Automatic Processing of Irrelevant Co-Speech Gestures with Human but not Robot Actors

Cory J. Hayes[1], Charles R. Crowell[2], and Laurel D. Riek[1]
[1] Department of Computer Science and Engineering
[2] Department of Psychology
University of Notre Dame
Notre Dame, IN, USA
{chayes3,ccrowell,lriek}@nd.edu

***Abstract*— Non-verbal, or visual, communication is an important factor of daily human-to-human interaction. Gestures make up one mode of visual communication, where movement of the body is used to convey a message either alone or in conjunction with speech. The purpose of this experiment is to explore how humans perceive gestures made by a humanoid robot compared to the same gestures made by a human. We do this by adapting and replicating a human perceptual experiment by Kelly et al., where a Stroop-like task was used to demonstrate the automatic processing of gesture and speech together. 59 college students participated in our experiment. Our results support the notion that automatic gesture processing occurs when interacting with human actors, but not robot actors. We discuss the implications of these findings for the HRI community.**

***Index Terms*—robot gesture perception, human perception of robots, nonverbal interaction, social robotics**

I. Introduction

Non-verbal communication is an important aspect of human social interaction as well as human-robot interaction [1]. One goal of HRI research focuses on the design of social robots that can communicate with and support humans to the point where they may positively affect human lives and well-being [1][2]. The trust and attitudes that people harbor toward robots also play an important role in HRI, and the key factors of human trust are predictability and a means for social exchange [3]. Social exchange helps to convey the meaning behind why an action is performed or a decision is made and helps to facilitate human trust; erratic and unexpected behavior can damage human trust [3]. Therefore, in human-robot communication it is important to design robot behaviors that both conform to people's expectations, and contingently convey social exchange cues to express intention, attitudes, and emotions [4][5][6]. In this paper, we specifically explore the communication channel of gesture.

Gesturing is an essential mode of communication in human social interaction and is an easy way to convey messages in various situations, such as those where speaking to another person is not desired or possible [7]. Furthermore, gestures are an essential part of multimodal, collaborative dialogue, and aid in both the production of human speech and understanding the speech of others [8][9].

Previous research suggests that the physical appearance of a robot plays an important role in the human tendency to attribute human-like qualities to (i.e., anthropomorphize) humanoid robots and thereby treat them as if they were human [8][10]. Furthermore, this attribution can be influenced by gender; Eyssel and Hegel [11] recently discovered that people apply gender stereotypes to anthropomorphic machines, which can significantly affect social perception in HRI scenarios.

Other work has hinted that there may be limits to the extent to which a humanoid robot can convey information when using gestures along with simultaneous speech (i.e., co-speech gestures). Specifically, the suggestion has been made that when robot co-speech gestures appear to be too human-like, communication with humans can be adversely affected [12][13].

Our work explores the effect that robot co-speech gesturing has on communication compared to human co-speech gesturing, which informs several of these larger questions about anthropomorphism in HRI. This extends our previous work on both perception of robot gesture as well as gendered perceptions of robots [6][14].

II. Background and Research Questions

In this work, we present an experiment that is primarily an adaptation and replication of a study done by Kelly et al. [15] (herein referred to as "the reference study"). The main problem addressed in the reference study was to determine if gesture and speech, when used together, are processed automatically for language comprehension. In their experiment, participants viewed videos where either a male or female actor performed a common gesture accompanied by either a male or female voiceover. The actors in the video performed the gestures while sitting at a table, with only their torso visible. The co-speech gesture performed in each video was either congruent or incongruent with the voiceover. For example, the voiceover "hammer" may have a congruent "hammering" gesture being acted out on the video or an enacted incongruent gesture of "twisting." (See Fig. 1).

In the reference study, participants were given a Stroop-like task of identifying the gender of the speaker in each voiceover as quickly and as accurately as possible. In a Stroop test, participants receive either consonant or conflicting information to process, which in the reference study was either a co-speech

Fig. 1. Example of an incongruent gesture from the Kelly et al. study we are replicating.

gesture congruent or incongruent with the gesture named in the voiceover. Accuracy rates, response times, and event-related potentials (ERPs), were recorded for each response.

The results of the reference study revealed that incongruent gestures led to longer response times, reduced accuracy rates, and higher brain activity. These findings were interpreted by Kelly et al [14] to indicate that participants were compelled to look at/interpret (i.e., automatically process) co-speech gestures even though doing so was not necessary for the task they were completing. These results intrigued us, and we wondered if we could replicate the congruency effects in a behavioral experiment using both human and robot actors.

Our experiment replicates the reference study as closely as possible within both human-human and human-robot interaction contexts. To accomplish this, we substituted the male / female actor variation from the reference study with a humanoid robot / matched-gender human actor. We also substituted the male / female voiceover variation with a synthetic (robotic) / matched-gender human voice. Thus, our experiment conforms to a 2 (actor) x 2 (voice) x 2 (congruency) x 9 (gesture type) within-subject factorial design.

Just as in the reference study, the primary measures in our experiment were participant accuracy and response time (RT). However, we did not record ERP data due to not having access to an ERP instrument.

