Efficiency of Speech and Iconic Gesture Integration For Robotic and Human Communicators - a Direct Comparison

Paul Bremner[1], Ute Leonards[2]

Abstract— **Co-verbal gestures are an important part of human communication, improving its efficiency for information conveyance. A key component of such improvement is the observer's ability to integrate information from the two communication channels, speech and gesture. Whether such integration also occurs when the multi-modal communication information is produced by a humanoid robot, and whether it is as efficient as for a human communicator, is an open question. Here, we present an experiment which, using a fully within subjects design, shows that for a range of iconic gestures, speech and gesture integration occurs with similar efficiency for human and for robot communicators. The gestures for this study were produced on an Aldebaran Robotics NAO robot platform with a Kinect based tele-operation system. We also show that our system is able to produce a range of iconic gestures that are understood by participants in unimodal (gesture only) communication, as well as being efficiently integrated with speech. Hence, we demonstrate the utility of iconic gestures for robotic communicators.**

I. Introduction

Humanoid robots are thought to have a number of advantages over non-humanoid robots, one of which is the possibility of communicating with a person in a naturalistic manner, i.e., in a way that is intuitively understood by humans without learning processes. Naturalistic communication is thought to be achievable by mimicking human communication. Though naturalistic communication can be achieved with only speech, human communication is often multi-modal utilising gestures to improve its efficacy and efficiency [1]. Hence, humanoid robots have a potential advantage as the human-like form enables them to produce hand/arm-gestures to accompany speech (co-verbal gestures), in a similar manner to humans. This gives the possibility of people applying mental models of human communication to a humanoid robot. Further, a number of studies revealed that hand gestures improve user perceptions of robots on scales such as likability, and competence, and future contact intentions (e.g. [2][3][4]). A possible explanation for this finding is that humanoid robots engender humanlike behavioural expectations in people they interact with, thus, when these expectations are met, the interaction is viewed more positively.

In human-human communication studies, co-verbal gestures have been shown to add communicative value for listeners, disambiguating speech they accompany, be it for semantic [5] or paralinguistic information [6]. Also, multimodal communication is more efficient and effective at conveying information between speaker and listener than unimodal communication [5]. However, in order for gestures to add communicative information to the speech they accompany, information contained in both modes of the communication must be integrated by the listener [5][7]. This integration process is evidenced by an increase in listener understanding of the information conveyed by the speaker, relative to unimodal communication.

Our knowledge about whether this integration process between perceived action and perceived speech can occur when the information comes from a non-human agent such as a robot, is, as yet, very limited. First studies using co-verbal pointing gestures to improve robot communication seem to suggest that, at least for pointing, such integration can occur [8][9][10]. Other types of gestures, however, have not yet been examined. More importantly, even for those studies in which information integration had been shown for non-human communicators, it remains unclear whether this integration process is as efficient (i.e., occurs as reliably) as when communicating with other humans.

Hence there are two key questions we seek to address here, whether speech and gesture integration of non-pointing gestures occurs for robot communicators, and whether it as efficient as for human communicators. To do so we investigate the integration process of speech and hand gestures for iconic gestures (i.e. gestures that depict aspects of spatial images or actions) when produced by a (live) NAO robot (Aldebaran Robotics, [11]), and compare it directly to the integration obtained when information is derived from a human communicator (appearing on video). In particular, we use a tele-operation system which uses human speech and motion to produce communicative behaviour for the robot, which we then record. To match conditions as closely as possible, the same actor is used for human stimuli recorded on video, and to produce the robot stimuli. Given the previous findings of high recognition and integration rates (close to ceiling) using video stimuli of human performers for the type of iconic gestures used in our study [7], we reasoned videos of human gestures would be comparable in efficacy to live performances, with the advantage of being identical across participants.

The tele-operated approach has a number of advantages over either hand-scripted or autonomously produced speech and gestures in robots, as the robot's gestures can be closely matched in both form and timing to the original human

This research grant is funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood, grant ref: EP/L00416X/1.

[1]Paul Bremner is with Bristol Robotics Laboratory, University of The West of England, Bristol, BS16 1QY, UK. paul.bremner@brl.ac.uk

[2]Ute Leonards is with School of Experimental Psychology, University of Bristol, Bristol, BS8 1TU, UK. ute.leonards@bristol.ac.uk

gestures. Further, this approach allows us to keep the speech identical for human and robot performers.

While the outcomes of the study should be interesting in their own right, they are also an important step in the development of embodied methods of telecommunication with remote users. Whether a tele-presence system in which a NAO robot is used as an avatar for communication by a remote user has advantages over more conventional screen-based approaches such as video conferencing (whether stationary or on a mobile platform such as the Giraff [12]) will improve human-human communication, will depend strongly on people's ability to integrate robotic gestures with the human voice of their remote partner. Hence, we were motivated to compare human video communication with a live robot to evaluate the relative merits of each as a telecommunication medium.

