for each annotation category since each analysis includes block periods only from storytelling episodes where at least one instance of the behavior type was observed.

perception. But which ones do children perceive, understand, and know to respond to? In our next set of analyses, we examine which speaker cues, taken singly or in combination, were observed to elicit a contingent backchannel from child listeners. We marked a backchannel as being contingent if the listener responded within [0.5−3.0] seconds[4] after the emitted cue with any of the previously found attentive behaviors. Those attentive behaviors were the onset of partner-gazes, forward-leans, brow-raises, nods, and utterances as well as prolonged smiles.

To further refine our proceeding analyses, we considered the situation where a speaker cue occurred but during a period when the listener was not paying attention to the storyteller. Their lack of a contingent response in this situation does not add relevant information to determining which cues children know to respond to. As such, our analyses only included data from moments when listeners were marked as attentive. This way, we can reason that when an attentive listener is unresponsive to a particular speaker cue, it is because the listener does not know to respond to this type of cueing signal.

4.4.1. As Individual Signals

A logistic regression analysis was performed to determine which speaker cues predict that an attentive listener will contingently backchannel. The overall logistic regression model was statistically significant [$\chi^2(6) = 45.9$, $p = 3.15e^{-08}$], and the speaker cues—gaze, pitch, filled pause, and long utterance taken singly—can elicit a response from young listeners (see **Table 4**). As expected, some of the speaker cues—energy and pause—do not offer significant predictive ability when examined in isolation. However, young children have been previously observed to respond more often in stronger cue contexts where two or more cues are co-occurring (Hess and Johnston, 1988).

4.4.2. As Co-Occurring Signals

Using the set of multimodal cues (extracted in *Section 4.2.3*), we examined the ability of cue combinations to predict that an attentive listener will contingently backchannel. The likelihood of observing a combination of cues is much smaller than individual cues, resulting in small sample sizes for each unique combination. Rather than performing a logistic regression analysis, we use the binomial exact test to determine whether the response rate of a cue combination is greater than an expected rate of 0.5. As shown in **Table 5**, the one-sided binomial test indicates that the response rates of the co-occurring cues Pitch-Energy, Gaze-Pause, Gaze-Pitch, Gaze-Pitch-Pause, and Gaze-Pitch-Energy are significantly higher than the expected rate. Interestingly, as the number of co-occurring cues increases (1 → 2→ 3), the likelihood of receiving a response also increases (0.68 → 0.82 → 0.93)[5]. Stronger the cue context, the more likely a listener will respond.

[4]We found that children positively respond on average 1.77 seconds (SD = 1.30) after an emitted cue. As such, we considered only the listener behaviors within a standard deviation from this average response time.

[5]Averaged response rates of only significantly predictive cues.

TABLE 4 | Descriptive statistics and the logistic regression model for individual speaker cues.

Logistic Regression Model						
Predictors	**Gaze**	**Pitch**	**Energy**	**Pause**	**Filled pause**	**Long utterance**
b	1.89	0.65	0.08	0.09	1.33	1.05
t-stat	5.35	2.16	0.22	0.31	2.13	2.25
p-value	$p^* = 8.67e^{-08}$	$p^* = 0.03$	$p = 0.82$	$p = 0.76$	$p^* = 0.03$	$p^* = 0.02$
N	174	147	52	122	27	17
rate	0.76	0.59	0.58	0.51	0.59	0.76

The logistic regression model predicts the likelihood of a contingent response from an attentive listener based on the emitted speaker cues. N is the collective frequency counts found in the dataset. Rate is the likelihood of a response from listeners.

TABLE 5 | Descriptive statistics and the one-sided binomial exact test for co-occurring speaker cues.

