Personality Perception of Robot Avatar Tele-operators

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

Oya Celiktutan* and Hatice Gunes
Computer Laboratory
University of Cambridge
Cambridge, CB3 0FD, UK
Email: oya.celiktutan@gmail.com, haticeg@ieee.org

***Abstract*—Nowadays a significant part of human-human interaction takes place over distance. Tele-operated robot avatars, in which an operator's behaviours are portrayed by a robot proxy, have the potential to improve distance interaction, e.g., improving social presence and trust. However, having communication mediated by a robot changes the perception of the operator's appearance and behaviour, which have been shown to be used alongside vocal cues in judging personality. In this paper we present a study that investigates how robot mediation affects the way the personality of the operator is perceived. More specifically, we aim to investigate if judges of personality can be consistent in assessing personality traits, can agree with one another, can agree with operators' self-assessed personality, and shift their perceptions to incorporate characteristics associated with the robot's appearance. Our experiments show that (i) judges utilise robot appearance cues along with operator vocal cues to make their judgements, (ii) operators' arm gestures reproduced on the robot aid personality judgements, and (iii) how personality cues are perceived and evaluated through speech, gesture and robot appearance is highly operator-dependent. We discuss the implications of these results for both tele-operated and autonomous robots that aim to portray personality.**

***Index Terms*—humanoid robot; telepresence; personality; multi-modal communication**

I. Introduction

Telecommunication is omnipresent in today's society, with people desiring to be able to communicate with one another, regardless of distance, for a variety of social and practical reasons. While video enabled communication offers a number of benefits over voice only communication, it is still lacking compared to face-to-face interactions [1]. For example remotely located team members are less included in co-operative activities than co-located team members [1], and have fewer conversational turns and speaking time in group conversations [2]. Suggested reasons for these disparities are a lack of social presence of these remote group members, reduced engagement, and reduced awareness of actions [3].

An alternative means of tele-communication that might address some of these issues is tele-operated robot avatars, whereby an operator is given a physical presence in a remote location through a robot that renders their actions. Such systems have been demonstrated to offer benefits with regards to engagement of conversational partners [4], social presence [5], group interaction [4], and trust [6].

A key characteristic of robot avatars is that the operator is represented with a different physical appearance, much as computer generated avatars do in virtual environments. It has been observed that personality perception can be influenced by appearance [7], and this can extend to virtual avatars [8], [9]. How this might manifest with robot avatars, in particular in the interaction between a robot appearance and human voice communication, remains unclear and is yet to be explored.

Personality perception is an important facet of communication. Researchers in psychology have shown that personality plays a key role in forming interpersonal relationships, and predicting future behaviours [10]. These findings have motivated a significant body of work for how people judge others' personalities based on their observable behaviours. Speaking style, gaze, head movements and hand gestures have been frequently reported to be significant predictors of personality [11], [10], [12]. Beyond its importance in social interactions, personality perception is crucial where tele-operated robots are used in a service capacity such as for elderly care [13], and search and rescue [14]. In these settings, perception of the operator will effect system utility for carrying out the desired service and achieving the desired outcome.

In this paper, we are motivated to investigate if and how personality judgements are affected for judges observing targets whose communication is mediated by a robot avatar. We use a robot tele-operation system where human gestures are captured using a motion capture system and portrayed on a humanoid robot. We designed an experiment in which participants performing different communication tasks (targets) are recorded. These recordings are then shown to external observers (judges) for personality assessment. Judges observe three communication conditions: audio-visual, audio-only and tele-operation, which enables us to compare judgements for robot-mediated communication and for video-/audio-mediated communication. Our results show that the appearance of a tele-operated robot influences personality judgements made of its operators, especially for *neuroticism*. While gestures of operators replicated on the robot aid personality judgements, we find that how appearance and gestural cues are utilised together with verbal cues is highly operator-dependent.

* Both authors contributed equally to this article.