The main contribution of this paper is that speech and gesture integration occurs for co-verbal iconic gestures performed using the tele-operated NAO robot, in the same way that they are for a video of a human performer. Additionally, we demonstrate the efficacy of the tele-operation system for producing comprehensible gestures, and hence a range of gestures that can be understood when performed by a humanoid robot.

II. Background and Related Work

A. Iconic Gestures in Human Communication

In studies of human co-verbal gestures, gestures are typically classified according to their form and function [13][1]. Within the commonly applied classification scheme proposed by Kendon [13] a key class of gestures are termed iconic gestures, which have a clear meaning, and bear a close formal relationship with the semantic content of the speech they accompany. They are described as gestures of the concrete, displaying in their form, and manner of execution, a homology to aspects of the events or objects being verbalised [13]. Note that such gestures often communicate additional information to the words spoken, particularly of the sort that is more easily and effectively conveyed using the gestural channel; for example to convey relative locations of referents, or how a particular action was performed (termed manner gestures) [14]. An example iconic gesture is putting both hands together, then moving the top of the hands apart while keeping the base together as if they are the pages of a book, used with the phrase 'I read' to convey reading a book.

For human-human communication a number of studies established the communicative value of iconic gestures, by examining how such gestures affect the understanding of information in multi-modal communications. One approach to test effectiveness is to measure participants' ability to recall details of speech delivered in dependence on being accompanied by different gestures (e.g. [15][14]). Analysis of participant responses for such experiments is non-trivial, and depends strongly on the memorability of the stimulus material content.

An alternative approach was suggested by Beattie et al. [5], whereby participants were asked questions about short multi-modal vignettes, the answers to some of which were only contained in the gestural channel. However, in such an approach it might be difficult to distinguish between speech and gesture integration, and contextual inferences [16].

To avoid confounds such as the ones potentially inherent in the approaches described above, we decided to base our experiments on a seminal study presented by Cocks et al. [7]. In their study, participants were presented with a series of actions conveyed either through speech alone, speech accompanied by an iconic (manner) gesture, or the gesture alone, and asked to choose the appropriate action from a set of action images. This fully-balanced within subjects design, allowed them to clearly distinguish between action understanding based on unimodal as compared to multi-modal communications. We adapted Cocks et al.'s design for use with the NAO robot and our tele-presence control scheme (see section III).

B. Gestures in Human-Robot Interaction

A primary focus of previous work on robot-performed gestures has been observers' comprehension of these gestures. Results were variable and seemed to depend on the type of gestures used. For example, participants were reported to be able to identify co-operative robot gestures in timed response trials, even when performed with non-humanlike velocity profiles [17], but not iconic, emotive and emblematic gestures [18][19][20]. As gestures in the above studies were hand-scripted and presented in isolation (without speech), it cannot be excluded that instead of robotic gestures themselves being difficult to comprehend, it was the way they were scripted (and the inherent difficulties in scripting) that was the actual problem. In our study, we tried to overcome any scripting-related issues by using our tele-operation control scheme to copy both the shape and the timing of human movement. Note, however, that even a tele-operation control system is limited by the design, and the degrees of freedom of the robotics system used.

When presented with speech, robot-performed pointing (deictic) gestures [8][9][10], revealed that better understanding of relative locations of referents could be achieved by supplementing speech information with such gestures. Thus providing evidence for speech and gesture integration.

Robot performed gestures have also been observed to have effects beyond information comprehension. Huang and Mutlu found that participants' recall of items in a factual talk presented by a robot was reliably improved if the robot used deictic gestures, while other types of gesture had little impact [21]. Bremner et al. found that parts of a monologue accompanied by (metaphoric and beat) gestures were not recalled any better than those without, though higher certainty in the information recalled by the gestures was observed [22]. By contrast, van Dijk et al. found that recall was improved for actions accompanied by redundant iconic gestures [23]. Moreover, Chidambaram et al. [24] reported that participants were significantly more likely to be persuaded by a robot when it used a variety of non-verbal cues including gestures.

To the best of our knowledge, we are the first to directly compare participants' comprehension of speech and iconic gesture integration for human and robot performers in a single experiment, therefore eliminating a range of confounds that make it difficult to compare findings within the literature.

III. Experimental Methods

We conducted an experimental study with 22 participants (12 male, 10 female), aged 18-55 ($M = 34.80 \pm 10.88SD$), all Native English speakers. Participants gave written informed consent to participate in the study which was in line with the revised Declarations of Helsinki (2013), and approved by the Ethics Committee of the Faculty of Science, University of Bristol.