2 Cues	N	Rate	p-value	3 Cues	N	Rate	p-value
..PE..	14	0.57	p = 0.40	.CPE..	10	0.70	p = 0.17
.CP...	64	0.56	p = 0.19	GCP...	14	0.93	$p^* = 9.16e^{-04}$
..P.F.	19	0.63	p = 0.18	GC.E..	15	0.93	$p^* = 4.88e^{-04}$
.C...W	12	0.75	p = 0.07				
.C.E..	44	0.66	p* = 0.02				
G.P...	18	0.89	$p^* = 6.56e^{-04}$				
GC....	39	0.90	$p^* = 1.68e^{-07}$				

We show the most frequently observed cue combinations in our dataset. Cue combinations are specified through the presence of the cue's symbolic letter. G: Gaze, C: Pitch, P: Pause, E: Energy, F: Filled Pause, W: Wordy (for long utterances). A dot represents the absence of that cue. N is the total occurrences of the cue combination found in the dataset. The one-sided binomial exact test determines whether the response rate of a cue combination is greater than an expected rate of 0.5.

4.5. Analysis of Cues and Responses to Predict State

Our primary goal is to better understand the relationship between speaker cues and listener responses and how their joint meaning can influence perceptions about listeners' attention. To fully model how the unique combinations of cue-response pairs effect this perception, we need much more data. Given our dataset, we instead create similarity heuristics to form groups that define a smaller range of possible behavior combinations.

Based on the relationship between cue-strength and response rate from our prior analysis (*Section 4.4.2*), we categorize multimodal cues based on their number of co-occurring cues as either weak, moderate, or strong cue contexts. For example, a Gaze-Pitch-Energy multimodal cue is represented with a value of 3, or a strong cue context.

Based on our analysis in *Section 4.3*, listener response combinations are grouped based on their overall valence score. Measured as a sum of individual valences, a forward-lean (+), prolonged smile (+), and an away-gaze (−) observed from the listener within [0.5−3.0] seconds after an emitted cue is represented as a total valence value of +1, an overall weak positive response. By accounting for both positive and negative behaviors, we roughly measure the magnitude and direction of listeners' overall response during this time window.

We also recorded whether annotators marked listeners as being attentive or inattentive by the end of this time period. This captured whether the perception about a listener changed (or remained the same) after witnessing the emoted response to the cue context.

In sum, for the following analyses, we use data tuples of <Cue, Response, State>:

- **Cue:** number of co-occurring speaker cues representing strength of weak, moderate, or strong [1 to 3].
- **Response:** measure of listener response to a cue as an overall valence rating [−3 to +4].
- **State:** perception of listener's attentive state sampled immediately after the response window [0 or 1].

We first examine the ability of cues and responses as independent predictors to explain state by themselves (*Section 4.5.1*: Only Main Effects) and then compare what happens when we add an interaction term that represents the relationship between cues and responses (*Section 4.5.2*: With Interaction Effects).

4.5.1. Only Main Effects

We examined the ability of cue-strength and response-valence to predict listeners' attention. The overall logistic regression model was statistically significant, [$\chi^2(2) = 71.4$, $p = 3.2e^{-16}$], where response-valence was the primary predictor in estimating state (see **Table 6A**). One unit increase in the response-valence makes the listener 2.91 times more likely to be paying attention. This result is not surprising since listeners' behaviors are, of course, good predictors of their attentive state. But this analysis also serves as a means to validate our method of measuring listener response as an overall valence rating.

4.5.2. With Interaction Effects

In adding an interaction term to our previous logistic regression model, we find that the overall model is again statistically significant, [$\chi^2(3) = 78.4$, $p = 6.74e^{-17}$], but can explain more of the variance $R^2 = 19.5\%$ compared to $R^2 = 17.8\%$ of the previous model. As shown in **Table 6B**, the interaction term is significant ($p = 0.02$), which indicates that the predictive power of listener response is modified by the cue context.

As shown in **Figure 6**, the strong-cue curve approaches areas of higher likelihood (i.e., y-axes limits) more quickly than the other curves, especially in comparison to the weak-cue curve which has less steep tails. This means, that for the same valence of listener response, stronger cues facilitate higher levels of certainty regarding listener's attentive state.