II. Background and Related Work

A. Thin Slice Personality Analysis

First impression or thin slice personality analysis, is a body of research that studies the accuracy with which people are able to make personality judgements of others based only on short behavioural episodes (termed thin slices). Analysing these judgements is reported to provide insight into the assessments people make in everyday interactions [15], [10].

In such studies, targets are typically asked to perform a range of communication tasks, either monologues or conversations with confederates, and are filmed while doing so. Judges then observe the video clips and complete personality assessment questionnaires. Ratings of judges can be compared with targets' self ratings, acquaintance ratings, and for between-judge consistency.

Between-judge consistency is typically higher than accuracy of judge ratings to self/acquaintance ratings. For many traits there is sufficient between-judge consistency for the method to be useful in assessing the impressions a person creates on those they interact with [10]. Accuracy is somewhat dependent on task [10] and slice duration [16]. Varying slice duration only affects accuracy for some traits, and where there is an effect, there is typically peak in accuracy at 60s. Further, judgements are also influenced by the availability of cues displayed by the target, i.e., some targets are easier to judge than others [17].

B. Tele-Operation

A common approach to embodied telecommunication is the use of mobile remote presence (MRP) devices: a screen displaying the operators face mounted on a stalk attached to a wheeled base. Studies investigating the utility of MRPs have largely focused on task specific interaction analysis, using measures such as trust [18], and social norms [19]. There are some implications of being able to perceive the personality of the user, but it has not been explicitly studied as we do here.

Previous work on robot avatars has largely focused on their efficacy for interaction (e.g., [5], [13], [4]). However, there are two studies that look directly at personality perception of tele-operators. Kuwamura *et al.* [20] examined an effect that they term *personality distortion*, demonstrated by reduction in internal consistency of the personality questionnaire they used, for two different robot platforms and communication using video. They use 3 tasks: (1) an experimenter talks freely with the participant, (2) a different experimenter introduces and talks about themselves, and (3) a third experimenter interviews the participant. They only observed *personality distortion* for one of the robot platforms, for *extroversion* in the interview task, and for *agreeableness* in the introduction task. Using a single fixed person for each task, particularly members of the experimental team who are aware of the goals of the study, greatly reduces the ecological validity of their results. In contrast, here we use a large number of naïve targets performing naturalistic communication, and conduct far more in-depth data analysis.

In a study with a tele-operated, highly humanlike robot, Straub *et al.* [21] examined both how participant tele-operators incorporate the fact they are operating a robot into their presented identity, and how interlocutors at the robot's location blend operator and robot identities. They used language analysis to make their assessments. They observed that many operators pretended they themselves were a robot, and interlocutors often referred to the operator as a robot. A likely reason is that interlocutors at the robot's location were not told if the robot was autonomous or not and, consequently, they often assumed that the robot was autonomous.

More recently, personality inferences from virtual avatar appearance have been examined. It was observed that judges made relatively consistent inferences based on appearance alone [8], [9], and more attractive avatars were rated more highly in an interview scenario [22]. However, how these inferences interact with communicative behaviours of operators has yet to be examined. Additionally, behaviours in virtual worlds are likely to differ from that of real life.

III. Research Questions

In this study we investigate four research questions. When comparing communication mediated by a robot avatar with video or audio only communication from the same targets:

RQ1. Are there differences in judges' consistency in assessing personality traits (within-judge consistency)?

RQ2. Are there differences in how much judges agree with one another on personality judgements (between-judge consistency)?

RQ3. Are personality judgements less accurate compared to self ratings (self-other agreement)?

RQ4. Are perceived personalities systematically shifted to incorporate characteristics associated with the robot's appearance (personality shifts)?

Inconsistent personality judgements, whether within judges (RQ1), or between judges (RQ2), are an indication of either a lack of cues (causing judges to guess) [23], or conflicting cues (causing judges to utilise different cues for their judgement) [20]. When using a robot avatar it is important to ascertain whether robot appearance cues are utilised, and whether behavioural cues transmitted via the robot can be used to make judgements.