Participants observed a series of pre-recorded communications which were comprised of either speech, gesture or both speech and gesture, performed by either a person (on video) or the NAO robot (physically present). Human video stimuli were used to enable validation of the experimental procedure, and allow comparative analysis between the person and a tele-operated robot. Hence, the experiment used a 3(communication mode) x 2 (performer) within-subjects design.

A. Tele-operation System

We have designed a tele-operation system to reproduce gestures from a tele-operator, on the NAO humanoid robot platform from Aldebaran Robotics (see Fig. 1, for specifications see [11]). The system is built using the ROS framework [25], with nodes to gather kinematic information of the human tele-operator. The gathered information is then published as ROS messages that are processed by a NAO control node that calculates the required commands and sends them to the robot. Fig. 1 shows a schematic of the system architecture.

To ensure that during gestures joint coordination and link orientations are correctly maintained, along with the desired hand trajectory, arm link end points are tracked on the tele-operator. For this purpose a Microsoft Kinect sensor, combined with the Nite skeleton tracker API from OpenNI is used. The Nite skeleton tracker is encapsulated in a ROS

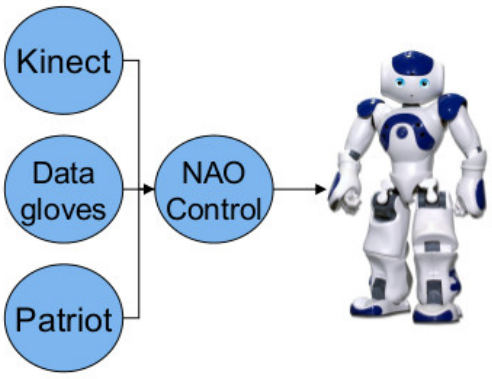

Fig. 1. Tele-operation control architecture. Each circle is a separate ROS node.

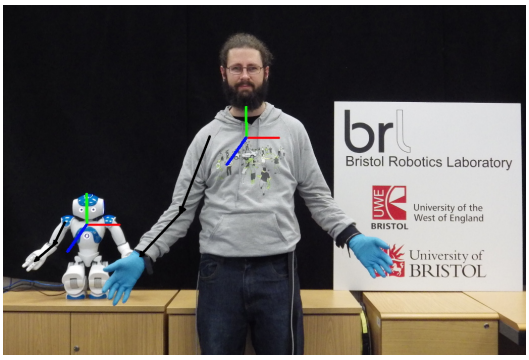

Fig. 2. A matched pose between the tele-operator and the NAO robot. The directions of the arm unit vectors are indicated with black arrows, torso coordinate frames in RGB (XYZ).

node which uses the arm link end points to calculate unit vectors for the arm links relative to the torso coordinate frame of the operator[1], which are then sent as ROS messages. Sensor update rate is 30Hz.

The NAO control node uses the arm unit vectors to calculate the required angles for the robot's arm joints so as to align the robot arm links with equivalent unit vectors in the robots own torso coordinate frame[2]. Fig. 2 gives an example mapping between the human and robot positions. The resulting joint angles are smoothed using a moving average filter with a ten frame window, as the data from the Kinect are subject to high levels of noise.

One limitation of the skeleton tracking data provided by the Nite API (as a result of limitations of the resolution of the Kinect when viewing the full body) is that it is unable to provide tracking information for rotation of the hand relative to the forearm (radial rotation), or finger tracking. To allow tracking of these additional degrees of freedom (DoF) a Polhemus Patriot (for radial rotation) and 5DT data gloves (for finger tracking) are used. ROS nodes were written that package this sensor data as ROS messages, again with an update rate of 30Hz. The NAO node calculates the needed joint angles for these additional DoF, and coordinates them with the other calculated joint angles to send a single command for all controlled DoF each command cycle[3].

In order to stream audio to the robot a NAO module based on the Gstreamer media framework was created, with a corresponding program on the controlling PC.

B. Materials and Procedure

Stimuli for the experiment consisted of a set of 10 verb phrases, depicting common actions (e.g., I paid, I read), chosen from among those used in the Cocks et al. study [7]. Each verb phrase was performed twice, each time accompanied by a different iconic (manner) gesture that conveyed how the action was carried out. In an aim to replicate conversational gesturing the gestures performed were comparatively vague,

[1]calculations omitted here for brevity as they are relatively trivial

[2]see footnote 1

[3]Video of the tele-operation system in action is available in the supplemental material uploaded by the authors, at http://ieeexplore.ieee.org

