Robot Nonverbal Behavior Improves Task Performance In Difficult Collaborations

Henny Admoni
Dept of Computer Science
Yale University
New Haven, CT 06520 USA
henny@cs.yale.edu

Thomas Weng
Microsoft Corporation
Redmond, WA 98052 USA
thomasweng@aya.yale.edu

Bradley Hayes
Dept of Computer Science
Yale University
New Haven, CT 06520 USA
bradley.h.hayes@yale.edu

Brian Scassellati
Dept of Computer Science
Yale University
New Haven, CT 06520 USA
scaz@cs.yale.edu

***Abstract*—Nonverbal behaviors increase task efficiency and improve collaboration between people and robots. In this paper, we introduce a model for generating nonverbal behavior and investigate whether the usefulness of nonverbal behaviors changes based on task difficulty. First, we detail a robot behavior model that accounts for top-down and bottom-up features of the scene when deciding when and how to perform deictic references (looking or pointing). Then, we analyze how a robot's deictic nonverbal behavior affects people's performance on a memorization task under differing difficulty levels. We manipulate difficulty in two ways: by adding steps to memorize, and by introducing an interruption. We find that when the task is easy, the robot's nonverbal behavior has little influence over recall and task completion. However, when the task is challenging—because the memorization load is high or because the task is interrupted—a robot's nonverbal behaviors mitigate the negative effects of these challenges, leading to higher recall accuracy and lower completion times. In short, nonverbal behavior may be even more valuable for difficult collaborations than for easy ones.**

I. Introduction

People use nonverbal behaviors (NVBs) to augment spoken references, clarify ambiguous language, and convey attention, among many other functions [1], [2], [3]. Joint activity, which involves coordinating action among two partners, requires NVBs that direct attention to particular objects or regions of space [4]. These actions can take the form of pointing (i.e., *deictic*) gestures, which can be enacted with the hand, the head, or other body parts [2], [3], [4]. In this paper, we focus on two specific deictic NVBs: pointing with the hand and looking with the head.

Robots can take advantage of deictic NVBs to improve human-robot collaborations. For example, imagine a robot assistant on a factory floor that is training an employee to construct an assembly out of component parts. The robot can look and point to the parts as it refers to them in order to clarify the references. This is especially important when there are multiple parts that can be described the same way, but need to be placed in a particular order, for example, a left and right version of the same bracket piece. Instead of saying "the left bracket piece," the robot can say "that bracket piece" and use pointing to disambiguate the reference to "bracket piece."

Human-robot interaction research has shown that people can benefit from this kind of deictic NVB from robots. Pointing and gaze from robots during object references allows people to more quickly locate objects and to disambiguate object references, increasing the efficiency of the collaborations [5], [6], [7]. People also have more positive evaluations of a robot when it uses gestures along with speech [8].

Fig. 1. This paper investigates how deictic nonverbal behavior from a robot mediates a task's difficulty in terms of information recall and time to task completion in a human-robot collaboration.

In this paper, we present a computational model for generating robot NVBs, and then use a real-time implementation of this model in a human-robot interaction study (Figure 1). We show that our model generates helpful NVBs that improve people's understanding of a robot's communication, and we reveal new insights into the use of NVB in collaborative HRI.

Generating NVBs for robots is not trivial. A naïve NVB controller might always select all possible nonverbal behaviors, looking and pointing at every possible reference. But there is a benefit to being selective about generating NVBs. Frequent nonverbal behavior is undesirable when it engages effectors that the robot might otherwise need, such as hands for object manipulation and head for vision. Additionally, in human collaborations, people use nonverbal behaviors as subtle, implicit mechanisms of communication [9], so excessive NVBs may be visually or cognitively distracting to a viewer. For robots powered by batteries, energy expenditure from moving effectors to perform NVBs might also be a concern.

For these reasons, our behavior model is selective about when to generate NVBs. The model considers elements of the scene and the task to select the most communicative and least expensive NVBs for the particular reference and environment

978-1-4673-8370-7/16/$31.00 © 2016 IEEE

at hand. This model is described in Section III.

Next, we use our novel model to generate deictic NVBs for a human-robot collaboration that investigates the effectiveness of NVBs under different task difficulties. In particular, we explore whether the difficulty of a task affects how well a robot's deictic NVBs serve to communicate spatial references. To answer this research question, people are asked to complete a memorization task based on instructions provided by a humanoid robot. We manipulate task difficulty in two ways: by increasing the number of steps people need to memorize, and by introducing an interruption that distracts people momentarily from their task. We hypothesize that:

- H1 Using nonverbal behaviors while providing spatial task instructions will improve recall accuracy and reduce task completion times,
- H2 When the task difficulty increases, the effect of nonverbal behaviors will increase, and
- H3 A robot that displays nonverbal behaviors will be rated more positively than a robot that only uses speech for communication.

Spatial collaboration, like the task employed in this study, involves manipulating and moving objects in the environment. Because the position of these objects is not restricted, the model cannot simply pre-script the NVBs for each object. Instead, our real-time robot behavior model continually calculates the best NVB for each object reference as the objects in the environment are manipulated.