The overarching goal for our experiment was to address the following question: Is human-humanoid communication similar to human-human communication? If so, then we would expect that irrelevant, incongruent, co-speech gestures made by *either* actor (human or robot) will lead to automatic gesture processing. This would be evidenced, as in the reference study, by slower RTs in the voice identification task when either actor performed a gesture incongruent with the action named in the voiceover.

A. Main Research Question: Are people as affected by robot gesture congruity as they are by human gesture congruity?

Our main research question was to determine if people respond to congruent and incongruent gestures made by a robot in a similar manner to gestures made by another person, and if the findings from the reference study with regard to automatic language comprehension hold for both robot and human co-speech gestures.

This question led to several preliminary questions that we addressed through pilot studies before we performed the main experiment. We were particularly concerned with ensuring a rigorous adaptation the reference study when switching from a male / female paradigm to a human / robot one.

B. Preliminary Question 1: Perceived Robot Gender (Visual)

Our first preliminary question was whether people considered the humanoid robot we intended to use in our experiment (Nao) as either male or female based on its visual appearance alone. This question was important because it would inform our choice for the gender of the human gesturer and voiceover in our main experiment.

In the reference study, the human actors in the stimulus videos differed only in gender, and the Stroop-like manipulation exploited this difference through a gender-based Stroop-like manipulation of both visual and aural channels. For our experiment, we were interested instead in doing a human/robot – based Stroop-like manipulation. (i.e., one gesturer would be human while the other would be a Nao).

Thus, this preliminary question was important in order to ensure that the implied gender of the robot did not contrast with the implied gender of the human subject in our stimulus videos. Given the results found by Eyssel and Hegel [11], maintaining a consistency between the perceived gender of the human and robot actors in our experiment potentially may help reduce variability during our Stroop-like test.

C. Preliminary Question 2: Perceived Robot Gender (Aural)

Since our main experiment used both aural and visual cues to trigger participant responses, a second preliminary question was: Given a human or synthetic voice sample, do people consider the voice to be masculine or feminine?

D. Preliminary Question 3: Perceived Robot Action

A third preliminary question was: given a robot gesture, what do people think it signifies? This question of gesture labeling has an even more substantial influence on the experiment compared to the first two preliminary questions since gestures form the basis of the entire experiment. If participants are unable to determine what the robot is doing when gesturing, then the congruency manipulation will be meaningless.

E. Preliminary Question 4: Comparability of human and robot stimuli labels

The fourth preliminary question asks if a person would interpret a gesture made by a robot in a similar way to how the same gesture made by a human would be interpreted. This question was important to determine which gestures would be used in the main experiment since, for each gesture type, the interpretation needs to not vary depending on the actor.

To address these research questions, we conducted a pilot study to identify visual and aural perceptions of our robot's gender and to know which human actor gender (male or female) would be most comparable to the perceived gender of our robot actor. We then used these findings to create an initial set of human and robot gestures for the purpose of a second pilot study. In this second study, we determined ground truth labels and comparability of the human and robot gesture

stimuli. Finally, we used these findings to prepare our main experimental stimuli videos.

Section III provides an explanation of the materials used for creating our stimulus videos, Section IV describes the results from our pilot studies and main experiment, and Section V discusses our findings and their relevance to the HRI community.

III. Materials

This section describes the creation of stimuli used in both our pilot studies and main experiment. We created two sets of visual stimuli (human and robot gesturers), and two sets of aural stimuli for the voiceovers (human and robot speech).

A. Robot Stimuli

For creating our robot visual and aural stimuli, we used the Aldebaran Nao robot, a 57 cm tall humanoid [16]. Nao has 6 degrees of freedom on each arm, one in the wrist, two in the elbow, two in the shoulder, and one in the hand, where all three fingers are manipulated together to either open or close the hand. This restriction of finger movement prevented us from using a number of the stimulus gestures used in the reference study, so we compensated by programming additional comparable iconic gestures.

We also discovered that the Nao has a limited maximum motor speed and will simply skip over certain programmed movements if that speed is exceeded during programming. However, this limitation did not play an important role for stimulus creation; any robot gestures that were visibly slower than their human-performed counterparts were speed up slightly on video.

To program the Nao's gestures we used Choreographe, a development environment provided by Aldebaran [17]. The interface is mainly drag and drop, and allows the programmer to create a sequenced combination of predefined or custom behavior boxes to manipulate the Nao's joints or attributes, such as its voice or LED colors. Using Choreographe, and the videos we obtained from reference study authors as a guide, we programmed the following 14 gestures and actions: clapping, beckoning, juggling, knocking, sawing, scrubbing, shaking, squeezing, steering, stirring, turning, twisting, wiping, and chicken (the gesture where a person moves their hands near their shoulders and moves their elbows up and down typically to signify cowardice). We deactivated the Nao's LEDs for all stimulus recording.

The robotic aural stimuli also were created using Choreographe, and the Nao generated the utterances through its speaker, which we then audio recorded.