Phrase	Integration Target
I Cleaned	1. Dusting a lamp 2. Scrubbing a pan
I Cut	1. Cutting with a craft knife 2. Chopping into a melon
I Fixed	1. Hammering a nail 2. Sticking paper with tape
I Lit	1. Pulling a light pull 2. Pressing a light switch
I Measured	1. Pouring liquid into a measuring jug 2. Using a tape measure
I Opened	1. Pulling open a door 2. Opening a book
I Paid	1. Signing a cheque 2. Handing over cash
I Played	1.Playing chess 2. Playing a cello
I Read	1. Reading a newspaper 2. Reading a book
I Rubbed	1. Using a pencil eraser 2. Rubbing a balloon

TABLE I

THE SET OF PHRASES AND THE INTENDED MEANING WHEN COMBINED WITH EACH OF THE TWO MANNER GESTURES (INTEGRATION TARGET)

and less detailed than pantomime gestures would be, e.g., one gesture for 'I paid' is one hand tracing a curling path, and the intended meaning is paying by cheque (see table I for the list of phrases and multi-modal meanings[4]).

Two sets of stimuli were recorded, one for the human performer using a digital video camera, and one for the robot using the tele-operation system. Both sets of stimuli were performed by the same human actor to avoid inter-individual variability in action performance. However, the data-gloves necessary for tele-operation were thought likely to distort participant perceptions if videos of tele-operation were used as stimuli; hence, the two stimulus sets were recorded separately. To control for stimulus set-related biases, a video of each tele-operation performance was reviewed by the actor prior to the recording of each video stimulus, and compared during recording (with repeat performances as needed) to ensure performance was as similar as possible.

To create the robot communication stimuli, the messages from the sensor nodes were recorded to a file using the built in recording capabilities of ROS, as well as being directly streamed to the robot to allow verification during recording. Similarly, the audio (captured using a lapel microphone) was recorded and streamed simultaneously.

The human video stumli and recorded tele-operation stimuli were then edited to produce a set of presentations lasting approximately five seconds each, in three conditions: verbal-gesture condition (VG; verbal phrase heard and gesture seen); gesture only condition (G; gesture performed but audio not played); verbal only condition (V; audio only no performer movement). In both VG and G conditions, there were two versions for each verb phrase, one for each of the different manner gestures; hence, each action phrase came

[4]A video showing some of the gestures used, and the response image sets is available in the supplemental material uploaded by the authors, at http://ieeexplore.ieee.org

Fig. 3. Response image set for "I paid": A and B are the gesture only matches; C and E are the integration targets, and either of them match the speech only condition; D and F are the incongruent foils.

in five different versions per performer (V, G1, G2, VG1, VG2).

Note that for each phrase, the same audio was used for both human and robot performers in all conditions including a verbal component. The human stimuli were created by adding the audio recorded during the robot performances to the videos of the human performance, editing out the original audio on the videos. The original audio on the videos was used only to ensure the correct relative timing between speech and gesture. This overriding of audio-information in the video was seen as necessary to ensure that the audio information provided was absolutely identical between human and robot performer. To ensure there was no unwanted effect of facial gestures, and to prevent any lip-synching issues, the human performer's face was occluded in the video. The speech and gesture timing for the robot conditions was based on the video taken of the robot captured during the original recording with the tele-operation system.

In total there were ten experimental conditions: five communication mode conditions (V, G1, G2, VG1, VG2) for each of the two performer conditions. Ten action phrases were used in each experimental condition; hence, each participant responds to a hundred different trials. Average experiment time was 20 minutes. The trials are split into ten blocks each containing all ten phrases and all ten experimental conditions. To prevent ordering effects, trial presentation order was counterbalanced by means of pseudo-randomisation using partial latin squares across blocks for both condition and sentence order.

After each stimulus presentation, participants were asked to select one out of six colour photographs of people performing actions presented on the 12.1 inch screen of a response laptop by clicking with the laptop's mouse curser on the image they thought matched best what had been communicated; doing so advances to the next trial. Participants' responses were recorded[5]. The layout of the images, and hence the location of the target(s) on the response screen, were randomised between conditions and between phrases.

[5]Presentation of the response images, and recording of responses was done using the PsychoPy software [26]

Fig. 4. Experimental Setup

The sets of pictures for each phrase consisted of: an integration target for each of the two manner gestures, which matched the corresponding speech and gesture combination, a gesture only target for each gesture, that matched the according gesture but not the speech; a pair of unrelated foils, each one semantically linked to one of the gesture-only images, but not matching either the speech or the gesture (Fig. 3 shows an example set, that used for 'I paid'). The integration targets were both semantically congruent with the speech so should have been chosen equally likely in the V condition. For a particular gesture, the integration target and one gesture only image were both semantically congruent with it, so should have been equally likely selected in the G condition. One of the integration targets was the only congruent choice in each of the VG conditions.