Section IV details the implementation of the model and the experiment. Section V describes the results of the study, and Section VI discusses these results and future research.

II. Related Work

People use deictic gestures like pointing to focus attention on a target spatial region [10]. As pointing becomes more precise (because the pointing targets are closer), people rely more on pointing and less on language for references [10]. Deictic gestures are especially useful in communicating how to assemble objects [11], which is the task we have selected for the present study.

Human-robot interaction (HRI) research has shown the benefit of deictic gestures in human-robot collaboration. Implicit nonverbal communication increases the efficiency of task performance and reduces the impact of errors from miscommunication [12]. When robots are providing instructions or referencing objects, people use robots' deictic gestures to improve their task speed and efficiency [5], [6], [7]. Robots can even use deictic gaze to subtly influence people's selections of objects without those people realizing it [13]. Robots that gesture along with their speech elicit more attention and better recall [14], as well as higher ratings [8], than robots that do not show co-verbal gestures. Cooperative gestures are most effective when they are presented frontally and with machine-like "abrupt" motion [15]. Multiple deictic gestures may be better than individual gestures [16].

Computational models of NVB allow robots to generate their own gaze and gestures in response to the context of the interaction. Some of these models are based on empirical examples for human performance, such as data-driven models of tutoring [17] and narration [18]. Others are based on contextual and semantic knowledge [19].

In this paper, we use an NVB model that accounts for a user's perspective to select the correct deictic behavior for object references. Some robot behavior generators also model the user's perspective to select deictic behaviors for providing route directions [20] or references to people nearby [21]. A robot that simulates human cognition when selecting deictic behaviors can more effectively convey the region of space to which it refers [22]. Our model is different from prior work because it applies to object references, not to people or spatial areas, and because it uses both top-down and bottom-up cues from the scene to model the user's perspective.

Our robot behavior model is inspired by psychology's understanding of how people direct visual attention. Visual attention involves both bottom-up mechanisms to isolate objects of interest from their background, and top-down mechanisms to select task-relevant objects [23]. Both top-down and bottom-up attentional processes are important components of fluid joint action between people and robots [24].

Bottom-up processing involves visual features like color, orientation, and shape, which can be combined into a saliency map indicating the conspicuity of objects in the visual scene [25]. The effect of bottom-up features is mitigated by top-down contextual cues like scene understanding and object recognition that influence where visual attention is directed [26]. Gaze and pointing are top-down cues, serving to draw attention to a particular visual region [10]. However, gaze and pointing only refer to approximate spatial zones rather than to precise linear vectors [27]. Therefore, these nonverbal behaviors can be seen as directing a cone of attention out toward the scene, rather than a single line, a construct we employ in our robot behavior model.

III. Robot Behavior Model

In this study, our robot uses a real-time behavior model [28] to generate appropriate, task-relevant looking and pointing behaviors. This model accounts for elements of the scene—such as the visual configuration of objects from the participant's perspective—and contextual information about the task to select when to gaze or point at objects in the environment.

The model uses both bottom-up and top-down information to decide when to produce nonverbal spatial references. To do so, it considers all possible objects that might be referred to, and calculates a *referential likelihood score* for each of them. This score represents how much the model expects that the user will see that particular object as the target of the robot's reference, given the features of the scene, context of the interaction, and the robot's nonverbal behaviors.

To select the best nonverbal behavior for an object reference, the model simulates referential likelihood scores for all possible verbal and nonverbal behaviors toward the target object. It selects the behavior that maximizes the user's attention toward the target while minimizing the cost of that action.

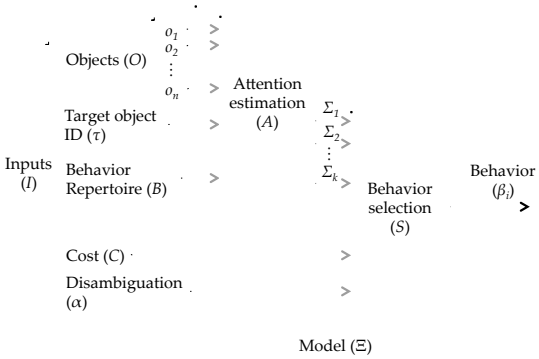

Fig. 2. The behavior model accounts for features of the scene, context of the task, and the robot's capabilities, and outputs a set of verbal and nonverbal behaviors that maximizes user attention to the target object while minimizing behavior cost.

In this section, we provide a mathematical definition of the referential behavior model (Figure 2). In general, the model takes a set of inputs $I = (O, \tau, B, C, \alpha)$ and outputs a single behavior β,

$$\Xi : I \rightarrow \beta \tag{1}$$

The set of objects $O = \{o_1, \ldots, o_n\}$ represents all possible objects that can be referred to. In this experiment, O is the set of objects (blocks and bins) on the table. Of these, o_τ represents the target object.