A common method for assessing personality judgements is comparison with targets' self-ratings of the same traits (RQ3). It is a useful measure of whether targets produce cues related to their self perception of personality in the tasks undertaken [23]. Hence, if the targets produce such cues, it is important to ascertain whether perceptions of these cues are altered by robot mediation.

If robot appearance cues are utilised in making judgements, we expect shifts of judgements (from video/audio judgements) to incorporate features related to the robot's appearance (RQ4). This is in light of previous work on personality judgements of virtual avatars based on their appearance [8], [9].

IV. Materials and Methods

We designed a two-stage experimental study to address the above research questions. Firstly, 20 participants (targets) were recorded performing three communication tasks, using both

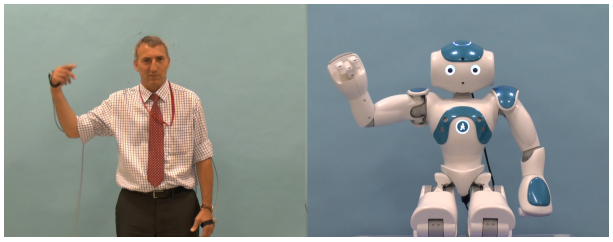

Fig. 1. An example of a matched gesture for a target and the NAO robot recorded during the performance of one of the tasks.

a motion capture based tele-operation system and on video camera. This ensures we have a large set of natural communication behaviours, and hence personality cues, for a range of personality types, that can be viewed directly or mediated by a robot. In the second stage of the study, the recorded data was used to create a set of clips for each target in three communication conditions: audio-visual (video recordings of the targets), audio only, and tele-operated robot (recorded on video). These clips were used to obtain personality judgements of the targets by independent observers (judges).

A. Tele-operation System

In order to reproduce the gestures of targets on the NAO humanoid robot platform from Aldebaran Robotics [24], we use a motion capture based tele-operation system. A similar system has been demonstrated to be capable of producing comprehensible gestures [25]. The arm motion of the targets is recorded using a Microsoft Kinect and Polhemus Patriot[2], and used to produce equivalent motion on the robot. Arm link end points at the wrist, elbow and shoulder are tracked, and are used to calculate joint angles for the robot so that its upper and lower arm links reproduce human arm link positions and motion. This method ensures that joint coordination, and hand trajectories are as similar as possible between the human and the robot within the constraints of the NAO robot platform. Figure 1 shows a gesture produced by one of the targets, and the equivalent gesture on the NAO.

B. Design and Tasks

In order to evaluate personality perception by independent judges, we recorded a number of targets with a range of personalities, behaving as naturally as possible. In order to engender communicative behaviour containing suitable personality cues, we developed three tasks based upon a subset of those used by Borkenau et al. [10]. Each of the three tasks was framed as an interaction with the experimenter who stood beside the video camera used in the recordings, and provided non-verbal feedback and prompt questions to ensure as natural communicative behaviours as possible. Targets were instructed to speak for as long as they felt able, with a maximum time of 2 minutes for each task. The majority of the targets talked for 30-60 s on each task, with occasional prompts for missing information. Prior to performing tasks, we asked the targets to introduce themselves and give some information about themselves, e.g., where they work, what they do, their family, etc. This stage was purely to help naturalise the target to the experimental setting, i.e., it was not used to produce clips for judge rating.

Task 1 (Hobby): This task asked targets to describe one of their hobbies, providing as much detail as possible. Suggested detail included what their hobby involves, why they like it, how long have they been doing it for, etc. Example personality cues we anticipated from this task include what targets have as their hobby, and what detail and the depth of detail they provide while describing their hobby.