The experimental setup is shown in Fig. 4. The NAO and the video screen were both positioned 57cm from the participant. A 32 inch wide-screen TV was used for playback of the video stimuli in order to make the human performer a similar size to the robot. Before each trial which presenter was next was displayed on the screen of the response laptop for 1s, and a tone sounded to indicate trial commencement. Then the phrase performance was played, after which the response images were automatically displayed. Playback of each performance was started by the experimenter from a laptop situated behind the screen so they were not watching the participant during the trials, but could initiate playback, and allow any breaks requested. Before the experimental trials began participants undertook two practice trials to familiarise themselves with the experimental procedure.

C. Data Analysis and Results

To test gesture comprehension we estimated the proportion of correct responses in the gesture only conditions. Each gesture was evaluated individually by comparing the proportion of correct responses against chance (0.33 as there are two possible correct answers for each gesture) using a chi-squared test. These results are shown in Fig. 5 significant results are indicated with * (alpha level 0.05). Almost all gestures (both the *I lit* gestures in the robot condition being the exceptions) were identified significantly better than chance in both human and robot conditions, with high average proportions of correct responses ($M_{human} = 0.943 \pm 0.065SD; M_{robot} = 0.802 \pm 0.17SD$). A Wilcoxon signed rank test showed there was a significant difference between performers ($p<0.001$) for the same gestures even excluding the *I lit* gestures ($p<0.001$).

To test for speech and gesture integration, we examined the proportion of correct integration target choices in the different conditions. The scores for all stimulus items were summed for each participant (incorporating scores for both gestures for each phrase), and the proportion of correct integration target choices was then calculated. Fig. 6 shows participants' group mean image choice for the image combining verbal and gestural information together, in dependence of the presented stimulus mode. Along with expectations that unimodal presentations had beside the integration image also an unimodal image as correct answer, reponses for unimodal conditions were around 50%, and a clear increase of selection of the integrated image can be seen for multimodal iconic information. Note, however, that no clear difference in performance is observed between robot and human presenter.

Accordingly, a 2(presenter) x 3(communication modus) repeated measures ANOVA revealed a significant main effect of communication mode ($F(2,42) = 282.57, MSEffect = 2.21, MSError = 0.008, p < 0.0001$). Post-hoc analysis (Tukey) confirmed that participants chose the 'integrated' images far less often when gestures were presented on their own ($M = 0.39 \pm 0.11SD$) than when verbal information was presented on its own ($M = 0.49 \pm 0.02SD, p < 0.0005$). More importantly, as to be expected if verbal and gesture information is correctly integrated, and therefore ambiguity is decreased, participants chose the image representing integrated verbal and gestural information when both were presented together ($M = 0.82 \pm 0.08SD; p < 0.0005$). Thus, participants were able to clearly integrate verbal and gestural information. There was indeed neither a significant main effect for presenter ($F(1,21) = 2.61, MSEffect = 0.01; MSError = 0.004, p = 0.12$), nor a significant

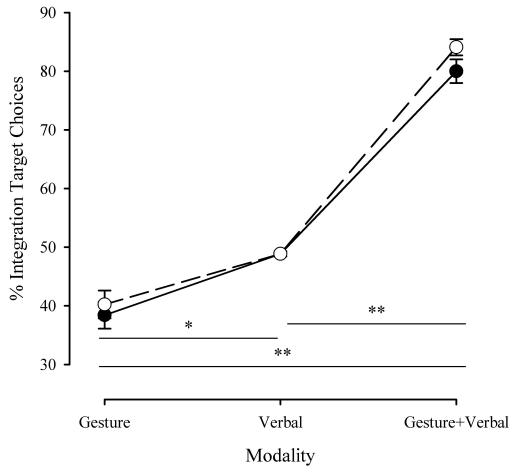

Fig. 6. Group averages for choosing the integrated targets for each communication mode, in dependence of the type of communicator. Filled symbols: robot communicator, open symbols: human communicator. Error bars represent ± 1SEM. * $p<0.0005$; ** $p<0.0001$.

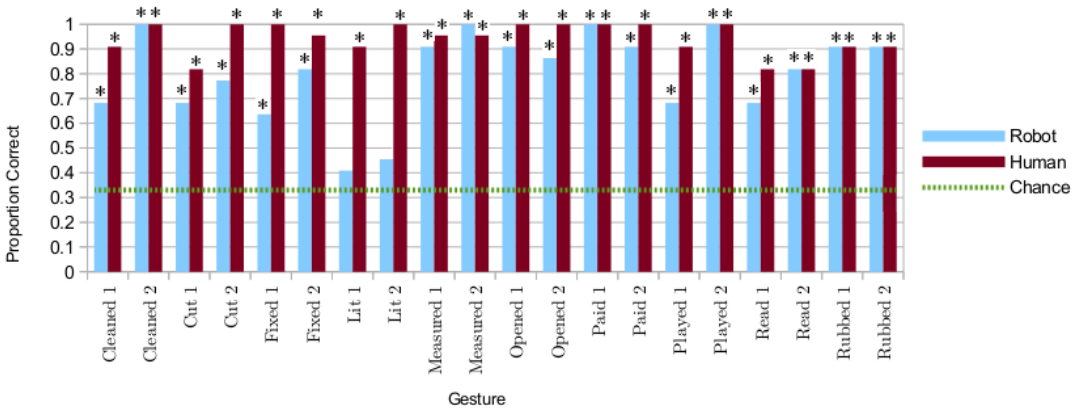

Fig. 5. Proportion of correct identifications of each gesture for the two performers. * p<0.05.