In this implementation, the robot's behavioral repertoire $B = \{\beta_1, \ldots, \beta_k\}$ contains three actions: speaking, looking, and pointing. Speaking uses the robot's text-to-speech generator. Looking is enacted by having the robot orient the center of its face toward the target location. For a robot that has fixed eyes—like the one used in this study—head orientation must take the place of true gaze, which involves both head and eye movement. Pointing involves the robot reaching its arm so that a ray extending from its shoulder joint to wrist joint intersects with the center of the target. The robot uses whichever arm is closer to the object to point.

C is a ranking of behaviors by relative cost. In this experiment, relative cost is determined by energy expenditure, with speaking as lowest cost, followed by looking, followed by pointing.

The value α represents a disambiguation level for a reference. It specifies how distinctly the referential behavior should indicate the target object, compared to other objects in the scene. A lower α means that that the nonverbal behaviors do not distinguish the target quite as completely from competing objects. In this study, $\alpha = 2.0$ based on pilot testing.

The model uses a composition of two functions to generate behaviors,

$$\Xi : I \xrightarrow{A} (\Sigma, C, \alpha) \xrightarrow{S} \beta \tag{2}$$

A. Attention Estimation (A)

The first function, A, performs *attention estimation*. It estimates how much the user's attention is expected to be drawn to each object in the scene. This function,

$$A : (O, \tau, B) \rightarrow \Sigma \tag{3}$$

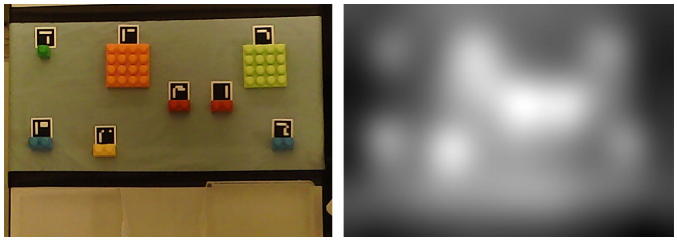

Fig. 3. The NVB model accounts for bottom-up saliency of objects in the scene from the user's perspective. The left image is a participant's view of blocks and table. The right image is a saliency map generated from this view.

takes the set of objects, target object, and robot behaviors, and calculates referential likelihood scores for each object. The set of all referential likelihood scores under all behaviors is Σ. The score for each object under a particular behavior j is $\Sigma_j = \{\sigma_{1,j}, \ldots, \sigma_{n,j}\}$. For example, if behavior β_1 is "pointing," the values in Σ_1 indicate likelihood scores for each object in the scene when the robot is pointing at the target object.

The model takes four cues into account when calculating referential likelihood scores. It considers the *visual saliency* of a scene based on low-level features like color, intensity, and orientation. It identifies top-down *verbal context* based on the descriptive words for each object compared to the words being spoken in the reference. It also recognizes *gaze* and *pointing* behaviors that further disambiguate object references.

The linear combination of these cues gives the likelihood score

$$\sigma_{i,j} = \omega_s S_i + \omega_v V_i + \omega_g G_{i,j} + \omega_p P_{i,j} \tag{4}$$

The likelihood score $\sigma_{i,j}$ of an object $o \in O$ under behavior $\beta_j \in B$ depends on the weighted sum of the visual saliency S_i, the verbal context V_i, the gaze behavior $G_{i,j}$, and the pointing behavior $P_{i,j}$. We use the following weights in our experiment, empirically determined via pilot study: $\omega_s = 0.25, \omega_v = 0.57, \omega_g = 0.91, \omega_p = 1.64$.

1) Saliency: Saliency identifies areas that draw visual attention. It depends on bottom-up features such as color and orientation. In this work, we capture the point of view of the scene from the user's perspective using a camera mounted above the table. We then calculate a saliency map from that image using the Ensembles of Deep Networks (eDN) algorithm [29] (Figure 3). An object's saliency score S_i is calculated as the sum of the above-average pixels in the region of the saliency map representing that object.

2) Verbal Context: Verbal context calculates the proportion of the features in a referential utterance that match descriptor words for an object. The robot's behavior can include utterances u of descriptive words such as "small" and "red." Each object o_i also has a set of descriptor words $D_i \in D$ that apply to it. We calculate the verbal context score as

$$V_{i,j} = \sum_{w \in u} \frac{\text{in}(w, D_i)}{|D|}, \qquad \text{in}(w, D_i) = \begin{cases} 1, & \text{if } w \in D_i \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The value of $V_{i,j}$ ranges from $[0, 1]$, where $V_{i,j} = 1$ indicates that all of the words in the utterance describe the object.

3) and 4) Gaze and Pointing: Deictic behaviors (gazing and pointing) are conceptualized as a cone of attention that is centered on a target. The cone is comprised of a set of rays beginning at origin h of the behavior (head or hand). Rays extend in the direction of the behavior with some pitch ϕ and yaw θ off the center ray, which is focused on the target object. The extent of ϕ and θ were determined visually during piloting: pointing rays extend 5° from center in all directions to represent the robot's apparent field of pointing, and gaze rays extend 15° degrees from center in all directions to represent the robot's useful field of view. If a ray intersects an object, that object's score increases by some base amount. An attenuation function modulates the base score for each ray, with rays closer to the center providing a greater impact on gaze or point scores than rays toward the edge of the cone.