Task 2 (Story): This task is based on Murray's thematic apperception test (TAT), where the target is shown a picture and is asked to tell a dramatic story based on a picture [26]. They are asked what is happening in the picture[3], what are the characters thinking and feeling, what happens before the events in the picture and what happens after. The picture is purposely designed to be ambiguous so that the target has scope to interpret the picture as they see fit, and has to be creative in their story telling. It is a projective test, where the details given by the target, and how they relate the actions of the characters, provide cues about their personality.

Task 3 (Mime): This task required the targets to mime preparing and cooking a meal of their choice. This was different from the mime task used by Borkenau et al. [10], where targets had to mime alternative uses for a brick. Our pre-tests showed little variability between targets for that task. Instead, the chosen task gave the desired variability, and the gestures were better suited to performance on the NAO robot. Which meal was selected, and the complexity of the mime, are example personality cues we anticipated from this task.

C. Participants and Self-assessment

26 participants were recorded as targets (16 female, Age: 22-49, M=30.85, SD=7.58), and gave written informed consent for their participation, they were reimbursed with a £5 gift voucher for their time. Recordings for 20 of the targets were used to create the clips used for judgements (6 targets were omitted due to recording problems). The study was approved by the ethics committee of Faculty of Environment and Technology of The University of the West of England.

After the recordings, all targets were asked to fill in a questionnaire that aims to gather a self-assesssment of personality along the widely known Big Five personality traits [27]. These five personality traits are *extroversion* (EX - assertive, outgoing, energetic, friendly, socially active), *agreeableness* (AG - cooperative, compliant, trustworthy), *conscientiousness* (CO - self-disciplined, organized, reliable, consistent), *neuroticism* (NE - having tendency to negative emotions such as anxiety, depression or anger) and *openness* (OP - having tendency to changing experience, adventure, new ideas). In our experiments, we used the BFI-10 [28] to assess the Big Five personality traits, each trait is measured using a pair of items scored on 10-point Likert scales.

[2]Product of http://polhemus.com/

[3]Image used was https://www.flickr.com/photos/bassclarinetist/, used under creative commons licence.

D. Recordings and Clips

All tasks were recorded by one RGB video camera and and one motion capture system used for tele-operation. The recorded motion capture data was then used to produce robot mediated versions of the targets' performances on the NAO robot using the aforementioned tele-operation system, which were also recorded on video.

The video showing each target was edited into separate clips for each task, hereafter the audio-visual condition. Video clips of the robot mediated performances of the tasks were also created, hereafter the tele-operated robot condition (these used the same audio as the audio-visual condition to ensure consistency). The audio from hobby and story tasks of the audio-visual condition was also used to create audio only clips, hereafter the audio-only condition. Hence, each target had a total of 8 clips split over 3 communication conditions: 3 clips for the audio-visual condition, 2 clips for audio-only condition, and 3 clips for tele-operated robot condition. This resulted in a total of 158 clips (two clips became corrupted).

To avoid confusion, prompt questions were edited out of the clips. Further, for the few tasks where performance exceeded 60 s, clips were edited to be close to this length as pre-tests showed a decrease in the reliability of judgements with overly long clips. Mean clip duration was 50 s (SD=20 s).

E. Personality Judgements

Employing participants (e.g., judges) via online crowd-sourcing services has recently gained popularity due to its efficiency and practicality as it enables collecting responses from a large group of people within a short period of time. Similarly to [29], we used a crowd sourcing service (i.e., CrowdFlower [30]) to have the clips assessed.

The clips were split up into jobs containing four clips: one of each task and one of the audio-only clips, each of a unique target. Communication condition was pseudo-randomised across the three tasks in each job, but always contained at least one of each communication condition.

Jobs containing different clips of the same target were launched on a time delayed basis, such that they were not live at the same time. This was seen as a sufficient precaution against repeat rating influence, as the CrowdFlower platform does not allow restricting jobs based on having done related work. Hence, the communication condition for each target can be considered as close to being varied between judges (within the constraints of CrowdFlower).