interaction between presenter and communication mode ($F(2, 42) = 1.23, MSEffect = 0.0046, MSError = 0.003p = 0.30$). This first analysis seems to indicate that there is no difference in integration efficiency of verbal and gesture information for human and robot communicators.

In order to give an estimate of the effect size of the VG condition, and confirm the results of the pairwise comparisons of the integration scores, we estimated the change in probability of integration target choices (ITC) between unimodal and multimodal conditions, termed the multimodal gain (MMG). Following the method described in Cocks et al. [7] this is estimated as the difference between the proportion of ITC in the VG conditions (P(Multi)) and an estimate of the proportion of ITC in unimodal communication (P(Uni)), as in (1).

$$MMG = P(Multi) - P(Uni) \tag{1}$$

The proportion of ITC in unimodal conditions (P(Uni)) is calculated as the weighted mean of ITC in the V (ITC_V) and G (ITC_G) conditions, as in (2). This calculation is based on the premise of how likely a given modality is taken into account, i.e., it is assumed that participants are more likely to be influenced by the modality that provides more accurate information. Hence, WV and WG are estimated as normalised proportions of trials in which correct choices were made (PCV and PCG, for V and G trials respectively), as in (3) and (4).

$$P(Uni) = WV * ITC_V + WG * ITC_G \tag{2}$$

$$(2)WV = PCV/(PCV + PCG) \tag{3}$$

$$(3)WG = PCG/(PCV + PCG) \tag{4}$$

Hence, MMG takes into account how often the integration targets were chosen in both unimodal conditions, and was calculated for each performer separately (the results for both gestures for each phrase were included to give one value per performer), shown in Fig. 7 as a percentage gain. By using two gestures per phrase we had an advantage over the original study of Cocks et al. [7] in that for some phrases in the verbal condition there was a clear preference for one of the congruent images; so MMG for that particular image was almost zero regardless of integration, whereas for the other image MMG was very high if integration occurred; hence, we got a clearer picture of the influence of integration by including both these scores in the calculation.

A two tailed t-test was conducted for both performers against the null hypothesis that MMG=0, both sample means ($M_{human} = 0.393 \pm 0.079SD; M_{robot} = 0.355 \pm 0.095SD$) were significantly different from 0 ($t_{human}(21) = 23.12, p < 0.001, r = 0.98; t_{robot}(21) = 17.405, p < 0.001, r = 0.97$). Note that maximum MMG can be estimated as 1-P(Uni), which as a percentage is 56% for the robot, and 55% for the human, so both MMG values are approaching ceiling.

A paired two tailed t test comparing the means of the two performers directly against each other revealed no significant differences between them ($t(21) = -2.005, Diff = 0.019, p = 0.058, r = 0.213$). Note, however there was insufficient statistical power to prove the hypothesis there is no difference between performers. Hence, we used a repeatability measure, the intraclass correlation coefficient, to test if the results are interchangeable between performers; we use ICC(2,k) as the MMG scores for each participant are calculated from multiple measurements (somewhat equivalent to a mean score) [27]. The results were found to be significantly correlated, and indicate fair to substantial reliability ($ICC(2, k) = 0.61, F(21, 21) = 2.8, p = 0.011$); thus making us confident that our Null hypothesis of no difference between performers was indeed the most likely interpretation.

IV. Discussion

The results show that for the gestures used in this experiment, almost all were identified significantly better than chance when they were presented in isolation (without speech) for both the human and the robot. Thus, people were able to identify a range of iconic gestures performed by a humanoid robot. Hence, the iconic gestures examined here

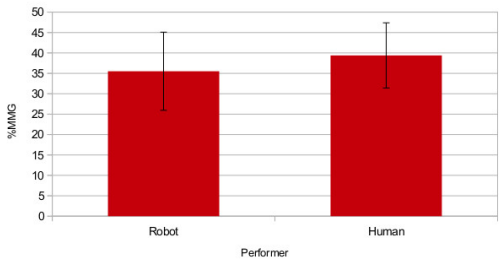

Fig. 7. Group average multimodal gain for each performer. Error bars represent ± 1SD.

represent a large repertoire of viable gestures which can be used for HRI. This observation clearly differs from earlier findings by Cabibihan el al. [18] and Zheng et al. [19] for whom the robotic gestures were difficult to identify on their own. It is hard to determine if the differences between studies are due to subtleties in gestures captured by the tele-operation scheme, the types of gestures used, or the different method of response-gathering (forced choice used here as opposed to open responses in previous work), or some combination of all of these. However, it is important to note that gestures were significantly easier to identify when performed by a person than by the robot, shown by an increase in mean recognition rate from 80.2% to 94.3%. This is most likely due to subtleties of hand movement not conveyed by the tele-operated NAO, that does not possess the same degrees of freedom in the hand as a human performer.