TABLE 6 | Logistic regression models predicting attention based on cues and responses.

(A) Main Effects Model

Predictor	Cue	Response
b	0.02	1.07
t-stat	0.09	7.20
p-value	$p = 0.93$	$p^* = 5.98e^{-13}$

(B) With Interaction Effects

Predictor	Cue	Response	Cue·Response
b	0.09	0.10	0.70
t-stat	0.38	0.24	2.41
p-value	$p = 0.71$	$p = 0.81$	$p^* = 0.02$

(A) *The first logistic regression model considers only the main effects of cue-strength and response-valence to predict state.* ***(B)*** *The second model adds an interaction term which represents the relationship between cues and responses.*

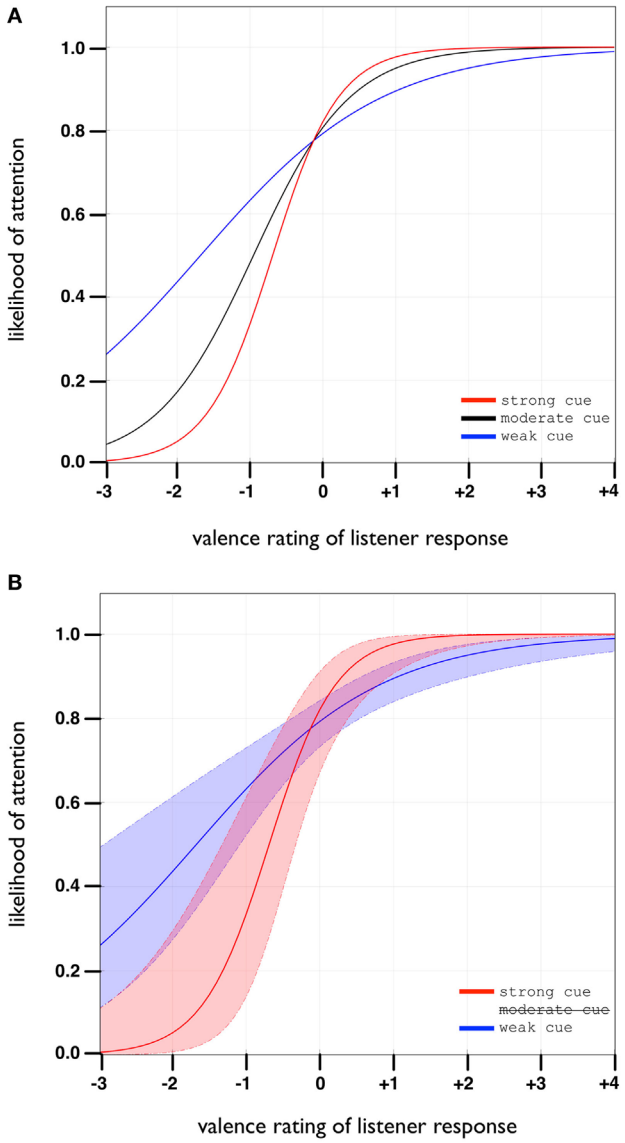

FIGURE 6 | Graphical representation of the logistic regression model from *Section 4.5.2*. **(A)** The model predicts the attentive state of listeners based on cue-strength and response-valence as well as their interaction. The x-axis represents overall listener's response as either very positive (+4) to very negative (−3). The y-axis represents the likelihood of attention, or inversely as inattention. **(B)** Shows the 95% confidence bounds of strong vs weak cue contexts. For the same valence of listener response (e.g., x = −2), there is a difference in interpretation if we observed it after a weak vs a strong cue. Strong cues buy us greater certainty that the listener is not paying attention (likelihood of 70–100% vs 50–70%).

