Speech and Gesture Emphasis Effects For Robotic and Human Communicators - a Direct Comparison

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

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

ABSTRACT

Emphasis, by means of either pitch accents or beat gestures (rhythmic co-verbal gestures with no semantic meaning), has been shown to serve two main purposes in human communication: syntactic disambiguation and salience. To use beat gestures in this role, interlocutors must be able to integrate them with the speech they accompany. Whether such integration is possible when the multi-modal communication information is produced by a humanoid robot, and whether it is as efficient as for human communicators, are questions that need to be answered to further understanding of the efficacy of humanoid robots for naturalistic human-like communication.

Here, we present an experiment which, using a fully within subjects design, shows that there is a marked difference in speech and gesture integration between human and robot communicators, being significantly less effective for the robot. In contrast to beat gestures, the effects of speech emphasis are the same whether that speech is played through a robot or as part of a video of a human. Thus, while integration of speech emphasis and verbal information do occur for robot communicators, integration of non-informative beat gestures and verbal information does not, despite comparable timing and motion profiles to human gestures.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors, software psychology*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*evaluation/methodology, user-centered design*

General Terms

Experimentation, Human Factors

Keywords

Human-robot interaction; Gestures; Humanoid Robots

HRI'15, March 2–5, 2015, Portland, Oregon, USA.
ACM 978-1-4503-2883-8/15/03.
http://dx.doi.org/10.1145/2696454.2696496.

1. 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 the way people communicate with one another. Humanoid robots have this potential advantage as the human-like form enables them to produce hand/arm-gestures to accompany speech (co-verbal gestures), a key feature of human communication [22]. Further, a number of studies revealed that hand gestures improve user perceptions of robots on scales such as likability, and competence (e.g. [26][1]).

In human-human communication studies, emphasis, by means of either pitch accents or beat gestures (rhythmic co-verbal gestures with no semantic meaning), has been shown to serve two main purposes: to provide information as to the intended meaning and syntactic structure in otherwise ambiguous sentences [20], and to indicate which elements of the speech are salient to the speaker [8]. Conveyance of such paralinguistic information makes communications more efficient and effective [22]. However, in order for beat gestures to result in the perception of emphasis in the speech they accompany, perception of speech and gesture must be integrated by the listener [13][34].

Our knowledge about whether perceived action and perceived language can be integrated when the information comes from a non-human agent such as a robot, is, as yet, very limited. First studies investigating the communicative value of beat gestures have found limited effects on recall of items accompanied by beat gestures (a possible proxy for the perceived salience of those items) [14][4]. However, more direct measures of the integration of beat gesture have yet to be examined, in particular when the timing is well controlled for. Previous work on gesture synthesis in robots [27][3] and in animated agents [21][6] has shown the importance of correct temporal alignment of gestures and speech. This tight synchronisation in multi-modal communication is based on work in human communication studies [22][17] as well as empirical observations of human communication by the authors of the aforementioned synthesis work.

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 as when derived from a human communicator. Further, whether the effect of speech emphasis is independent of the speaker also remains unclear.

In an attempt to answer all these questions, we here investigate the integration process of speech and beat gestures when produced by a NAO robot (Aldebaran Robotics, [12]), and compare it directly to that derived from a human communicator. In particular, we use a tele-operation system to produce speech and gesture stimuli for the robot, produced by the same actor used for human stimuli recorded on video, to match conditions as closely as possible. Given the previous findings of high integration rates [13] and salience identification [21][19] from beat gesture using video stimulus, we reasoned videos of human gestures would be comparable in efficacy to live performance, but would convey advantages such as exact stimulus reproducibility 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 gestures. Further, this approach allows us to keep the speech identical for human and robot communicators.

While the outcomes of the study are interesting in their own right, they are also an important step in the development of embodied methods of telecommunication. Whether a tele-presence system in which a NAO robot is used as an avatar for communication by a remote user will improve communication over more conventional screen-based approaches such as video conferencing, will depend strongly on interlocutors' ability to integrate robotic gestures with the human voice of their remote partner. Hence, we are motivated to compare human video communication with a real robot for thir merits as a telecommunication medium.

The main contributions of the work presented here are to investigate whether: i) speech and gesture integration occurs for co-verbal beat gestures performed using the tele-operated NAO robot, in the same way that they are for a video of a human communicator; and ii) the effects of speech emphasis are independent of speaker.