B. Human Stimuli

After performing our pilot studies and determining the robot was perceived as female (see Section IV), we enlisted a female colleague to be recorded as the subject in the human stimulus videos. The subject wore a plain dark blue shirt, matching the dark blue accents on the Nao robot. To avoid any unwanted effects due to distractions, the human subject did not wear any jewelry, and her hair was tied up so that it would not be visible in the videos. The subject was instructed to perform the same 14 gestures in a similar fashion to the Nao. We ensured a careful balance between making the human gestures appear similar to the Nao's while also keeping them human-like; an example of this is visible in Fig. 2 for the gesture "clapping".

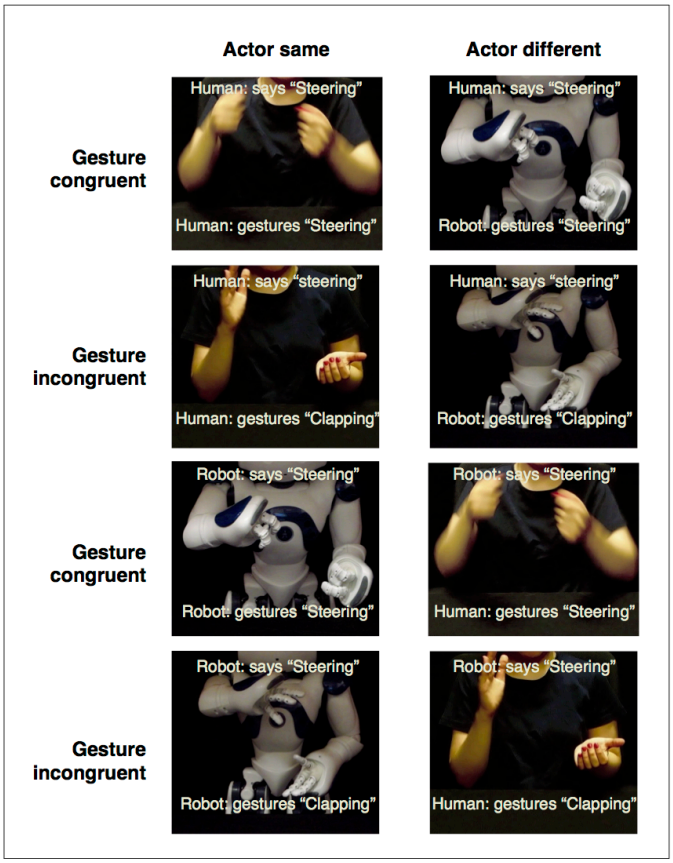

Fig. 2. Example still frames of the congruent and incongruent stimuli for the actor-same and actor-different videos.

For the aural stimuli, we had a female colleague, who is a native English speaker and also a voice actor, record the requisite gesture names. Her voice was recorded using the same audio recorder used for the Nao speech.

C. Video Recording and Processing

Both the Nao robot and the human were filmed in a similar manner to remove as much potential bias as possible. Both the Nao and human subject were filmed against a black fabric background using a Canon Powershot ELPH 300 HS 12 megapixel camera. The human subject performed the gestures while sitting at a table covered in black fabric. Due to its small stature and complexity in terms of programming balance and obstacle avoidance, we were unable to do the exact the same setup for the Nao; however, the stimuli were still nearly identical and comparable, as is visible in the Fig. 2. clapping videos. The human subject and Nao robot each were recorded in one continuous clip to reduce variation during recording, and then clipped in post-production to create single-gesture videos.

The Canon camera used for recording captures 1080p high-definition video. However, the video quality was slightly

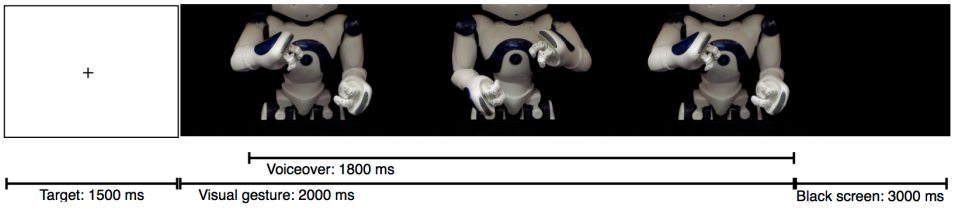

Fig. 3. The timeline of the stimuli videos. First a target was presented, then the visual gesture began. The voiceover started after a 200 ms delay. Finally, a black screen was displayed at the end.

degraded during post-processing where we darkened the scenes to remove potential distractions so as to shift focus as much as possible to the hands and upper torso in the videos. We cropped the videos as necessary to remove anything above the chin area for both the Nao videos and the human subject videos. Each gesture video was clipped to be exactly two seconds long. While there are three phases in gesturing: preparation, stroke, and retraction, the videos were cropped to the point where only the stroke phase is displayed [18]. All post-production was performed using iMovie.

D. Main Experiment Apparati

The application used in the main experiment was programmed in MATLAB using the Psychophysics Toolbox Version 3 (PTB-3) extension. PTB-3 allows for the precise displaying of audiovisual stimuli and interaction with a user. The main experiment was performed using MATLAB 2012a on a Mac Mini with the following specifications: Mac OS X version 10.7.3, a 2.3 GHz Intel Core i5 processor, and 4 GB of 1333 MHz DDR3 memory.