Each job consisting of 4 clips was assessed by 10 judges who filled in the BFI-10 personality questionnaire [31] (the same as used in self-assessment) for each target observed. All the jobs were restricted to 12 selected European countries, USA, Canada and South Africa, and were completed by a total of 143 workers within a duration of 15 days. 46% of workers were aged between 25 and 34 years old. The gender distribution was 34% females and 62% males.

V. Results and Data Analysis

To address the research questions introduced in Section III, we tested within-judge consistency, between-judge consistency, self-other agreement and personality shifts. We present these results with respect to different communication conditions, i.e., audio-visual (AV, targets only mediated by video), audio-only (AO) and tele-operation (TO, targets mediated by the robot and recorded on video), and different tasks, i.e., Hobby, Story and Mime. We also investigate the effects of removing low-quality judges on these results.

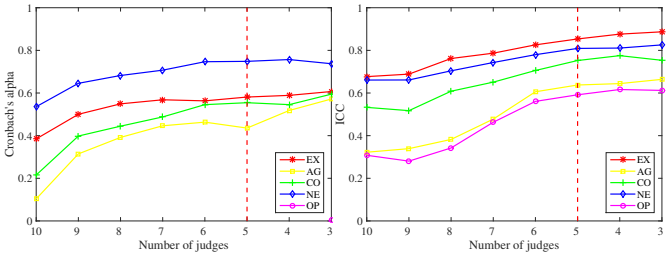

Fig. 2. Changes in Cronbach's α values (left) and ICC values (right) as a function of number of selected judges (k) for different traits for the AV condition.

A. Elimination of Low-quality Judges

Although crowd-sourcing techniques have many advantages, identifying annotators who assign labels without looking at the content (low-quality judges or spammers) is necessary to get informative results. As a first measure we eliminated judges who incorrectly answered a test question about the content of the clips. After this elimination mean-judges-per-clip was 7.9 (SD=1.5), with minimum judges-per-clip of 5.

To assess whether there remained further low-quality judges we calculated within-judge consistency for the AV clips using Cronbach's α, which measures whether the values assigned to the items that contribute to the same trait are correlated. The average value across all tasks was lower than we expected (less than 0.5), indicating some judges answer randomly. With no low-quality judges, we would expect values for the AV clips greater than 0.5, i.e., in line with values reported in the literature for the BFI-10 with video clips assessed by online judges [32]. We therefore used a judge selection method to remove these additional low-quality judges. Due to the small numbers of judges per clip, we used a ranking-based method based on pairwise correlations instead of standard methods for outlier detection. For each clip, we calculated an average correlation score for each judge from pairwise correlations (using all 10 questions in the BFI-10) with the remaining judges. Judges with low correlation scores are deemed to be spammers. The judges were then ranked in order of correlation score and the k highest ranked selected.

To evaluate the efficacy of this ranking procedure we calculated within-judge consistency results for the AV clips for different judge numbers ranging from $k = 10$ (without elimination) to $k = 3$. These values averaged over all tasks are presented in Figure 2. Selecting 5 judges per clip (based on pairwise comparisons) was found to be sufficient to increase reliability to acceptable levels for the AV clips (greater than 0.5) for all traits except for *openness*. We use 5 judges as it allows us to exclude all judges who failed the test question while having the same number of judges for all clips. Note

that 5 judges is common in this type of study, e.g., [12], [29].

B. Within-judge Consistency

Within-judge consistency was measured in terms of Cronbach's α. For the selected 5 judges per clip, the detailed results with respect to different communication conditions and tasks are presented in Table I-a, where α values that indicate sufficient reliability for the BFI-10 (greater than 0.5, in line with values reported in the literature [32]) are highlighted in bold. To compare α values between communication conditions we follow the method suggested by Feldt *et al.* [33]: 95% confidence intervals are calculated for each α value, and if the value from one condition falls outside the confidence intervals from a condition it is being compared to, this suggests it is significantly less consistent. Comparing AO with AV for the hobby task, values for all traits, except for *agreeableness*, fall outside the 95% confidence intervals of the AV values. Comparing TO with AV for the mime task, values for all traits, except for *conscientiousness*, fall outside the 95% confidence intervals of the AV values. This indicates AV is found to be more consistent as compared to AO for the hobby task (except for *agreeableness*) and TO for the mime task (except for *conscientiousness*).