2. BACKGROUND AND RELATED WORK

2.1 Emphasis Effects in Human Communication

In studies of human-human communication the communicative value of speech emphasis has been investigated by examining its effects on speech comprehension [29][20][30]. In particular the grammatical functions of pitch accents (changes in verbal pitch used to provide emphasis) were examined, i.e., their effect on the interpretation of ambiguous sentences. A typical example taken from Schafer et al. [29] and later used by Lee and Watson [20] is the sentence:

'The sun sparkled on the *propeller* of the *plane* that the mechanic was so carefully examining.'

Both propeller and plane are possible responses to the question 'What is the mechanic so carefully examining?', and which is chosen is dependent on which noun is interpreted to be attached to the final relative clause. In other words, there is high attachment if the first noun is selected and low attachment for the second. It was found that the probability of selecting either one was increased by a pitch accent on it [29][20]. A similar result has also been found for other types of sentences [30][7].

However, Lee and Watson reason that the apparent change in meaning might be due to the increased perception of salience of the emphasised word, rather than syntactic resolution; i.e, participants are more likely to give the salient word in response to a post-stimulus question [20]. Indeed, it has been shown that pitch accents often accompany new information in a sentence and this is viewed as more salient by listeners [8][18]. Whatever the cause of the observed effect, the method of examining participants sentence interpretation has been validated as an approach for assessing perception and processing of emphasis, and is adapted for use in the study presented here.

Another means by which people add emphasis to their speech is gestures [22][17]. Within the commonly applied classification scheme proposed by Kendon [17], beat gestures are rhythmic hand motions with no semantic meaning that match the prosody of speech they accompany; their primary functions is to mark salient elements of discourse [17].

Krahmer and Sverts observed that unaccented words accompanied by beat gestures were reported as being more prominent than those without gestural accompaniment [19]. Further, a recent study by Holle et al. showed that salient elements of discourse are distinguished more easily when accompanied by beat gestures, but not animated dots that moved with the same timing as the beat gesture [13]. This indicates that beat gestures were cognitively integrated, but dots were not. This motivates us to investigate whether robot-performed beat gestures are integrated with speech in the same way human-performed beat gestures are, or whether they are treated like the moving dots and thus perceived separately.

2.2 Gestures in Human-Robot Interaction

Previous work in human robot-interaction on how co-verbal gestures affect comprehension has largely focused on pointing (deictic) gestures [5][23][28], revealing that better understanding of relative locations of referents was achieved by supplementing speech information with deictic gestures. This provides some first evidence for speech and gesture integration. Here we examine speech and gesture integration for beat gestures instead. Note however, pointing gestures do carry information in their own rights in contrast to beat gestures that simply emphasise verbal information. Whether robot-produced beat gestures can be integrated with verbal information to disambiguate the latter is as yet unknown, and thus subject of the current study.

Robot performed gestures have been observed to have effects beyond information comprehension. For example, 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 [14]. Similarly, Bremner et al. found that parts of a monologue accompanied by (metaphoric and beat) gestures were not recalled any better than those without, though they reported higher certainty in the information recalled by the gestures [4]. By contrast, van Dijk et al. found that recall was improved for actions accompanied by redundant iconic gestures [9]. It can be considered that recall of conveyed information is an indicator for salience perception, as salient information is more likely to be remembered [10]. However, there are a number of factors that affect memory formation, so thus far there is little direct evidence for integration of speech and robot performed beat gesture; and hence its effect on emphasis perception.

To the best of our knowledge, we are the first to provide direct evidence for speech and beat gesture integration for

robot communicators through a direct comparison between human and robot communicators in a single experiment.

3. EXPERIMENTAL METHODS

To investigate the aforementioned issues of emphasis perception, speech and beat gesture integration, and whether a robot communicator is as effective as a human communicator, we designed a within-participants study. Participants observed a series of pre-recorded communications which had emphasis placed in one of two locations either with a pitch accent or with a beat gesture, performed by either a person (on video) or the tele-operated NAO robot; additionally a "no emphasis" baseline condition was used for comparison for each communicator. Human videos were used as stimuli instead of human live actors to enable exact replication of conditions between participants, to validate the experimental procedure, and allow comparative analysis between video of a person and a tele-operated robot. Hence, the experiment was a 2 (emphasis location) x 2 (emphasis mode) x 2 (communicator) within subjects design.