For recording reaction times, we used a Cedrus RB-530 response pad to ensure accuracy. PTB-3 provides built-in functionality with the RB-530, allowing the programmer to log both the response time and the name of the button pressed. Each button of the response pad has a removable top portion where a set of colored or transparent button covers can be fitted. In this experiment, we removed the covers of the top and bottom buttons and placed a small piece of paper stating "ROBOT" and "HUMAN" underneath transparent button covers for the left and right buttons respectively.

IV. Methodology

All pilot studies and the main experiment were approved by the IRB at the University of Notre Dame.

A. Pilot Study 1:Perception of Robot Gender

The preliminary questions regarding the visual and aural perception of Nao's gender were addressed simultaneously in our first pilot study. This study was administered online using SurveyMonkey. Participants first provided demographic information and took a short audio/visual test to ensure their volume was set correctly and they could play clips.

The remainder of the first pilot study was split into two parts. The first part addressed the question of perception of Nao's gender based on appearance alone, and as stimuli used both a still picture as well as muted videos of the Nao. Participants were presented with two videos and a picture and then asked whether the robot appeared to be a male or female. Ratings were obtained using a 5-point labeled scale with "confidently male" and "confidently female" on each end and "gender-neutral" as the center value.

The next part focused on the second preliminary question as to the implied gender attributes of a robotic voice. Participants were presented with four audio clips, with each clip accompanied by the same three rating scales: male, female, or gender-neutral. Using the provided software, we generated samples we believed to be a deep masculine voice, a regular masculine voice, a gender-neutral voice, and a feminine voice.

Eight participants, seven male and one female, completed the study. All participants were at least 18 years of age (mean age = 24.6 years). Given the muted videos and picture, the responses were: 25.0% for confidently male, 12.5% for somewhat male, 62.5% for gender-neutral, and 0% both for somewhat female and confidently female. The deep male voice and regular male voice each received complete consensus "male" votes while the female voice received complete consensus "female" votes. The default Nao voice received 4 "female" votes and 4 "gender-neutral" votes. However, for the last stimulus video where the Nao performed a gesture while speaking using its default voice, 0% voted confidently male or somewhat male, 25% for gender-neutral, 62.5% for somewhat female, and 12.5% voted for confidently female.

These results suggest that just by appearance alone, the gender of the Nao appears to be undefined with a slight lean towards masculine, but when accompanied by a voice that is not obviously masculine, the Nao is strongly believed to be feminine. Since all of the main experiment videos required an accompanying voiceover, we decided to use the default Nao voice as the robotic voice and to use a female subject for the human stimulus videos.

B. Pilot Study 2: Stimuli Selection and Labeling

The third and fourth preliminary questions, how people interpret robot gestures and whether or not the robot and human gestures are comparable, were also addressed simultaneously in another pilot study. In particular, we were interested in establishing ground truth labels of the gesture videos to be used as the aural stimuli in our main experiment.[1]

[1] One reviewer expressed concern over the number of participants in the labeling study, so we repeated it to ensure the same findings. 18 participants on mTurk completed the study (all US citizens, mean age 30.39, 8 women), and were compensated 5 USD. We again found high inter-rater agreement for the nine labels used in our experiment (Nearly all were greater than 12/18 raters agreeing; one gesture, rubbing, was 11/18 raters).

This study was also performed online using Survey Monkey, and participants were recruited using Amazon's Mechanical Turk. Five people participated (all U.S. residents, mean age = 30.2), and were compensated 5 USD for completing the study. The group consisted of three males and two females. One participant's responses were excluded due to being rather unusual and highly incongruous from the others.

Participants were asked to provide their demographic information and completed a short audio / video test. They then completed a short training task to instruct them on how to complete the main experiment. Next, they received all of the initial stimuli videos (14 robot, 14 human) in random order, and were asked two open-ended questions: "What action is being depicted?" and "Why might this action be performed?" The second question helped establish a context for the gesture being performed and provide more information on interpretation. This helped us to better determine which gestures were ambiguous and which gestures had a clear or universal meaning.

The responses were then analyzed side by side, using various dictionary and thesaurus resources to determine which gesture videos had majority agreement on interpretations. Here, we define majority agreement to be 3 out of 4 raters giving the same label to a stimulus video.

Of these videos, we also ensured inter-rater agreement between the labels for both the robot and human analogue (for example, both the human video and the robot video needed to be labeled as "clapping" in order for it to be included).

This pilot study resulted in the following labels: clapping, juggling, rubbing (renamed from "scrubbing"), sawing, shaking, squeezing, steering, turning, and twisting.

Both human and robot voiceovers for theses nine gesture labels were recorded and normalized in Audacity to have similar volume and pitch. In addition to these gesture voiceovers, we also recorded voiceovers for "knocking" and "stirring" to be used for training in the main experiment.