C. Between-judge Consistency

We computed between-judge consistency in terms of Intra-Class Correlation (ICC) [34]. ICC assesses the reliability of the judges by comparing the variability of different ratings of the same target to the total variation across all ratings and all targets. We used ICC(1,k) as in our experiments each target subject was rated by a different set of k judges ($k = 5$), randomly sampled from a larger population of judges ($K = 143$). ICC(1,k) measures the degree of agreement for ratings that are averages of k independent ratings on the targets. Our judge selection method uses the k most correlated judges so might bias the ICC results (see Section V-A). To evaluate this we calculated ICC for $k = (10, ...3)$ for the AV condition. Figure 2-b shows that, for *extroversion*, *conscientiousness* and *neuroticism*, ICC does not change meaningfully as the number of judges varies, while selecting the 5 most correlated judges slightly biases the results for *agreeableness* and *openness*. In Section VII we discuss the implications of this biassing effect on our results analysis.

The detailed results for the selected 5 judges per clip are presented in Table I-b. We obtained significant correlations for most traits in the AV condition, with values in the same range ($0.40 < ICC(1, k) < 0.81$) as reported in the literature for online judges using a 10-item test ($0.42 < ICC(1, k) < 0.76$) [29]. Exceptions were *agreeableness* in the mime task, and *openness* in hobby and mime tasks. Fewer significant correlations were observed in the other communication conditions, particularly in the story task for AO and the mime task for TO. *Extroversion* was the only trait that consistently maintained correlation across conditions.

D. Self-other Agreement

We examined the extent to which judges agree with the target's self-assessment. Pearson correlations between the self-ratings and the judge's ratings of conditions and tasks are reported in Table I-c for the selected 5 judges per clip. We observed that the judge's ratings bear a significant relation to the target's self-ratings for *extroversion* only ($r = 0.24 - 0.44$ and $p < 0.05$). However, we did not obtain any significant correlations in the TO condition (all $r < 0.2$ and $p > 0.05$).

E. Personality Shifts

We examined the extent to which people shifted from one personality class to another, in judges' perception, between AV and TO conditions for the selected 5 judges per clip. We did not examine shifts involving AO as the ICC scores indicated that personality ratings in this condition would be too unreliable. In order to measure shifts we first classified each target into low or high (e.g., *introverted* or *extroverted*) for each trait according to if their average judge rating over hobby and story tasks (mime in the TO condition also had ICC values that were too low to use those judgements) was above or below the mean for all targets in AV. For each trait, each target was grouped according to their classification in both conditions, creating 4 groups (e.g., AV: high and TO: high, AV: high and TO: low, etc.). These results are presented in Table II as 2x2 contingency tables. To determine if the shifts were significant we used McNemar's test with Edwards's correction [35]. To aid analysis we have also illustrated each shift as a proportional change (%) both from high to low (HIGH2LOW) and from low to high (LOW2HIGH) in Figure 3. We found a significant shift from high to low for *neuroticism* (70%). Although not statistically significant, we observed large shifts from low to high for *extroversion* (56%), *conscientiousness* (67%) and *openness* (57%). Note that the corrected McNemar's test is very conservative in estimating significance, particularly for small sample sizes.

VI. Discussion

A. Judgement Consistency

Consistency within judges for how each trait is judged (Table I-a) is used to address RQ1. With the exception of *openness*, judges were sufficiently consistent in their trait ratings in the audio-visual condition (AV) for us to conclude that the tasks and judges' behaviours were reliable, when rating targets observed directly. For the hobby task, judges remained consistent in both the audio-only condition (AO) and the tele-operated robot condition (TO), indicating they were able to use audio cues to make judgements for this task, and robot appearance had no effect on consistency.