Mathematically, the gaze score is calculated as

$$G_{i,j} = \int_{\theta}\int_{\phi} a^G(\theta,\phi) \cdot I(h,\theta,\phi,o_i)^{-1} \cdot r^G \ \mathrm{d}\phi\,\mathrm{d}\theta \quad (6)$$

where a^G is the attenuation function, I indicates the distance between ray origin h and intersection with object o_i (∞ if no intersection), and r^G is the base score for a gaze ray. The point score $P_{i,j}$ is computed similarly to Equation 6 but with an attenuation function a^P and base score r^P specific to pointing.

The attenuation function determines the impact of the ray on the overall score. Because eye gaze is the less precise deictic behavior, we design that attenuation function (Equation 7) to yield a larger but less focused cone of attention than pointing (Equation 8). The gaze attenuation function

$$a^G(\theta,\phi) = (1+\theta^2+\phi^2)^{-1} \quad (7)$$

indicates that gaze diminishes at a rate inversely proportional to the squared distance of the ray's angular deviation, while the pointing attenuation function

$$a^P(\theta,\phi) = (1+e^{\sqrt{\theta^2+\phi^2}})^{-1} \cdot 2 \quad (8)$$

indicates that pointing is more concentrated near the center, dropping off exponentially with angular distance.

B. Behavior Selection (S)

The model uses the likelihood scores Σ to select the behavior that maximizes the likelihood of the target object while minimizing behavior cost. Because likelihood scores represent the relative likelihood that one object is being referenced compared to another, what matters is the relative likelihood score of the target object compared to the other objects in the scene. We calculate a score b_j that represents how many standard deviations above the mean the likelihood score of the target object, $\sigma_{\tau,j}$, is under each behavior β_j,

$$b_j = \frac{\sigma_{\tau,j} - \bar{\sigma_j}}{\sqrt{\frac{\sum_{i=0}^{n}(\sigma_{i,j}-\bar{\sigma_j})^2}{n-1}}} \quad (9)$$

For a behavior to be selected, the value of b_j must be above the empirically-determined disambiguation threshold α. The system selects the behavior with the smallest cost $c_{\beta_j} \in C$ subject to the constraint $b_j > \alpha$.

1. Put one of the small red blocks on top of the large lime block.
2. Put the small green block next to the red block.
3. Then stack a small blue block on the red block.
4. Put that arrangement in the bin on your right.

Fig. 4. An example of steps for an assembly in the high memorization condition. Assemblies in the high memorization condition all have four steps. Low memorization assemblies have three steps, and are generated by removing the third step of a corresponding high memorization assembly.

IV. Experiment

To evaluate the effect of nonverbal behavior on people's performance during interrupted tasks, we conducted an in-person human-robot interaction study. In this study, we test people's memory for a set of assembly instructions given by a robot. For some participants, the robot uses NVBs to augment its spoken instructions. We compare people's performance with or without NVBs and at various levels of task difficulty to evaluate how NVBs affect instruction recall and task efficiency.

A. Design

1) Experimental variables: This study has three between-subjects independent variables.

- *NVB* is "present" or "absent" depending on whether or not the robot displays nonverbal behaviors when providing the assembly instructions
- *Memorization load* is "low" or "high" depending on the number of steps in the assembly to be memorized
- *Interruption* is "present" or "absent" depending on whether or not the user is interrupted during their completion of the task

Therefore, this study has a 2 (NVB) $\times$ 2 (memorization) $\times$ 2 (interruption) design, which results in eight conditions. Participants are randomly assigned to one of these conditions.

The nonverbal behaviors in the NVB condition are *looking* and *pointing*. These behaviors are autonomously generated in real time in response to object references using the model described in Section III. Details of the model implementation for this experiment are in Section IV-B.

We employ two strategies for changing the difficulty of the task. The first is an increase in memorization load. Each task in this study requires memorizing two assemblies at a time (Figure 4). Low memorization assemblies involve three steps for completion, and high memorization assemblies require four steps. In both cases, the final assembly step is always an instruction to place the assembly in a particular bin. The other assembly steps involve a subject, a spatial relation, and a target. For example, "put the small green block next to the red block" involves the green block (subject), next to (spatial relation), and red block (target).

The second strategy for changing task difficulty is interruption. In this study, an interruption involves completing a

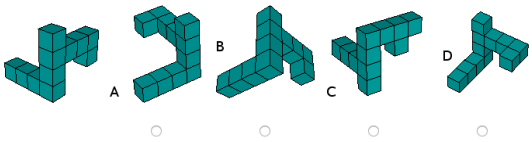

Fig. 5. An example of a mental rotation question used in the interruption. The correct answer is B. Courtesy of [30].

mental rotation test [30] (Figure 5). In the test, participants are shown pictures of a target shape and four possible rotations of that shape. They are asked to select the image that correctly represents what the target shape would look like when rotated. Participants who are interrupted complete eight such questions with a time limit of four minutes. We selected a mental rotation test as an interruption to try to interfere with people's spatial and visual memory for the assembly instructions.