3.1 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 Figure 1, for specifications see [12]). 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.

To ensure that, during gestures, joint coordination and link orientations are correctly maintained, arm link end points were tracked on the tele-operator. For this purpose a Microsoft Kinect sensor was used to track the arm link end points to calculate unit vectors for the arm links relative to the torso coordinate frame of the operator[1], which were then sent as ROS messages. Sensor update rate was 30Hz.

The NAO control node used 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]. Figure 1 gives an example mapping between the human and robot positions. The resulting joint angles were smoothed using a moving average filter with a ten frame window, as the data from the Kinect was subject to high levels of noise.

One limitation of the skeleton tracking data (due to 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, a Polhemus Patriot (for radial rotation) and 5DT data gloves (for finger tracking) were used. ROS nodes that package these sensor data did so with an update rate of 30Hz. The NAO node calculated the needed joint angles for these additional DoF and coordinated them with the other calculated joint angles to send a single command for all degrees of freedom each command cycle.

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

[2]see footnote 1

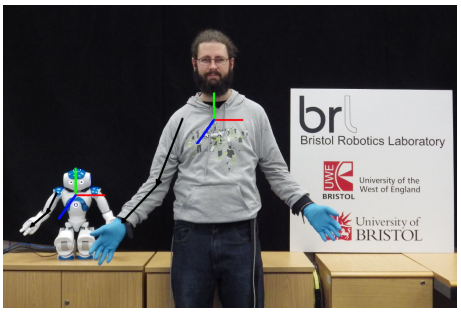

Figure 1: A matched pose between the tele-operator and the NAO robot. The directions of the arm unit vectors are indicated with black arrows, torso coordinate frames in RGB (XYZ).

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.

3.2 Materials and Procedure

3.2.1 Stimulus Material

Stimuli for the experiment consisted of a set of 10 globally ambiguous sentences (see appendix A for the list of sentences), chosen from among those used in Lee and Watson [20]. The sentences chosen were those described as leading to the strongest emphasis effect [20]. The sentences all included relative clauses (RC) preceded by the complex noun phrase consisting of two nouns, both of which could be potentially modified by the relative clauses (see 2.1 for an example, for full list of sentences see appendixA).

The beat gesture used for emphasis in this study was of a similar form to that used by Holle et al. [13] (exaggerated compared to normal conversational gesture): a downward vertical movement of the hand timed to coincide with the word to be emphasised (see Figure 2). This downward movement (the stroke phase of the gesture) is preceded by moving into position prior to the start of the emphasised word (preparation phase), and followed by a return to a rest position (retraction phase).

Two sets of stimuli were recorded, one for the human communicator 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 hand and wrist sensors necessary for tele-operation were seen as likely to distort participant perceptions if videos of tele-operation were used as stimuli; hence, the two sets were recorded separately. To control for any biases produced by this procedure, video of each tele-operation performance was reviewed by the actor prior to the recording of each video stimulus, and compared during recording 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.

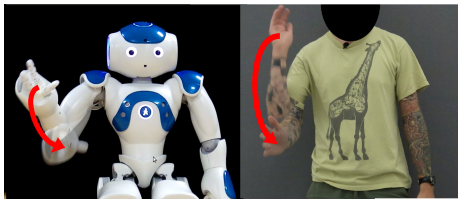

Figure 2: The beat gesture used, performed by NAO and the human actor.

Each sentence in both sets of stimuli was recorded in 3 versions: one each for verbal and beat gesture emphasis on each of the nouns that could be attached to the RC, and one with no emphasis. The stimuli were then edited to produce a set of presentations lasting approximately three seconds each, in five conditions: speech emphasis on noun 1 (S1G0), speech emphasis on noun 2 (S2G0), gesture emphasis on noun 1 (S0G1), gesture emphasis on noun 2 (S0G2), baseline with no emphasis (S0G0). Krahmer and Sverts reported that speakers naturally tend to produce verbal emphasis when producing beat gestures, even when instructed not to do so [19]. To avoid this possibility, the audio from the S0G0 condition was used in the gesture emphasis conditions to ensure that no speech emphasis was present.