4.6. Discussion

4.6.1. Attention-Related Backchannels of Young Listeners

We identified nonverbal behaviors that are indicative of a child either attentive or inattentive to their partner's storytelling (see summary **Table 7**). We determined the form in which these nuanced behaviors are emitted and differentiate the relevance of a behavior as either a prolonged expression or as a frequent occurrence. Of the behaviors identified, the most unexpected result was the opposing interpretations of frequent versus prolonged brow-raises. Prolonged brow-raises most often co-occurred when listeners were also looking away from storytellers (see Figure S1 in Supplementary Material for a correlation map); their joint emission can serve as a strong signal of a listener losing attention.

4.6.2. Response Rate of Multimodal Speaker Cues

By examining prosodic- and gaze- based cues, we identified multimodal speaker cues, taken singly or in combinations, that can elicit a response from listeners at different rates of success (see summary **Table 8**). Some prosodic cues such as pauses in speech or changes in energy seem to be too subtle for young children to perceive, but their cueing context can be strengthened by adding co-occurring behaviors such as a gaze cue. We confirm prior work by Hess and Johnston (1988) in demonstrating that children respond more often in stronger cue contexts. However, we differentiate our work by finding cues that young listeners know to respond to as well as employ themselves as storytellers.

4.6.3. Magnifying Certainty about Listeners' Attentive State

We found that speaker cues can modify the interpretation of backchannel responses. For the same valence and quality of listener response, there is a difference in interpretation if observed after a weak versus a strong speaker cue. We found that stronger cues buy us greater certainty that a listener is attentive or inattentive. We need both speaker cues and their associated listener responses for an accurate understanding of attention. Backchannels are more informative about the attentive state of listeners when we also know the manner in which they were elicited.

5. GENERAL DISCUSSION

Our primary contribution is introducing the role speaker cues can have in the process of attention inference. We found

TABLE 7 | Listener response summary.

Attentive Behaviors		Inattentive Behaviors	
Frequent	Partner-gazes	Prolonged	Away-gazes
Frequent	Forward-leans	Frequent	Away-leans
Frequent	Brow-raises	Prolonged	Brow-raises
Prolonged	Smiles		
Frequent	Nods		
Frequent	Utterances		

Summary of nonverbal behaviors, as prolonged expressions or frequent occurrences, that are indicative of an attentive or inattentive child listener.

TABLE 8 | Speaker cue summary.

Single Cue	rate	Dual Cues	rate	Tri Cues	rate
Pitch	0.59	Pitch-Energy	0.66	Gaze-Pitch-Energy	0.93
Filled Pause	0.59	Gaze-Pause	0.89	Gaze-Pitch-Pause	0.93
Long Utterance	0.76	Gaze-Pitch	0.90		
Gaze	0.76				

Summary of multimodal cues that children storytellers are observed to use and also can elicit a contingent response from children listeners with varying rates of success.

that speaker cues add interpretive value to attention-related backchannels and also serve as a means to regulate the responsiveness of listeners for those backchannels. Although these findings are based on human–human interaction studies, their implications are noteworthy toward our research goal of developing attention recognition models for social robots. We detail two major implications that will need further validation in an HRI context, which open promising directions for future research.

5.1. Design Implication 1: Modeling the Cueing Actions of Robots Can Increase Attention Recognition Accuracy

Since speaker cues and listener responses are both necessary for accurate attention inference, robot storytellers capable of accounting for their own nonverbal cueing behaviors in their attention models can form more accurate inferences about their human partners. Current approaches to attention recognition primarily focus on modeling the nonverbal behaviors of the sole individual, e.g., just the listener. As we saw in our video-based human-subjects experiment, this approach is akin to asking participants to form accurate inferences about listeners while removing the context of the storyteller. But, inference performance decreases when missing this interpersonal context.

Furthermore, we found that the interpretation of backchannels from listeners depends on whether it was observed after a weak, moderate, or strong speaker cue. A strong cue is a strong elicitation for a response. As such, the cue-response pair is more informative.