However, for the story task, judges were much less consistent in the AO than in the AV condition, for all traits except for *agreeableness*. This is in contrast to the tele-operated robot condition (TO), where they remained as consistent as in the AV condition. Gestures are the only additional cue available with the robot compared to audio only, indicating they are used to aid judgements in the same way that they do in the AV condition. This is somewhat unexpected, as in the personality literature gestures have only been previously linked to *agreeableness* [12] and *extroversion* [11]. A possible reason for this disparity is that the aforementioned studies were not able to isolate gesture in the way that we have here. Further,

TABLE I

ANALYSIS OF PERSONALITY JUDGEMENTS ACROSS 3 COMMUNICATION CONDITIONS AND 3 TASKS. (A) WITHIN-JUDGE CONSISTENCY IN TERMS OF CRONBACH'S α (ADEQUATE RELIABILITY ARE HIGHLIGHTED IN BOLD); (B) BETWEEN-JUDGE CONSISTENCY IN TERMS OF ICC(1,K) (AT A SIGNIFICANCE LEVEL OF $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$); (C) SELF-OTHER AGREEMENT IN TERMS OF PEARSON CORRELATION (AT A SIGNIFICANCE LEVEL OF $^{*}p < 0.05$, $^{**}p < 0.01$ AND $^{***}p < 0.001$).

	Audio-Visual (AV)				Audio-Only (AO)			Tele-Operation (TO)			
	Hobby	Story	Mime	All	Hobby	Story	All	Hobby	Story	Mime	All
(a) *Within-judge*											
EX	**0.64**	**0.56**	**0.63**	**0.62**	**0.57**	-0.15	0.34	**0.61**	0.39	0.19	0.47
AG	**0.54**	0.41	**0.60**	**0.52**	**0.61**	0.33	**0.52**	0.40	**0.56**	0.37	0.44
CO	0.47	**0.60**	**0.54**	**0.55**	**0.50**	0.21	0.39	**0.54**	**0.56**	**0.57**	**0.55**
NE	**0.76**	**0.76**	**0.78**	**0.78**	**0.75**	0.42	**0.63**	**0.66**	**0.54**	0.30	**0.50**
OP	-0.6	0.05	0.22	-0.04	-0.14	0.12	0.05	0.17	-0.24	-0.14	-0.07
(b) *Between-judge*											
EX	0.84***	0.81***	0.74***	0.81***	0.72***	0.51*	0.70***	0.72***	0.63**	-0.12	0.66***
AG	0.46*	0.61**	0.40	0.55***	0.25	-0.15	0.32	0.21	0.54**	-0.95	0.39**
CO	0.78***	0.67***	0.71***	0.72***	0.37	-0.10	0.22	0.32	0.65***	-0.35	0.36*
NE	0.80***	0.71***	0.55**	0.75***	0.57**	0.12	0.55***	0.70***	0.36	-0.56	0.44**
OP	0.12	0.67***	0.40	0.52***	0.49	0.40	0.55***	0.34	0.17	0.04	0.36*
(c) *Self-other*											
EX	0.34***	0.32**	0.26*	0.30***	0.44***	0.01	0.24***	0.12	-0.02	0.04	0.05
AG	0.04	0.13	0.04	0.07	0.28**	-0.05	0.12	0.08	-0.01	0.10	0.06
CO	-0.17	0.09	0.16	0.03	0.13	-0.13	0.01	0.05	0.16	-0.16	0.01
NE	0.00	-0.07	0.05	-0.01	0.07	0.09	0.07	0.02	-0.08	0.04	0.00
OP	0.06	0.03	0.00	0.03	0.10	0.04	0.07	0.16	0.07	0.03	0.09

TABLE II

CONTINGENCY TABLES FOR EACH TRAIT (AT A SIGNIFICANCE LEVEL OF $^{*}p < 0.05$)