Note that for each phrase, the same audio was used for both human and robot communicators 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 synchronise audio and video correctly. Further, to prevent unwanted effects of facial gestures and lip-synching issues, the human communicator's face was occluded in the videos. The speech and gesture timing for the robot conditions was matched as closely as possible to that observed in the human videos. Note that we used this editing procedure as, while we have developed modifications to the filter to reduce delays, performance is not identical to the original timing, which may confound direct comparisons between the two communicators.

To verify that the gestural stimuli were correctly edited, two judges external to the experimental team each viewed the gesture emphasis stimuli from each presenter and were asked to judge which word the gesture accompanied. Previous work has shown people are adept at making such judgements [21]. A similar process was followed for the speech emphasis conditions with the judges being asked to assess the most prominent word in the sentence. In all cases the intended word was identified correctly by both judges.

3.2.2 Beat Gesture Comparison

To verify the performance of the tele-operation system in producing robot beat gestures that are sufficiently similar to those performed by the human actor, we analysed the joint motion profiles recorded by the kinect and those on the NAO joint sensors. Figure 3 shows the change in angle for the major joints for the performers over the duration of a beat gesture: both joint motion and relative timing are similar for the two performers. There is, however, one minor difference as NAO is not able to bend its elbow as much as the actor. However, correct motion of the other joints resulted in a

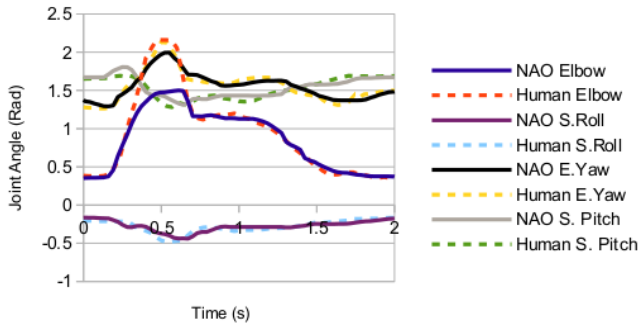

Figure 3: Joint profiles during a beat gesture. NAO joints recorded on the joint sensors, human joints recorded by the Kinect

vertical stroke distance of 71.5mm, approximately 68% of the length of the upper arm of the robot, a gesture large enough to convey the desired emphasis. Further, the velocity and acceleration of the hand during the stroke is similar, key features of how emphatic a beat gesture is [17].

To evaluate if the gestures are percieved similarly between the two actors we had three judges external to the experimental team evaluate the human and robot gestures. They were asked to rate the gestures (presented in a random order) on a 7 point Likert scales (i.e., ratings 0-6) for both suitability for use as a gesture to convey emphasis (human M=5.1 SD=0.71, NAO M=4.8 SD=0.92), and for how emphatic the gestures were (human M=4.5 SD=0.90, NAO M=4.4 SD=0.84). Results were compared using 3 x 2 (judge x performer) mixed ANOVAs [3], no main effect was found for judge or performer for either measurement scale. Combined with a low variance in each set of results (all $\sigma < 1$), gives us confidence the gestures are perceived similarly.

3.2.3 Participants

There were 22 participants (10 male, 12 female), aged 18-55 ($M = 31.8 \pm 9.04SD$), 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.

3.2.4 Setup and Procedure

There were ten experimental conditions: five emphasis conditions (S0G0, S1G0, S2G0, S0G1, S0G2) for each of the two communicator conditions, and ten complex sentences which were used in each experimental condition; hence, each participant responded to 100 different trials. The trials were split into ten blocks, each containing all ten sentences and all ten experimental conditions. To prevent ordering effects, trial presentation order was counterbalanced across participants by means of pseudo-randomisation, using partial Latin squares; sentence order and condition order within each block were both randomised.

As soon as the stimulus had finished, an audio-filed question was played that participants had to answer. The question probed which noun in the stimulus sentence was at-

[3]low variance in the results means standard measures of inter-rater reliability, such as ICC, cannot be meaningfully interpreted

Figure 4: Experimental Setup

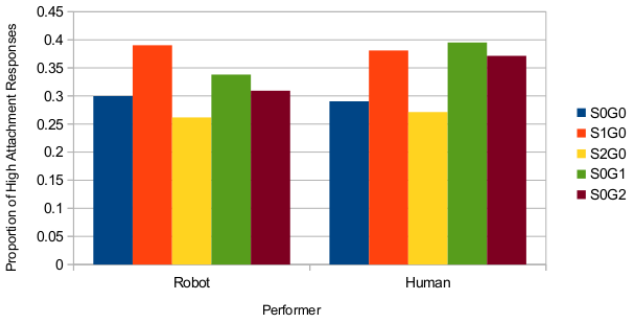

Figure 5: Proportion of high attachment responses for each communicator in the different conditions.