The two difficulty manipulations provide different types of challenges. Increasing memorization load puts greater strain on working memory. The number of steps to be memorized in each task goes from six in the low memorization condition to eight in the high memorization condition, approaching the 7 ± 2 limit to working memory [31]. The interruption, in contrast, presents an unexpected and rapid shift of attention. It was selected to mimic a distraction that might occur during any type of human-robot collaboration.

2) Measures: There are two objective measures and one subjective measure in this study. The first objective measure is *recall accuracy*, how well a participant follows a robot's instructions as measured by the number of correct steps the participant completes in each assembly. Each step is scored individually for accuracy, so participants can receive partial credit for an assembly even if some of the steps are completed incorrectly.

The second objective measure is the *completion time*, how long it takes the participant to put the blocks together once they are given the instructions. Completion time is measured from the moment the robot finishes its instructions to the moment the participant indicates that they are done with the task (see Section IV-C for details). Lower completion times mean more efficient interactions.

The subjective measure is people's perceptions of the robot's animacy, anthropomorphism, intelligence, and likability, using the Godspeed survey [32]. This standardized human-robot interaction questionnaire has five or six Likert-scale questions for each of the four perception items we are studying.

B. Apparatus

The robot in this study is a 58 centimeter tall humanoid called Nao. We used two degrees of freedom in Nao's head to enact looking behaviors and six degrees of freedom in Nao's arms for pointing behaviors. Nao's speech was generated using the robot's built-in text-to-speech system.

Participants constructed different assemblies using eight brightly colored Mega Blocks. A Microsoft Kinect v2 sensor provided real-time sensing capabilities for the Nao, enabling it to detect the blocks in real time and to track their positions in

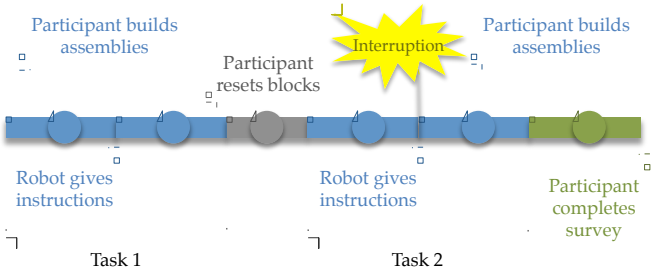

Fig. 6. A timeline of the interaction. If the participant was in the interruption condition, an interruption occurred between the robot's instructions and the participant's assembly in task 2.

3D space. Each block had a fiducial marker attached, so that the Kinect could uniquely identify blocks using the augmented reality library ArUco [33]. The markers were attached along each block's edge, so the center of a marker did not represent the center of a block. To ensure that the Nao's deixis would be correctly aimed at the center of the blocks, the Kinect found block centers using color segmentation and blob detection techniques from the OpenCV library [34], matching the marker closest to a given block center as that block's identifier.

Because object detection and nonverbal behavior modeling occurred in real time, the NVBs in this study were not pre-scripted. The NVBs a particular participant saw depended on the block layout. Though all participants began with the same block layout, after they manipulated the blocks, the NVB to each block was re-calculated based on its new position.

In every condition, Nao shifted its weight slightly from foot to foot to simulate animacy when it was not providing task instructions. When performing computationally expensive actions like calculating saliency scores for each object, which required several seconds at the start of each trial, Nao scanned left and right with its head to simulate looking at all of the objects on the table.

C. Methods

We collected data from 48 participants recruited from the Yale University campus (mean age 26; 25 male, 21 female, 2 other or preferred not to respond). Participants were compensated $5 for this 30 minute study. Participants were randomly assigned to an NVB, memorization, and interruption condition.

Figure 6 provides a visual timeline of the interaction. Participants performed two construction tasks, one after the other. Each task was comprised of two assemblies. For each assembly, Nao provided a set of pre-scripted verbal instructions (Figure 4). For participants in the NVB condition, these verbal instructions were augmented with simultaneous looking and pointing behaviors generated by the model described in Section III. All participants heard the same instructions; NVB and no-NVB trials were identical except that in NVB trials, the robot performed gazes and gestures which simply repeated what was already being conveyed verbally.

There was a timer on the computer screen next to the participant. After Nao was done giving its instructions for the task, it told participants to press "start" on the timer and begin

	Low Difficulty		High Difficulty	
	No Interruption	Interruption	No Interruption	Interruption
No NVB	0.976 (.04)	0.943 (.06)	0.675 (.21)	0.867 (.12)
NVB	0.929 (.11)	0.893 (.13)	0.917 (.05)	0.843 (.18)

TABLE I
AVERAGE RECALL ACCURACY ON TASK 2 FOR EACH OF THE EIGHT CONDITIONS, WRITTEN AS MEAN (STANDARD DEVIATION).