C. Stimuli Video Finalization

Each gesture video was slightly modified before being used in the main experiment. A white screen with a black crosshair was added to the beginning of each video and appeared for exactly 1.5 seconds to prime participants to prepare for the stimulus video. The gesture was then displayed for 2000 ms, and the audio began after a delay of 200 ms, matching the stimuli used by the reference study. Finally, the gesture portion of the video was followed by a black screen displayed for 3000 ms. This allowed any participants who did not respond within the 1800 ms timeframe where the audio began and the gesture portion ended to still have time to key in their response RB-530. (See Fig. 3).

TABLE I. LIST OF THE NINE CONGRUENT AND INCONGRUENT SPEECH-GESTURE PAIRS.

Congruent		***Incongruent***	
Gesture	***Speech***	***Gesture***	***Speech***
Clapping	Clapping	Clapping	Steering
Juggling	Juggling	Juggling	Turning
Sawing	Sawing	Sawing	Squeezing
Rubbing	Rubbing	Rubbing	Juggling
Shaking	Shaking	Shaking	Sawing
Squeezing	Squeezing	Squeezing	Scrubbing
Steering	Steering	Steering	Twisting
Turning	Turning	Turning	Shaking
Twisting	Twisting	Twisting	Clapping

Given there were nine gestures, the variation of the entity that provided the voiceovers (human vs. robot), the variation of the entity that performed the gesture (human vs. robot), and whether the performed gesture was congruent or incongruent with the voiceover, we had a total of 72 videos in the main experiment. Table I shows the congruent and incongruent gesture pairs.

D. Main Experiment

We advertised the main experiment by means of flyers on the Notre Dame campus and the surrounding area, through electronic means such as email and bulletin boards, and through announcements to an undergraduate general education psychology course. Participants from the psychology course could receive extra credit, and other participants could choose a $5 gift card for either Target or Starbucks.

Upon arrival, each participant was directed into a room where he or she was given a consent form, demographics questionnaire, and an instruction form. The participants were told that the purpose of the experiment was to determine how people respond to a synthesized robot voice versus a human voice. Participants were also told that they would watch a series of videos of either a robot or a human performing a gesture, and that their task was to respond as quickly and accurately as possible whether the voiceover in each video was either a robot or human. These statements were on the provided consent forms and also stated verbally by the experimenter as each form was handed out.

After reading and completing the forms, the participant was then led to the main experiment room where the experimenter provided an overview of what would happen and explained both the training phase as well as main portion of the experiment. The experimenter then informed the participant that pressing either of the labeled buttons would advance the screen on the computer, and the program would guide the participant through the experiment via multiple prompts. Then, the experimenter left the room, in order to avoid subject reactivity effects.

Participants received two training sessions to avoid learning effects. In the first training session, the participants practiced using the response pad, by simply responding to a text image on the screen that said either "ROBOT" or "HUMAN" and pressing the corresponding labeled button on the response pad.

For the second training phase, participants were presented with videos similar to the ones used in the main portion of the experiment in a randomized order, and were once again reminded to respond as quickly as possible as to whether the voiceover was a robot or human voice. The gestures and voiceovers used in these training videos were "knocking" and "stirring." It is important to note that in this training phase,

participants were not presented with incongruent speech, resulting in only eight videos used in this portion (human or robot acting out knocking or stirring accompanied by a congruent voiceover).

At the completion of the training phases, the participant was once again reminded of the experiment's main objective (to pay attention to the voice), and the main portion of the experiment began. Each participant encountered a randomized sequence of the 72 stimuli videos. The videos were divided into three 24-video blocks. We provided a 30-second relaxation video between each block to allow participants to take a break from responding. The relaxation videos were accompanied by soft music and did not contain any humans, animals, or robots; only nature scenes. At the conclusion of the relaxation video, the participant was prompted to resume the main experiment by pressing a button on the response pad.

After this main portion of the experiment was complete, the MATLAB program instructed the participant to retrieve the experimenter from the next room. The experimenter then administered an interpersonal sensitivity measure and an attitudinal questionnaire (DANVA2-POS [19] and NARS [20]). The participant then completed a final questionnaire asking whether he/she noticed anything unique about the human voices versus the robot voices while going through the videos, and whether or not they focused primarily on the voice or the entity performing the gesture. Finally, the participant was given a debriefing form, the advertised incentive, and that concluded the experiment.

V. Results

Fifty-nine people at the university served as participants in the main experiment, all but one of whom was an undergraduate (mean age = 20.24 years old). Of these, 31 participants were female. The undergraduate sample included a diversity of majors. All participants were fluent in English, and all but two had lived in the United States for most of their lives. Data from only 54 participants could be included in the final analysis due to equipment failure or other technical difficulties.

The dependent variables derived from the voice classification task were accuracy (i.e. classification of the voice as human or robot was either correct or incorrect) and response time (RT) based on the elapsed time to make the classification following the voice onset. In three instances, out of a total of 3888 trials across all participants, a subject pressed their classification button prematurely (i.e., before the audio voice-over started). No participant made this error more than once. These instances were coded in the data as missing values.