This was most clearly seen in the gestures that accompanied the phrase 'I lit' which were identified correctly in the human condition, but not significantly better than chance in the robot condition. For both gestures (pressing a light switch or pulling a cord as common in the UK), the unrelated foils were chosen with almost identical frequency to the correct answers. Visual inspection of the response images for 'I lit', revealed that the foils could be interpreted to differ from the target gesture images largely by hand orientation and shape. The NAO hand is not capable of hand shapes other than open and closed, and its design may make orientation harder to observe than in the human case. While we endeavoured to select gestures where hand shape was not critical (to ensure fair comparison between performers), naturally performed human gesture inevitably contain a hand shape and orientation component that influences their interpretation. Encouragingly however, the 'I lit' gestures were interpreted correctly when presented with speech, resulting in selection of the correct integration target in 82% and 95% of the time for gestures 1 (pressing light switch) and 2 (pulling a light cord), respectively. This finding seems to suggest that participants were able to infer the missing details of the gesture from the context provided by the speech (note, however, that evidence for such a suggestion is still limited).

The significantly larger proportion of ITC in the VG conditions suggest that speech and gesture are integrated when performed by the tele-operated NAO robot, as they are when performed by a human on video. In particular, the significant difference between the image choices in the two integration conditions (and large effect size) shows that the gestures were integrated with speech, and thus influenced the meaning ascribed to the communication. A clearer measure of the effect of speech and gesture integration was provided by MMG, which estimates, as a single variable, the change in probability that the integration target image was chosen compared to selections made in the uni-modal conditions, in particular the verbal one. The values found for MMG were shown to be highly significant for both performers, and there was no significant difference between them, indicating robot-performed gesture is integrated with speech, and this occurs in the same way as for human communication.

These findings have implications for both the use of a humanoid robot as a tele-communication avatar, and for the design of communicative behaviours for humanoid robots. The main implication is that semantic information in humanoid robot communications can be split across modalities, resulting in more efficient and accurate information conveyance. Further, our findings indicate that multi-modal communications are processed similarly to human ones, so should be used where possible to result in more natural human-robot interactions; this suggestion is in line with, and provides a possible explanation for, previous findings that people give higher subjective ratings of robots that perform gestures [2][3][4].

V. Conclusion

Using a within subject design, we show in this paper that iconic manner gestures conveyed on the NAO robot, using our Kinect based tele-operation system, are recognisable, and, more importantly, are integrated with speech that they accompany. Moreover, the multi-modal integration for robot performances is as efficient as human ones. Hence, with regard to multi-modal semantic information conveyance, a NAO tele-operated avatar can be close to a person in terms of efficacy.

In light of these findings, we suggest that robot communication should be multi-modal to disambiguate its meaning, improve its efficacy, and efficiency, in addition to the improvements in subjective (likability) ratings found in previous work [2][3][4]. Such improvement through multi-modal communication is not only encouraging for our future work on humanoid robot avatars, but also for the design of communication behaviours in autonomous robots: previous studies have found that participants treat avatars similarly to how they do autonomous systems [28], indicating the generality of our results. Indeed, one of the applications of humanoid tele-operation is as a tool to test what is important in terms of robot behaviour for successful HRI in so called super Wizard of Oz studies [29].

VI. Limitations and Future Work

While the work presented here provides initial insight into speech and iconic gesture integration for robotic communicators, it has a number of limitations which we hope to address in future work. Firstly, the range of tested gestures

was limited to manner gestures where hand shape was not expected to be critical. In future work, we hope to further investigate the idea that correct integration occurs even in gestures which cannot be fully realised as humans would, and thus easily understood, due to differences in degrees of freedom between the robot and a human (as occurred for the 'I lit' gestures). If so, one could generalise results far more easily across different robot platforms than is currently possible. Secondly, all gestures used were tested in a laboratory setting. Future work will have to investigate more naturalistic environments such as conversational, interactive settings (extending the ideas in [30]). Thirdly, how close the gestures were to the original human gestures was not directly investigated, and it would be instructive to examine the similarity required for comprehension and integration, for robot design and control requirements (extending the ideas in [17]).

REFERENCES

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

[2] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, May 2013.