By including both the *robot* storyteller's cues and the *human* listener's backchannels, attention recognition models can achieve more accurate predictions especially when used in social situations.

5.2. Design Implication 2: Social Robots Can Pursue a Proactive Form of Attention Recognition in HRI

Since compounded cue contexts have a higher likelihood of eliciting a response from listeners, robot storytellers can manipulate their production of nonverbal speaker cues to deliberately gain more information. In moments of high uncertainty about the listener, a social robot can plan to emit an appropriate cue context to strongly elicit a response that can reduce state uncertainty. Through cueing actions, social robots can pursue a proactive form of inference to better understand their partner's emotional state. Toward this, an immediate extension of this work is to validate whether robot-generated speaker cues result in similar response rates from children listeners. To develop a robot capable of producing these nonverbal cues, we refer readers to our prior work in modeling prosodic-based cues through a rule-based method (Park et al., 2017).

5.3. Limitations

Admittedly, our work does not explicitly include a robot in the studies. But strong evidence exist in demonstrating the readiness of the human mind to respond to technology as social actors—capable of evoking the same social responses as they would with a human partner (Reeves and Nass, 1996; Desteno et al., 2012). We expect our finding from studying human–human interactions will carry over to human-robot interactions. However, further experimental validation is necessary to confirm the effectiveness of robot-generated speaker cues to boost attention recognition accuracies when incorporated into the model and evaluated in a human–robot interaction context.

CONCLUSION

Socially situated robots are not passive observers, but their own nonverbal behaviors contribute to the interaction context and can actively influence the inference process. We argue for a move away from the contextless approaches to emotion recognition, especially for human–robot interaction. A robot's awareness of the contextual effects of its own nonverbal behaviors has an important role in affective computing.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of MIT's Committee on the Use of Humans as Experimental Subjects (COUHES) with written informed consent from all subjects' legal guardian including the publication of subjects' photos. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by MIT's COUHES.

AUTHOR CONTRIBUTIONS

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in

the work to take responsibility for the content, design, analysis, interpretation, writing, and/or critical revision of the manuscript. More specifically, CB and JJL developed the concept of robot actions in affective computing and designed human-subjects studies with DD. JJL collected the storytelling demonstrations and performed the set of analyses under the guidance of DD. The manuscript was drafted by JJL and edited by CB and DD for important intellectual merit. JJL, CB, and DD give approval of the final version to be published.

ACKNOWLEDGMENTS

We thank Dr. Paul Harris for the crucial insights regarding children behaviors and Dr. Jesse Gray for advice on video-editing and data-processing.

FUNDING

We acknowledge the support of the National Science Foundation (grant IIS-1138986), Dolores Zohrab Liebmann Fellowship, and William Asbjornsen Albert Memorial Fellowship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/article/10.3389/frobt.2017.00047/full#supplementary-material.

VIDEO S1 | The FALSE condition for Listener-2 in **Figure 1A**.

VIDEO S2 | The ABSENT condition for Listener-2.

VIDEO S3 | The TRUE condition for Listener-2.

REFERENCES

Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Curr. Dir. Psychol. Sci.* 20, 286–290. doi:10.1177/0963721411422522

Desteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., et al. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychol. Sci.* 23, 1549–1556. doi:10.1177/0956797612448793

Dittmann, A. (1972). Developmental factors in conversational behavior. *J. Commun.* 22, 404–423. doi:10.1111/j.1460-2466.1972.tb00165.x

D'Mello, S., and Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surveys* 47, 1–36. doi:10.1145/2682899

Duncan, S., and Fiske, D. W. (1977). *Face-to-Face Interaction: Research, Methods, and Theory*. Cambridge University Press.

Ekman, P. (1984). "Expression and the nature of emotion," in *Approaches to Emotion*, eds K. Scherer and P. Ekman (Hillsdale, NJ: Lawrence Erlbaum), 319–343.

Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). "Creating Rapport with Virtual Agents," in *Proceedings of the International Conference on Intelligent Virtual Agents* (Paris, France), 125–138.