Each dependent variable was analyzed using an appropriate within-subjects factorial ANOVA. In these analyses, significant effects were those with p-values $\leq$ .05. Effect sizes for all ANOVA main effects and interactions were calculated as partial eta squared (η^2_p). Values of η^2_p between .01 and .06 are considered small effects, between .06 and .14 are considered medium effects, and above .14 are large effects [21]. The Greenhouse-Geisser correction was used where necessary to adjust degrees of freedom and p-values for violations of the sphericity assumption in the repeated measures ANOVA.

A. Accuracy

The proportion of correct classifications was computed across each of the nine gesture types within each of the eight Voice, Actor, and Congruency combinations to which each subject was exposed. A 2 (Actor: human vs. robot) x 2 (Voice: human vs. robotic) x 2 (Congruency: congruent vs. incongruent gestures) ANOVA was applied to these data.

Only the main effect of Voice, $F(1,53)=5.43$, $p=.023$, $\eta^2_p=.09$, and the Voice x Actor interaction, $F(1,53)=6.48$, $p=.013$, $\eta^2_p=.11$, emerged as significant from this analysis. The interaction effect, depicted in Fig. 4, shows the mean proportion of correct classifications across the nine gesture types within each of the Voice and Actor conditions. This figure indicates that, regardless of congruency, participants were more accurate in classifying the spoken voice when it was human provided that the video also depicted co-speech gestures being made by a human, whereas the opposite was true when the spoken voice was robotic. Fisher's Least Significant Difference (LSD) follow-up tests confirmed that participants were significantly more accurate with the human actor than with the robot actor under the human voice condition (black vs. gray bars on left; $p=.05$), but the arithmetically opposite difference under the robotic voice condition (black vs. gray bars on right) was not significant. Further comparisons showed that participants were significantly more accurate in their voice classifications when viewing the human actor if the voice was human rather than robotic (left black bar vs. right black bar; $p<.01$), whereas there was no significant difference attributable to voice when the actor was robotic (comparison of gray bars).

B. Response Time

RTs were subjected to a 2 (Actor: Human vs. Robot) x 2 (Voice: Human vs. Robotic) x 2 (Congruency: Congruent vs. Incongruent) x 9 (Gesture Type) within-subjects ANOVA. Since the effect of different gesture types was not the main focus of this study, results reported here will focus on the influence of Voice, Actor, and Congruency variables, collapsed over Gesture Type. Of these factors, only the main effects of Voice, $F(1,50)=17.99$, $p<.001$, $\eta^2_p=.26$, along with the Voice x Actor interaction, $F(1,50)=3.82$, $p=.05$, $\eta^2_p=.07$, and the Actor x Congruency interactions, $F(1,50)=4.78$, $p=.03$, $\eta^2_p=.09$, emerged significant from this analysis.

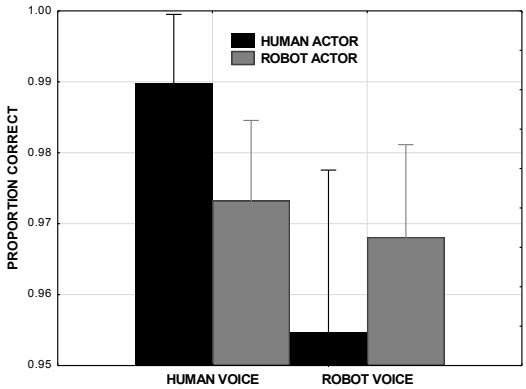

Fig. 4. Mean proportion of correct classification across the nine gesture types within each of the Voice and Actor conditions.

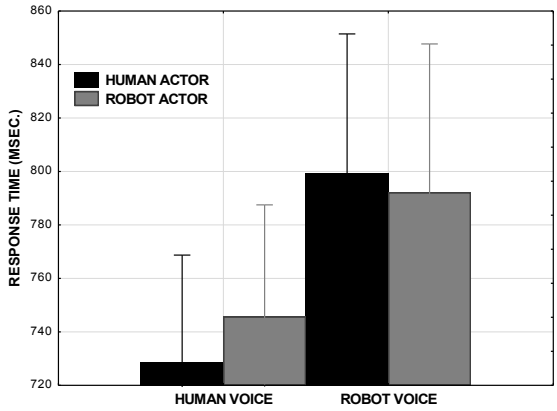

Fig. 5. Mean RTs, collapsed across Gesture Type and Congruency, as a function of Voice and Actor conditions.

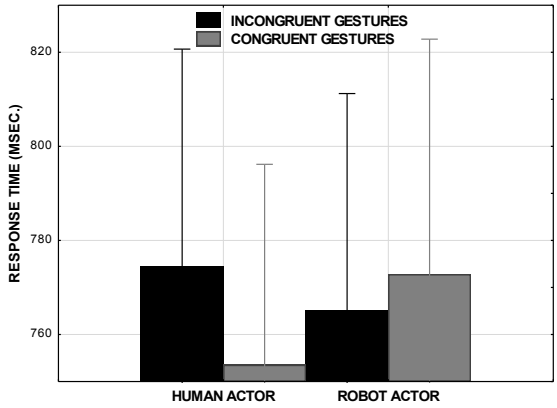

Fig. 6. Mean RTs, for voice classification as a function of both congruent and incongruent gestures being performed by either the human or robotic actor, collapsed across Gesture and Voice types.