[3] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *HRI '13 Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, Mar. 2013, pp. 325–332.

[4] J. Han, N. Campbell, K. Jokinen, and G. Wilcock, "Investigating the use of Non-verbal Cues in Human-Robot Interaction with a Nao robot," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, Dec. 2012, pp. 679–683.

[5] G. Beattie and H. Shovelton, "Why the spontaneous images created by the hands during talk can help make TV advertisements more effective," *British Journal of Psychology*, vol. 96, no. 1, pp. 21–37, Feb. 2005.

[6] L. Wang and M. Chu, "The role of beat gesture and pitch accent in semantic processing: an ERP study." *Neuropsychologia*, vol. 51, no. 13, pp. 2847–55, Nov. 2013.

[7] N. Cocks, G. Morgan, and S. Kita, "Iconic gesture and speech integration in younger and older adults," *Gesture*, vol. 11, no. 1, pp. 24–39, 2011.

[8] J.-J. Cabibihan, W.-C. So, S. Saj, and Z. Zhang, "Telerobotic Pointing Gestures Shape Human Spatial Cognition," *International Journal of Social Robotics*, vol. 4, no. 3, pp. 263–272, Apr. 2012.

[9] T. Ono, T. Kanda, M. Imai, and H. Ishiguro, "Embodied communications between humans and robots emerging from entrained gestures," in *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation.*, vol. 2. IEEE, 2003, pp. 558–563.

[10] A. Sauppé and B. Mutlu, "Robot deictics," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*. ACM Press, Mar. 2014, pp. 342–349.

[11] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of NAO humanoid," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, May 2009, pp. 769–774.

[12] *Giraff, Giraff Technologies AB*. http://www.giraff.org, accessed: Sept. 2014.

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

[14] S. D. Kelly, D. J. Barr, R. 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, no. 4, pp. 577–592, May 1999.

[15] J. Cassell, D. McNeill, and K.-E. McCullough, "Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information," *Pragmatics & Cognition*, vol. 7, no. 1, pp. 1–34, 1999.

[16] G. Beattie and H. Shovelton, "An exploration of the other side of semantic communication: How the spontaneous movements of the human hand add crucial meaning to narrative," *Semiotica*, vol. 184, no. 1-4, pp. 33–51, 2011.

[17] L. Riek, T. Rabinowitch, P. Bremner, A. Pipe, M. Fraser, and P. Robinson, "Cooperative gestures: effective signaling for humanoid robots," *5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010.

[18] J.-J. Cabibihan, W.-C. So, and S. Pramanik, "Human-Recognizable Robotic Gestures," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 4, pp. 305–314, Dec. 2012.

[19] M. Zheng and M. Q.-H. Meng, "Designing gestures with semantic meanings for humanoid robot," in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, Dec. 2012, pp. 287–292.

[20] J. Li, M. Chignell, S. Mizobuchi, and M. Yasumura, "What Do Robot Gestures Tell Us? Emotions and Messages in Simple Robotic Movement," in *Human-Computer Interaction. Novel Interaction Methods and Techniques*, 2009, pp. 331–340.

[21] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14*. ACM Press, Mar. 2014, pp. 57–64.

[22] P. Bremner, A. G. Pipe, C. Melhuish, M. Fraser, and S. Subramanian, "The effects of robot-performed co-verbal gesture on listener behaviour," in *11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, Oct. 2011, pp. 458–465.

[23] E. T. Dijk, E. Torta, and R. H. Cuijpers, "Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 491–501, Sep. 2013.

[24] V. Chidambaram, Y.-H. Chiang, and B. Mutlu, "Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 2012, pp. 293–300.

[25] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "{ROS}: an open-source Robot Operating System," in *Open-Source Software workshop of the International Conference on Robotics and Automation (ICRA)*, 2009.

[26] J. W. Peirce, "PsychoPy–Psychophysics software in Python." *Journal of neuroscience methods*, vol. 162, no. 1-2, pp. 8–13, May 2007.

[27] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability." *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[28] A. M. von der Pütten, N. C. Krämer, J. Gratch, and S.-H. Kang, ""It doesn't matter what you are!" Explaining social effects of agents and avatars," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1641–1650, Nov. 2010.

[29] G. Gibert, M. Petit, F. Lance, G. Pointeau, and P. F. Dominey, "What makes humans so different? analysis of human-humanoid robot interaction with a super wizard of oz platform," in *"Towards social humanoid robots: what makes interaction human-like?" Workshop at International Conference on Intelligent Robots and Systems*, 2013.

[30] H. Z. Hossen Mamode, P. Bremner, A. G. Pipe, and B. Carse, "Cooperative tabletop working for humans and humanoid robots: Group interaction with an avatar," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, May 2013, pp. 184–190.