tached to the relative clause. The two possible options participants could choose from to answer were presented on the 12.1 inch screen of a response laptop in 5cm high white letters on a grey background, one at the top of the screen, and one at the bottom. The noun location was randomised and counterbalanced between all trials, so the first noun appeared similarly often at the top and the bottom of the screen. Participants were requested to answer as quickly as possible by pressing one of two response keys on the laptop to ensure intuitive rather than considered responses.

The experimental set-up is shown in Figure 4. The NAO and the video playback screen were both 57cm from the participant, a 32 inch monitor was used for playback of the video stimuli, in order to make the human communicator a similar size to the robot. Before each trial which presenter was next was displayed on the response laptop for 1s, and a tone sounded to indicate trial commencement. The clip was played, after which the response question followed automatically[4]. Playback of each clip was started by the experimenter from a laptop situated behind the screen so they were hidden from the participant during the trials (to prevent observation effects), 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.

3.3 Results and Data Analysis

Evaluation of participant responses is measured by means of proportion of high attachment responses, i.e., proportion of responses in each condition where the first noun was selected as the subject of the relative clause. Before data were entered into statistical analysis, Grubbs' test was applied to detect outliers on numbers of high attachment responses; hence, the results of one participant had to be removed.

Figure 5 shows the results in terms of the proportion of high attachment responses in the different conditions. Following the analysis methods employed by Lee and Watson [20] the data was analysed using mixed logit models, which is an extension of logistic regression, that includes simultaneous modelling of participants and items as random effects, and condition and communicator as fixed effects [15]. The random effects structure was justified by means of likelihood ratio tests [2]. Random effects parameters that significantly improved the model's goodness of fit were included in the model (all p<0.05).

[4]Presentation of the response question and answers, and recording of responses was done using the PsychoPy software [24]

Mixed logit modelling allows us to estimate the change in probability of choosing a high attachment response between levels in the fixed effect conditions, i.e., between communicators, and between emphasis conditions. Hence, we can evaluate not only whether there is a significant change in probability of choosing a high attachment in one condition level compared to another, but also the magnitude of the probability change. In setting up the modelling process a reference condition is chosen for each of the fixed effects, and model parameters (β values) are estimated for each of the other levels of the fixed effects which indicate the change in probability of making a high attachment response between each level (condition) of the fixed effects and the reference. The model parameters are estimated in log-odds space, so we took their exponent to get the odds ratio, which informs us about how many times more likely high attachments are in a given condition compared to a reference condition; e.g., an odds ratio of 2 for a particular condition indicates high attachments are twice as likely in that condition than the reference. Figure 6 shows the estimated odds-ratios for the full model.

There was no significant effect for communicator using human as the reference condition ($\beta = -0.112, SE = 0.099, z = -1.139, p = 0.255$), but a clear effect for S0G1 ($\beta = 0.364, SE = 0.156, z = 2.335, p = 0.019$), and S1G0 ($\beta = 0.455, SE = 0.155, z = 2.935, p = 0.003$) when compared to the baseline no emphasis condition as reference. No other comparisons were significant.

Before concluding that there was indeed no difference in performance for human and robot communicators, we examined differences in efficacy of the two communicators more closely, by evaluating mixed logit models for each communicator type individually. We reasoned that difference in only one condition might be masked by similarity in the others in the full model; Figure 5 indicates such differences.