	Low Difficulty		High Difficulty	
	No Interruption	Interruption	No Interruption	Interruption
No NVB	12.5 (2.6)	15.8 (2.0)	23.2 (5.3)	32.7 (7.9)
NVB	13.6 (2.1)	21.9 (5.3)	24.5 (6.5)	25.3 (10.9)

TABLE II
AVERAGE COMPLETION TIME IN SECONDS FOR TASK 2 IN EACH OF THE EIGHT CONDITIONS, WRITTEN AS MEAN (STANDARD DEVIATION).

putting together the blocks, and to press "stop" when they were finished. Task completion time is measured from when the robot finishes its instructions to when the participant pressed "stop" on the timer, in order to account for time they spent thinking before pressing the "start" button.

As illustrated in Figure 6, Nao gave a complete set of instructions first; only after instructions ended were participants allowed to begin construction. Instructions took on average 23 seconds for low and 30 seconds for high memory tasks. Nao did not have to synchronize its behaviors to moving blocks, since blocks were always stationary (i.e., not being manipulated) during the instructions.

For participants in the interruption condition, an interruption occurred just after Nao finished providing instructions to task 2 but before participants could start assembling the blocks. During the interruption, the experimenter came into the room, placed the robot in an idling mode by tapping its head once, and asked the participant to complete a mental rotation test (detailed in Section IV-A1). The test itself had a four minute time limit, and the total interruption time was approximately five minutes, though it varied based on how quickly the participant completed the test. After the interruption, the robot was taken out of idling mode with a second head tap. It then prompted participants to begin the task 2 assembly.

Though we did not inform participants, task 1 is a practice trial that allows people to familiarize themselves with the task and the robot. Results from task 2 are analyzed in Section V.

At the end of the experiment, participants were asked to complete a questionnaire detailing their impressions of the robot and the task. They also provided demographic information at this time.

V. RESULTS

Two participants were excluded for noncompliance, so we examined data from 46 participants.

A. Objective Measures

To evaluate the behavioral effects of our manipulations, we examine the effect of the three experimental variables (memorization load, interruption, and NVB) on the two behavioral metrics (accuracy and completion time, both measured from the second task). Results are shown in Table I for accuracy and Table II for time.

We conducted a three-way analysis of variance (ANOVA) to measure the effects of our three independent variables on recall accuracy. The test revealed a statistically significant effect of memorization load ($F(1,38) = 9.137, p < 0.01$) and a statistically significant interaction between memorization and NVB ($F(1,38) = 4.713, p < 0.05$). Figure 7 illustrates this significant interaction. There was also a borderline significant interaction between interruption and NVB ($F(1,38) = 3.397, p < 0.1$) and a borderline significant three-way interaction among memorization, interruption, and NVB ($F(1,38) = 3.278, p < 0.1$).

We investigate this three-way interaction with tests of simple effects, which reveal how one variable influences the others. First, we conduct a test for simple two-way interactions between interruption and NVB for each level of memorization. This simple two-way interaction yielded a significant effect for high memorization ($F(1,38) = 6.554, p < 0.05$), but not for low memorization ($F(1,38) = 0.001, p = ns$). This indicates that in the high memorization case, the effect of NVB on accuracy depends on whether an interruption occurs. Investigating the interaction further, we run a test of simple simple main effects. We find a statistically significant effect of NVB on accuracy in the interruption absent condition ($F(1,38) = 9.466, p < 0.01$) but not in any other conditions.

To evaluate the effect of our second objective measure, completion time, we conducted a similar three-way ANOVA. For this test, we excluded the timing data from one participant whose response time (89 seconds) was an extreme outlier (> 3 SD from the mean). The test revealed a significant effect of interruption ($F(1,37) = 8.629, p < 0.01$) and memorization ($F(1,37) = 31.490, p < 0.001$). It also identified a borderline significant interaction between memorization and NVB ($F(1,37) = 3.161, p < 0.1$) and a borderline significant three-way interaction between memorization, interruption, and NVB ($F(1,37) = 3.362, p < 0.1$).

As with accuracy, we further investigate this three-way interaction with a test of simple effects. Testing for a simple two-way interaction between interruption and NVB did not yield significance for high memorization ($F(1,37) = 2.639, p = ns$) or low memorization ($F(1,37) = 0.917, p = ns$) conditions. However, a test of simple simple main effects showed a statistically significant influence of NVB on completion time for participants in the high memorization condition when an interruption occurred ($F(1,37) = 4.330, p < 0.05$), but not without an interruption ($F(1,37) = 0.101, p = ns$). In other words, in the high memorization condition, NVB mitigated the effects of the interruption on task completion time (Figure 8). There was no effect of NVB on the low memorization condition, either with or without an interruption.

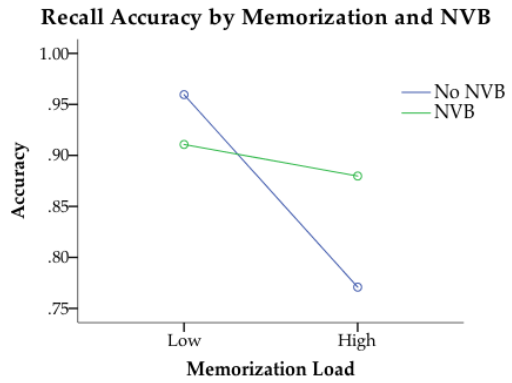

Fig. 7. Accuracy of recall by memorization and NVB conditions. The interaction is significant ($p < 0.05$), indicating that NVB helped mitigate the difficulty of the task.