Gravano, A., and Hirschberg, J. (2009). "Backchannel-inviting cues in task-oriented dialogue," in *Proceedings of the International Conference of INTERSPEECH* (Brighton, UK), 1019–1022.

Hassin, R. R., Aviezer, H., and Bentin, S. (2013). Inherently ambiguous: facial expressions of emotions, in context. *Emot. Rev.* 5, 60–65. doi:10.1177/1754073912451331

Hess, L., and Johnston, J. (1988). Acquisition of back channel listener responses to adequate messages. *Discourse Process.* 11, 319–335. doi:10.1080/01638538809544706

Huang, L., Morency, L.-P., and Gratch, J. (2010). "Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior," in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems* (Toronto, Canada), 1265–1272.

Jaques, N., McDuff, D., Kim, Y. L., and Picard, R. W. (2016). "Understanding and predicting bonding in conversations using thin slices of facial expressions and body language," in *Proceedings of the International Conference on Intelligent Virtual Agents* (Los Angeles, CA), 64–74.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol.* 26, 22–63. doi:10.1016/0001-6918(67)90005-4

Knapp, M., and Hall, J. (2010). *Nonverbal Communication in Human Interaction*. Boston, MA: Wadsworth Publishing.

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Front. Psychol.* 4:893. doi:10.3389/fpsyg.2013.00893

Maynard, S. (1997). Analyzing interactional management in native/non-native English conversation: a case of listener response. *Int. Rev. Appl. Linguist. Lang. Teach.* 35, 37–60.

Miller, L., Lechner, R., and Rugs, D. (1985). Development of conversational responsiveness: preschoolers' use of responsive listener cues and relevant comments. *Dev. Psychol.* 21, 473–480. doi:10.1037/0012-1649.21.3.473

Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Auton. Agents Multi Agent Syst.* 20, 70–84. doi:10.1007/s10458-009-9092-y

Nowicki, S., and Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: the diagnostic analysis of nonverbal accuracy scale. *J. Nonverbal Behav.* 18, 9–35. doi:10.1007/BF02169077

Otsuka, K., Sawada, H., and Yamato, J. (2007). "Automatic inference of cross-modal nonverbal interactions in multiparty conversations," in *Proceedings of the International Conference on Multimodal Interaction* (Nagoya, Aichi, Japan), 255–262.

Park, H. W., Gelsomini, M., Lee, J. J., and Breazeal, C. (2017). "Telling stories to robots: the effect of backchanneling on a child's storytelling," in *Proceedings of the International Conference on Human-Robot Interaction* (Vienna, Austria), 100–108.

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY: Cambridge University Press.

Sariyanidi, E., Cavallaro, A., and Gunes, H. (2015). Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 1113–1133. doi:10.1109/TPAMI.2014.2366127

Scassellati, B. (1999). "Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, ed. C. L. Nehaniv (Berlin, Heidelberg, Germany: Springer), 176–195.

Schegloff, E. (1982). "Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences," in *Analyzing Discourse: Text and Talk*, ed. D. Tannen (Washington, DC: Georgetown University Press), 71–93.

Thórisson, K. R. (2002). Natural turn-taking needs no manual: computational theory and model, from perception to action. *Multimodality Lang. Speech Syst.* 19, 173–207. doi:10.1007/978-94-017-2367-1_8

Ward, N., and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *J. Pragmat.* 32, 1177–1207. doi:10.1016/S0378-2166(99)00109-5

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). "ELAN: a professional framework for multimodality research," in *Proceedings of the International Conference on Language Resources and Evaluation* (Genoa, Italy), 1556–1559.

Yu, Z., Gerritsen, D., Ogan, A., Black, A. W., and Cassell, J. (2013). "Automatic prediction of friendship via multi-model dyadic features," in *Proceedings of the SIGdial Meeting on Discourse and Dialogue* (Metz, France), 51–60.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.