Figure 5 depicts mean RTs, collapsed across Gesture Type and Congruency, as a function of the Voice and Actor conditions. This figure reveals that RTs, like accuracy, showed a reversal of the effects of Voice across the two Actor types. LSD follow-up tests indicated that the participants responded significantly faster in the voice classification task when the voice was human provided that irrelevant co-speech gestures were being performed by the human rather than the robot actor (black vs. gray bars on left; p=.05).

In contrast, the opposite was true under the robotic voice condition (black vs. gray bars on right), though the LSD comparison of these conditions was not significant. Additional comparisons revealed that the RTs were slower to the robotic voice than to the human voice, both for the human (comparison of black bars; p<.001) and the robot (comparison of gray bars; p<.001) actors. These findings exactly parallel those reported for accuracy in showing that better voice classification task performance when Voice and Actor matched (e.g., human actor with human voice) than when they did not (e.g., the human actor but robotic voiceover). Taken together, these comparable findings for accuracy and RTs with respect to the Voice x Actor interaction strongly suggest that participants were in fact attending to the video-based gestures even though such attention was irrelevant to completion of the voice classification task.

Similar evidence that participants indeed were attending to the video-based gestures while performing the voice classification task was revealed by the significant Actor x Congruency interaction for RTs reported above. Figure 6 depicts this interaction, collapsed across both Gesture Type and Voice, by illustrating mean participant RTs for the voice classification task as a function of both congruent and incongruent gestures being performed by either the human or robotic actor. This figure reveals that, regardless of gesture type and voice type, participants responded to the voice classification task more quickly when a human actor was performing an action congruent rather than incongruent with the gesture being named in the voice classification task, whereas the opposite was true for gestures performed by the robotic actor. LSD follow-up tests indicated that the former comparison was significant (black vs. gray bars on left; p=.02), but not the latter (black vs. gray bars on right). Additional comparisons also revealed that participants responded more quickly when congruent gestures were performed by the human rather than robot actor (comparison of gray bars; p=.04), but not when incongruent gestures were performed (comparison of black bars).

VI. Discussion

Our results show that the voiceovers significantly impacted participant response times. Our post-experiment survey did not show that understandability of the human or robot voice was a problem. Some participants stated that they occasionally had issues telling the difference between the robot and human voices, but not a single subject commented on any difficulty understanding the robot voice. Comments about the robot voice quality were solely about the clearly artificial tone and odd emphasis on certain syllables. Furthermore, the Nao's voice synthesizing software has been used successfully in various HRI studies such as [22][23][24].

As the Actor x Voice interaction shows, when the actor and voice matched (i.e. robot-synthetic or human-human), response time and accuracy performance was superior to the mismatched condition, regardless of congruency. These results strongly suggest that subjects viewed the synthetic voice as "appropriate" for the robot and the human voice as "appropriate" for the human. However, the fact that it took longer for participants to react to the robot voice may play a role in concomitant gesture processing with robot actors.

These results seem to indicate that humans do automatically attend to (i.e., process) irrelevant gestures performed by human actors. That is, similar to what was shown in the reference study, we observed a Stroop-like effect in the present study when incongruent co-speech gestures were performed by a human actor. However, such automatic processing did not seem to happen with our robot actor, and if anything the effect of congruency was reversed (arithmetically though not statistically) in this condition. The latter finding may suggest that, unlike human communicators, robot communicators do not provoke automatic processing of their co-speech gestures.

These apparent differences in processing of co-speech gestures enacted by human or robot actors possibly can be

attributed to factors related to past experience. That is, as a normal part of our socialization, we come to know that humans have intentions (i.e. cognitive goals) that underlie both speech and gestures during communication. Therefore, when the words of a human communicator conflict with his or her gestures, a conflict of intentionality might occurs that could slow the processing and reactions of the communication receiver. However, the same conflict between the words and gestures for a robot communicator may just be perceived as an error; i.e., a person may not perceive a conflict of intentionality with a robot. More work needs to be done to reconcile this possibility with other work that has been done on the perception of robot intentionality (c.f. [25]).

Yet another dimension of past experience that may be relevant here is that humans have a great deal of communication experience with other humans, but not so with robots. More work needs to be done to identify the role that past experience, or perhaps even evolution, might play in contributing to our finding of differential processing of co-speech gestures with humans and robots.

A limitation of our study to note here is that our sample is young and almost entirely uni-cultural, with the oldest participant being age twenty-six. While we attempted to draw in adult participants of all ages from the local community, the experiment was conducted on a college campus where community accessibility may have been limited.

We plan to expand upon this work in future projects. Given our findings about robot gesture interpretation, we would like to run another version of this experiment with a more diverse sample of participants and observe the effects discovered. A thorough analysis of the interpersonal sensitivity and attitudinal data collected may also lead to insightful findings.