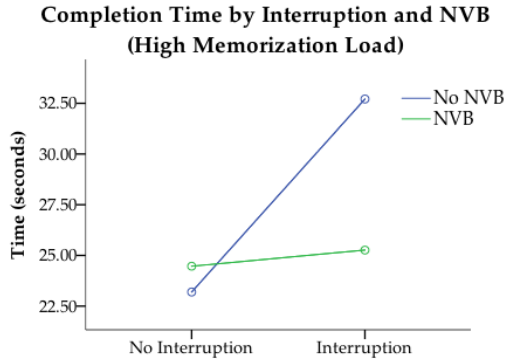

Fig. 8. Completion times for interruption and NVB conditions, shown for the high memorization condition only. There is a significant simple simple main effect of NVB on completion time when an interruption occurs ($p < 0.05$) but not without an interruption.

B. Subjective Measures

Our subjective measure was user perception of the robot in terms of anthropomorphism, animacy, likeability, and perceived intelligence. Each of these four items was measured by five or six Likert-type questions provided in a questionnaire. The items all had high internal consistency as determined by a Chronbach's alpha greater than 0.7.

We conducted a one-way ANOVA to determine if there were differences in people's responses to the four questionnaire items depending on which of the eight conditions they experienced. None of the experimental variables had statistically significant effects on these items (anthropomorphism: $F(7, 38) = 0.510, p = ns$; animacy: $F(7, 38) = 0.159, p = ns$; likeability: $F(7, 38) = 1.096, p = ns$; intelligence: $F(7, 38) = 1.621, p = ns$).

VI. Discussion

Objective measure results show that NVB has little effect on tasks that are already easy, but that when tasks become challenging, NVB improves people's performance by increasing recall accuracy and decreasing completion times.

Hypothesis H1 predicted that task performance (in terms of recall accuracy and completion times) would improve when a robot provided task instructions both verbally and with deictic NVBs, over only providing instructions verbally. H1 is not supported in the general case, because the results do not show a statistically significant effect of NVB across all memorization and interruption conditions.

However, there is a significant interaction effect between NVB and memorization load for both recall accuracy and completion time. This supports H2: in low memory (easier) cases, NVB has no effect on recall or completion time, but in high memory (harder) cases, NVB increases recall accuracy and lowers completion time significantly.

The interaction between NVB and interruption is also significant for completion time and borderline significant for recall accuracy. Again, this supports H2, because NVBs had a positive effect on performance only when there was an interruption that made the task difficult.

Unexpectedly, the hardest condition (interruption present, high memory load) had high recall accuracies in the NVB absent case, leading to statistically indistinguishable performance on NVB present and absent trials. A t-test comparing recall accuracy between NVB present and absent conditions shows that NVB significantly increases accuracy when interruption is absent ($p < 0.05$), but not when interruption is present ($p = ns$). The fact that there is no statistical difference between NVB present and absent for the interruption condition does not invalidate H2, but it fails to provide additional support for that hypothesis.

Subjective measure results were inconclusive. H3 predicted that subjective evaluations of a robot's anthropomorphism, animacy, likeability, and intelligence would increase when the robot used deictic NVBs over when it didn't. Our results do not support H3, because none of the items on the questionnaire reached significance.

This result is in contrast with other studies, which have shown that subjective perceptions of a robot are improved when the robot uses NVBs [8], [18]. While the current study only uses deictic NVBs, however, the previous studies also used expressive NVBs such as iconic or metaphoric gestures [3]. These types of gestures involve producing a visual representation of physical or abstract concepts, such as moving the hand up and down for "chopping" or signaling over the shoulder for "a long time ago." It may be that deictic behaviors do not elicit the same kind of perceptions of agency in a robot as other, more expressive gestures.

From the results, the difficulty manipulations used in this study ("low" or "high" memorization load, and "present" or "absent" interruption) seem approximately equally difficult. Recall accuracy is slightly worse for low memory, interrupted tasks (89%) than for high memory, uninterrupted tasks (92%), while task completion times are slightly worse for high memory, uninterrupted tasks (24.5 seconds) than for low memory, interrupted tasks (21.9 seconds). One limitation of our study is that it only uses two levels for the two difficulty manipulations. Future work could investigate a range of difficulty levels to

identify whether NVB helps even more with more difficult tasks and whether the effect plateaus at any point.

One novel feature of this study is the real time nonverbal behavior model that controlled the robot's actions. Because the model recalculated attention likelihood scores when blocks moved, the NVB a participant saw was specifically targeted toward the scene in front of them. As the results show, this NVB was effective in mediating the effects of a difficult task.

A primary principle of the behavior generation model is that too much NVB can be a hindrance to comprehension. This experiment did not evaluate this claim directly. A future study comparing NVBs produced by the model to other NVB generation models would elucidate how the scene-based model used here compares to other systems that potentially produce more NVBs during an interaction.