For human communicators significant differences were observed for S1G0 ($\beta = 0.456, SE = 0.220, z = 2.081, p < 0.05$) and for S0G1 ($\beta = 0.524, SE = 0.219, z = 2.392, p < 0.05$) compared to S0G0 as reference. No other comparisons were significant. However, for the robot communicator significant results were only observed for S1G0 ($\beta = 0.457, SE = 0.226, z = 2.027, p < 0.05$) compared to S0G0 as reference, but not for S0G1 ($\beta = 0.199, SE = 0.222, z = 0.895, p = 0.371$). No other comparisons were significant. Note though, that there is a small increase in odds for S0G1 in the robot condition. Estimating a new model using S0G1

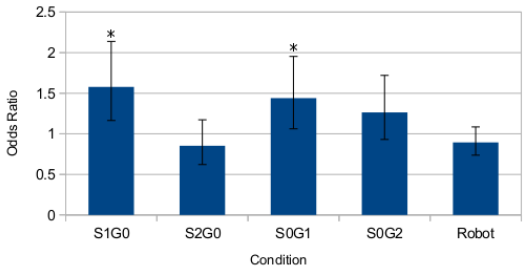

Figure 6: Odds ratios of high attachment responses, as estimated by mixed logit modelling, for each emphasis condition with S0G0 as reference, and for the robot condition with the human condition as reference. Calculated as exponential of the β values. Error bars indicate 95% confidence interval. * p<0.05

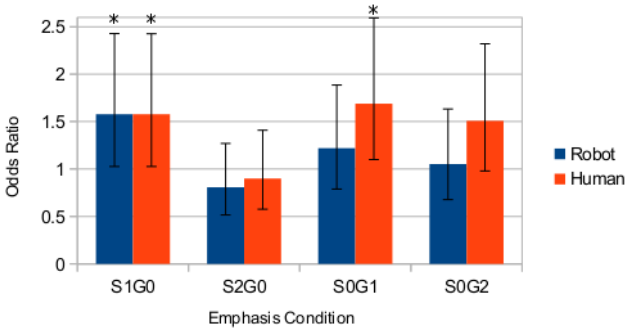

Figure 7: Odds ratios of high attachment responses, as estimated by mixed logit modelling, for each emphasis condition with S0G0 as reference. Calculated as exponential of the β values. Error bars indicate 95% confidence interval. * p<0.05

as the reference condition, there was no significant effect for S1G0 ($\beta = 0.258, SE = 0.216, z = 1.194, p = 0.232$); thus S1G0 and S0G1 are not really different. This indicates that gesture emphasis evokes a similar effect as speech emphasis on the first noun, but far weaker. Odds-ratios calculated for the two separate models are shown in Figure 7.

4. DISCUSSION

To evaluate the relative efficacy of emphasis perception, and speech and beat gesture integration for human and robot communicators, we constructed a number of mixed logit models from our data. Firstly, we constructed a full model that included the results from both communicators, modelling communicators as well as emphasis conditions as fixed effects. No difference between communicators was found. Speech emphasis on the first noun in the stimulus sentences increased the probability of making a high attachment response relative to the no emphasis condition. A similar, but weaker, effect was found for beat gesture emphasis on the first noun, indicating that integration of speech and gesture does occur. Our results for the first noun emphasis condition were similar to those reported by Lee and Watson [20], but our remaining results differ from theirs as we did not observe the significant decrease in probability for high attachments between the emphasis on the second noun (irrespective of emphasis mode) and the baseline.

To better examine the similarities and differences between the results from the two different communicators, models were created for each one separately. Verbal emphasis has the same effect whether played from the video showing a human communicator or through the robot; not a surprising result as both communicators used the same audio, but it nevertheless underlines the value of prosody in robot communication. However, the impact of beat gestures on the first noun (S0G1) differed for the two communicators, revealing an increase in high attachment probability for the human performance but not for the robot. Estimating a new model for the robot using S0G1 as the reference condition we showed that there is no difference in performance for the S0G1 (beat gesture emphasis on noun 1) and S1G0 (pitch emphasis on noun 1); hence, it seems reasonable to conclude that beat gesture emphasis is weakened for robot communicators as compared to human communicators. This was contrary to our expectation that speech and gesture integration would be as effective for the robot communicator as for the human.

At the current stage, one can only speculate why this might be the case. Similarities between our study and a neuroscientific study by Kelly et al, [16] provide a first explanation. Kelly et al. found that a gender mismatch between speech and gesture performers led to reduced speech and gesture integration as compared to a gender match, though brain activity involved in the task was similarly high for matched and unmatched situations. They suggest that while this process is partially automatic (hallmarked as a fast, low-level, obligatory process), it is modulated by some control, and this is triggered by the mismatch. As we used human speech for our robot, this might have caused a similar mismatch as in Kelly et al.'s study. Previous work has shown that mirror neurons fire when observing robot actions [11], so it seems reasonable to suggest that the beat gesture in the robot led to a similar, though reduced, brain activity pattern as the beat gesture of a human. Further work utilising fMRI or EEG measurements of participants observing robot co-speech gesture is needed to investigate these ideas.