ACKNOWLEDGMENT

We would like to thank Spencer Kelly for providing example videos from which we based our stimuli for this experiment as well as Maryam Moosaei, Elise Eiden, Ninabeth Cabahug, Michael Gonzales, Leonard Hall, Alexandra Janiw, Michael Clark, Michael Villano, and members of the Notre Dame eMotion and eCognition Lab.

REFERENCES

[1] L. En, S. Lan. "Applying Politeness Maxims in Social Robotics Polite Dialogue", 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

[2] V. Chidambaram, Y. Chiang, B. Mutlu. "Designing Persuasive Robots: How Robots Might Persuade People Using Vocal and Nonverbal Cues", 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

[3] M. Lomas, R. Chevalier, E. Cross, R. Garret, J. Hoare, M. Kopack. "Explaining Robot Actions", 7th ACM/IEEE Int'l Conf. on Human-Robot Interaction, 2012.

[4] I. Poggi, C. Pelachaud, F. Rosis, V. Carofigliom, B. Carolis. "Greta. A Believable Embodied Conversational Agent", Multimodal Communication in Virtual Environments, 2005.

[5] C. Breazeal. "Emotion and Sociable Humanoid Robots", International Journal of Human-Computer Studies, vol. 59, 2003.

[6] L. Riek, T. Rabinowitch, P. Bremner, A. Pipe, M. Fraser, P. Robinson. "Cooperative Gestures: Effective Signaling for Humanoid Robots", 5th ACM/IEEE International Conference on Human-Robot Interaction, 2010.

[7] D. Halpern, J. Katz. "Unveiling Robotphobia and Cyber-Dystopianism: The Role of Gender Technology and Religion of Attitudes Towards Robots", 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

[8] J. Bavelas, N. Chovil, L. Coates, L. Roe. "Gestures Specialized for Dialogue". Personality and Social Psychology Bulletin, vol. 21, 1995.

[9] H. Boukje, S. Kita, Z. Shao, A. Ozyurek, P. Hagoort. "The Role of Synchrony and Ambiguity in Speech-Gesture Integration during Comprehension", Journal of Cognitive Neuroscience, vol. 23, 2011.

[10] C. Liu, C. Ishi, H. Ishiguro, N. Hagita. "Generation of Nodding, Head Tilting, and Eye Gazing for Human-Robot Dialogue Interaction", 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

[11] F. Eyssel, F. Hegel. "(S)he's Got the Look: Gender Stereotyping of Robots," in Journal of Applied Social Psychology, vol. 41, issue 9, 2012.

[12] A. Kim, H. Kum, O. Roh, S. You, S. Lee. "Robot Gesture and User Acceptance of Information in Human-Robot Interaction", 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012.

[13] A. Saygin, T. Chaminade, B. Urgen, H. Ishiguro. "Cognitive Neuroscience and Robotics: A Mutually Beneficial Joining of Forces", RSS 2011.

[14] C. Crowell, P. Shermerhorn, M. Scheutz, M. Villano. "Social presence effects of gendered voice and robot entities: perceptions and preconceptions, IEEE/RSK International Conference on Intelligent Robots and Systems, 2009.

[15] S. Kelly, P. Creigh, J. Bartolotti. "Integrating Speech and Iconic Gestures in a Stroop-like Task: Evidence for Automatic Processing," in J. Cog. Neurosci, 2010.

[16] Key Features – Aldebaran Robotics: Available at: http://www.aldebaran-robotics.com/en/Discover-NAO/Key-Features/hardware-platform.html

[17] Choregraphe – Aldebaran Robotics: Available at: http://www.aldebaran-robots.com/en/Discover-NAO/Software/choregraphe.html

[18] A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge Univ. Press, 2004.

[19] H. Pitterman, S. Nowicki. "A Test of the Ability to Identify Emotion in Human Standing and Sitting Postures: The Diagnostic Analysis of Nonverbal Accuracy-2 Posture Test", Genetic, Social, and General Psychology Monographs, 2004.

[20] T. Nomura, T. Suzuki, T. Kanada, K. Kato. "Measurement of Negative Attitudes Towards Robots", Interaction Studies, vol. 7, 2006.

[21] J. Cohen. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[22] L. Ismail, S. Shamsudin, H. Yussof, F. Hanapiah, N. Zahari. "Robot-based Intervention Program for Autistic Children with Humanoid Robot Nao: Initial Response in Stereotyped Behavior", International Symposium on Robotics and Intelligent Sensors, 2012.

[23] S. Shamsuddin, H. Yussof, L. Ismail, F. Hanapiah, S. Mohamed, H. Piah, N. Zahari. "Initial Response of Autistic Children in Human-Robot Interaction Therapy with Humanoid Robot Nao", 8th IEEE International Colloquium on Signal Processing and its Applications, 2012.

[24] H. Cuayahuitl, I. Kruijff-Korbayova. "Towards Learning Human-Robot Dialogue Policies Combining Speech and Visual Beliefs", in proceedings of Paralinguistic Information and its Integration in Spoken Dialogue Systems, 2011.

[25] J. Hart, M. Vu, B. Scassellati. "No Fair!!! An interaction with a Cheating Robot", 5th ACM/IEEE International Conference on Human-Robot Interaction, 2010