We do not claim that our model provides optimal behavior generation for spatial references. However, our model performs at least a subset of the optimal behaviors for nonverbal communication, as determined by the improvement of recall accuracy and completion times. Better results may be possible with a different behavior generation model, and future studies comparing such models would help identify what kinds of NVBs are useful in human-robot collaborations.

Additions to the model might improve its performance. For example, once a robot has named an object by pointing to it, the need to deictically refer to that object again may decrease for a short time afterward. This would add a "prior reference" factor in the likelihood equation, which would increase the likelihood of an object if it has been recently referenced. This factor can decay over time to capture the temporal dynamics of attention. This and other modifications to the model could generate even more natural NVBs.

Acknowledgment

Funding was provided by grants from NSF #113907 and ONR #N00014-12-1-0822. Thanks to Nicole Salomons for help preparing the paper and reviewers for their suggestions.

References

[1] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Oxford, England: Cambridge University Press, 1976.

[2] S. D. Kelly, D. J. Barr, R. B. Church, and K. Lynch, "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory," *Journal of Memory and Language*, vol. 40, pp. 577–592, 1999.

[3] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago: The University of Chicago Press, 1992.

[4] H. H. Clark, "Coordinating with each other in a material world," *Discourse Studies*, vol. 7, no. 4, pp. 507–525, Oct. 2005.

[5] H. Admoni, C. Datsikas, and B. Scassellati, "Speech and gaze conflicts in collaborative human-robot interactions," in *CogSci*, 2014.

[6] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. F. Dominey, and J. Ventre-Dominey, "I Reach Faster When I See You Look: Gaze Effects in Human-Human and Human-Robot Face-to-Face Cooperation." *Frontiers in Neurorobotics*, vol. 6, no. May, pp. 1–11, Jan. 2012.

[7] C.-M. Huang and B. Mutlu, "The Repertoire of Robot Behavior: Designing Social Behaviors to Support Human-Robot Joint Activity," *Journal of HRI*, vol. 2, no. 2, pp. 80–102, Jun. 2013.

[8] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 201–217, 2012.

[9] J. Shah and C. Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human-robot teaming." *Human Factors*, vol. 52, no. 2, pp. 234–245, 2010.

[10] A. Bangerter, "Using pointing and describing to achieve joint focus of attention in dialogue," *Psychological Science*, vol. 15, no. 6, pp. 415–419, June 2004.

[11] S. C. Lozano and B. Tversky, "Communicative gestures facilitate problem solving for both communicators and recipients," *Journal of Memory and Language*, vol. 55, pp. 47–63, 2006.

[12] 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," *IROS*, pp. 708–713, 2005.

[13] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *HRI*, 2009, pp. 69–76.

[14] P. Bremner, A. Pipe, C. Melhuish, M. Fraser, and S. Subramanian, "The effects of robot-performed co-verbal gesture on listener behaviour," in *Humanoids*, 2011, pp. 458–465.

[15] L. D. Riek, T.-C. Rabinowitch, P. Bremner, A. G. Pipe, M. Fraser, and P. Robinson, "Cooperative gestures: Effective signaling for humanoid robots," in *HRI*, 2010, pp. 61–68.

[16] A. St. Clair, R. Mead, and M. J. Matarić, "Investigating the effects of visual saliency on deictic gesture production by a humanoid robot," in *RO-MAN*, 2011, pp. 210–216.

[17] H. Admoni and B. Scassellati, "Data-driven model of nonverbal behavior for socially assistive human-robot interactions," in *ICMI*, 2014.

[18] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *HRI*, 2014, pp. 57–64.

[19] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *IROS*, 2010, pp. 4617–4624.

[20] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: Design of robot's utterance, gesture, and timing," in *HRI*, 2009, pp. 53–60.

[21] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "It's not polite to point: Generating socially-appropriate deictic behaviors towards people," in *HRI*, 2013, pp. 267–274.

[22] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita, "Pointing to space: Modeling of deictic interaction referring to regions," in *HRI*, 2010, pp. 301–308.

[23] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, pp. 193–222, 1995.

[24] G. Hoffman and C. Breazeal, "Robotic partners' bodies and minds: An embodied approach to fluid human-robot collaboration," in *AAAI Workshop: Cognitive Robotics*. AAAI Press, 2006.

[25] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[26] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, no. 2, pp. 194–203, 2001.

[27] G. Butterworth and S. Itakura, "How the eyes, head and hand serve definite reference," *British Journal of Developmental Psychology*, vol. 18, no. 1, pp. 25–50, 2000.

[28] H. Admoni, T. Weng, and B. Scassellati, "Modeling communicative behaviors for object references in human-robot interaction," in *ICRA*, in submission.

[29] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *CVPR*, 2014.

[30] Alaska Research Group, "Spatial ability test," http://psychometrics.akresgr.org/spatialtest/, accessed: 2015-09-01.

[31] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 101, no. 2, pp. 343–352, 1956.

[32] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[33] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and M. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[34] G. Bradski *et al.*, "The opencv library," *Doctor Dobbs Journal*, vol. 25, no. 11, pp. 120–126, 2000.