An alternative explanation, based on the work by Holle et al. [13], is that robot gestures might not be processed as having communicative intent, a requirement Holle et al. suggest as necessary to improve ease of emphasis perception; as measured using event related potentials (ERPs): comparing the effects of beat gestures and animated dots (with similar timing and movement profiles), Holle et al. found that dots did not aid perception in the way that beat gestures did. It is thus tempting to speculate that robot gestures are processed more like the animated dots than like human gestures, and are therefore not properly integrated with speech. This could be due to inherent differences between human and robot performed gesture, such as differences in ranges of motion, and noise from the robot motors, factors that need to be investigated in detail in future studies.

The reduced impact of speech and gesture integration for robot performed beat gestures has important implications for future work on multi-modal human-robot interaction. The main implication is that an attempt to use beat gestures in robots to improve human-robot communication seems not worthwhile from a salience conveyance perspective if used without speech emphasis. However, it might still be impor-

tant from a more naturalistic communication point of view, as gestures in general are thought to improve subjective (likability) ratings of communicating robots [26][1], engagement into the conversation [4] and personal rapport [32]. At first glance, our results seem to suggest that beat gestures could be added to robot communicators without concern of their exact timing relative to salient elements of the speech; we believe, however, that such a conclusion is premature and that the effects of relative timing of gesture and speech prosody should be investigated explicitly.

Further, our results help to explain previous findings that in robots beat gestures have little or no impact on memory of speech that they accompany [14][4], whereas they have been found to do so in humans [31].

4.1 Limitations and Future Work

While the work presented here provides initial insight into speech and beat gesture integration for robot communicators, it has a number of limitations which we hope to address in future work. In order to separate the effects of speech and beat gestures, we have tested them independently of each other in the same participants. In future work we aim to investigate the combined effects of prosody and beat gestures, and how relative timing affects information processing, engagement, rapport and subjective ratings such as likability and competence. Further, the temporal dynamics such as delays that may be introduced by the tele-operation system need further investigation.

Another limitation in our study is the use of an unaltered human voice for the verbal component of the communications. While this aided direct comparison between performers, and was important for our aims, it does limit the generalisability of the results. In future work, we aim to investigate if there is any effect on integration of a synthetic voice that is more closely matched to the robot's form.

Finally, as changes in attachment selection only provide an indirect measure of integration, we will use event-related brain potentials as in Holle et al. [13] to provide a more direct physiological measure of the similarities and differences between processing of human and robot co-verbal gestures.

5. CONCLUSION

Using a within-subject design, we show in this paper that speech emphasis has a similar impact on speech understanding from a human or robot performer; emphasising the value of prosody for robot communication. More importantly however, we showed that beat gesture emphasis has significantly less effect when performed by a robot than when performed by a human, indicating integration of speech and beat gestures is less effective for a humanoid robot.

In light of these findings, we suggest that salience information in robot communication should not only be conveyed by beat gestures, but needs to be included in the speech pattern through pitch emphasis. Further, future studies involving robot co-verbal gesture should be mindful of the difference in cognitive workload between human and robot multi-modal communications. While humanoid robots such as the NAO are modelled to mimic human communication strategies, how the communication output are processed by the receiving human appears to be different. Further, these findings may provide some explanation as to the differences in the literature in effects of beat gestures on information recall between human [31] and robot [14][4] studies.

Such differences in multi-modal communication are not only interesting 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 biological autonomous systems [33], making us confident that our results should be generalisable to other robot platforms and situations.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] 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*, pages 325–332. IEEE Press, Mar. 2013.

[2] R. H. Baayen. *Analyzing Linguistic Data A Practical Introduction to Statistics using R*. Cambridge University Press, 2008.

[3] P. Bremner, A. G. Pipe, M. Fraser, S. Subramanian, and C. Melhuish. Beat gesture generation rules for human-robot interaction. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1029–1034. IEEE, Sept. 2009.

[4] 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*, pages 458–465. IEEE, Oct. 2011.

[5] J.-J. Cabibihan, W.-C. So, S. Saj, and Z. Zhang. Telerobotic Pointing Gestures Shape Human Spatial Cognition. *International Journal of Social Robotics*, 4(3):263–272, Apr. 2012.

[6] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, pages 477–486. ACM Press, Aug. 2001.

[7] L. F. C. Clifton. Comprehension of Sluiced Sentences. *Language and Cognitive Processes*, 13(4):499–520, Aug. 1998.

[8] D. Dahan, M. K. Tanenhaus, and C. G. Chambers. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47(2):292–314, Aug. 2002.

[9] 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*, 5(4):491–501, Sept. 2013.

[10] S. H. Fraundorf, D. G. Watson, and A. S. Benjamin. Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of memory and language*, 63(3):367–386, Oct. 2010.

[11] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers. The anthropomorphic brain: the mirror neuron

system responds to human and robotic actions. *NeuroImage*, 35(4):1674–84, May 2007.

[12] 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*, pages 769–774. IEEE, May 2009.

[13] H. Holle, C. Obermeier, M. Schmidt-Kassow, A. D. Friederici, J. Ward, and T. C. Gunter. Gesture facilitates the syntactic analysis of speech. *Frontiers in psychology*, 3:74, Jan. 2012.

[14] 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*, pages 57–64. ACM Press, Mar. 2014.

[15] T. F. Jaeger. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language*, 59(4):434–446, Nov. 2008.

[16] S. D. Kelly, P. Creigh, and J. Bartolotti. Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing. *Journal of cognitive neuroscience*, 22(4):683–94, Apr. 2010.

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

[18] E. Krahmer and M. Swerts. Testing the effect of audiovisual cues to prominence via a reaction-time experiment. *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)*, 2006.

[19] E. Krahmer and M. Swerts. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3):396–414, Oct. 2007.

[20] E.-K. Lee and D. G. Watson. Effects of pitch accents in attachment ambiguity resolution. *Language and cognitive processes*, 26(2):262–297, Jan. 2011.

[21] T. Leonard and F. Cummins. The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10):1457–1471, Dec. 2011.

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

[23] 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.*, volume 2, pages 558–563. IEEE, 2003.

[24] J. W. Peirce. PsychoPy–Psychophysics software in Python. *Journal of neuroscience methods*, 162(1-2):8–13, May 2007.

[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] 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*, 5(3):313–323, May 2013.

[27] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics*, 4(2):201–217, Feb. 2012.

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

[29] A. Schafer. Focus in Relative Clause Construal. *Language and Cognitive Processes*, 11(1-2):135–164, Apr. 1996.

[30] A. Schafer, K. Carlson, H. Clifton, and L. Frazier. Focus and the Interpretation of Pitch Accent: Disambiguating Embedded Questions. *Language and Speech*, 43(1):75–105, Mar. 2000.

[31] W. C. So, C. Sim Chen-Hui, and J. Low Wei-Shan. Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, 27(5):665–681, June 2012.

[32] L. Tickle-Degnen and R. Rosenthal. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, 1(4):285–293, Oct. 1990.

[33] 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*, 26(6):1641–1650, Nov. 2010.

[34] L. Wang and M. Chu. The role of beat gesture and pitch accent in semantic processing: an ERP study. *Neuropsychologia*, 51(13):2847–55, Nov. 2013.

APPENDIX

A. STIMULUS SENTENCES

The complex noun phrases used in the stimulus material. Each sentence may have emphasis placed on either of the capitalised words.

- Tyler met the HUSBAND of the WOMAN who the secretary saw in the waiting room on the second floor.
- Brandon interviewed the SON of the LADY who the man worked with for five years in Germany.
- Ashley watched the BROTHER of the WOMAN who the dean interviewed in his office for almost two hours.
- Blake approached the SISTER of the MAN who the president accidentally bumped into at the station.
- Adam greeted the WIFE of the MAN who the landlord complained to about the other tenants
- The sun sparkled on the PROPELLER of the PLANE that the mechanic was so carefully examining.
- The squirrels raced through the LEAVES of the TREE that had recently fallen down in the forest.
- The magazine article failed to mention the LIBRARY of the SCHOOL that had just been built by the contractors.
- The local newspaper described the CEREMONIES of the CLUB that people seemed to find so ridiculous.
- The insurance inspector photographed the COVER of the BOAT that john saw was covered with graffiti.