EX	TO: high	TO: low
AV: high	16	**6**
AV: low	**10**	8

AG	TO: high	TO: low
AV: high	16	**11**
AV: low	**5**	8

CO	TO: high	TO: low
AV: high	13	**9**
AV: low	**12**	6

NE	TO: high	TO: low
AV: high	6	**14***
AV: low	**1***	19

OP	TO: high	TO: low
AV: high	13	**6**
AV: low	**12**	9

as the utility of gestures appears to be task dependent (only of apparent benefit in the story task), differences might be explained by variation in tasks used.

The use of gesture to aid personality judgements appears to be dependent on it accompanying speech, as ratings in the TO condition are far less consistent than in the AV condition for the mime task. That is to say, gestures alone do not provide sufficient information for judging personality. Either other behaviour cues not transmitted by the robot are needed, or appearance cues are used which conflict with gesture cues in TO condition.

In addressing RQ2 we found consistency between judges (Table I-b) bears some similarities to the findings of consistency within judges. The results for the audio-visual condition reinforce our assertion that our tasks and judging method are reliable, and we can observe distortions by using the AV condition as a baseline.

Although lower than in the AV condition, there is agreement for *extroversion* and *neuroticism* in the hobby task for AO and TO, indicating that audio cues alone are enough for judges to give consistent ratings for these traits. Similar values in AO and TO indicate that the robot appearance has no effect on agreement when making these judgements. In the story task we observed that judges were in agreement for *agreeableness* and *conscientiousness* in the TO condition, in contrast to the lack of agreement for those traits for the same task in the AO condition. This again suggests gesture cues are being used to make these judgements. *Agreeableness* has been previously linked to gestural cues [12], though *conscientiousness* has not. It is important to note that the correlations for the story task are weaker for *extroversion* and *agreeableness* in the TO condition than in the AV condition, indicating that the cues are less consistently interpreted by the judges viewing the robot.

For the mime task judges were unable to provide a consistent rating for any trait in the TO condition, in contrast to the consistent ratings for *extroversion*, *conscientiousness*, and *neuroticism* in the AV condition. There are likely two contributing factors: firstly, the lack of availability of physical cues, resulting in judges being forced to guess; secondly, judges do not have consistent stereotypes relating to robot appearance, while they do for human appearances as observed in the AV judgements of targets [36].

Taken together, the findings from both consistency measures indicate that when vocal cues are present a robot appearance does not hinder making consistent judgements. Indeed, robot appearance seems to be ignored and judges try and use the verbal cues to make their judgements, though these are insufficient in many cases. Hence, multi-modal communication cues (i.e., speech and gesture together), even when presented via robot mediation, have a more common interpretation by judges than speech cues alone. On the other hand, in the absence of speech, arm movements alone are not sufficient cues to make judgements, and it seems likely judges use the robot appearance in this case which can conflict with the gestural cues.

B. Accuracy of Judgements

In order to assess RQ3 we analysed the extent to which judge ratings correlated with self ratings provided by target participants (Table I-c). In general there was very little correlation between self and other ratings. This is in contrast to previous findings where they found low, but significant, self-other correlation $(0.11 - 0.42)$ [16]. This suggests that participant targets did not present cues relating to their self-perception in the tasks we used.

However, we did find self-other correlation for *extroversion* in the AV condition. *Extroversion* is commonly reported as the trait with the most available cues; hence, we use the *extroversion* judgements to draw some tentative conclusions on RQ3 as follows. Audio cues were sufficient for this correlation to be maintained in the hobby task in the AO condition, but not in the story task. In the TO condition judges were no longer able to provide any accuracy for *extroversion*, indicating that the robot appearance altered their judgements.

C. Personality Shifts

In order to address RQ4 we analysed the extent to which targets' personalities shifted from low to high, or high to low classifications for each trait between audio-visual (AV) and tele-operated robot (TO) conditions (Table II). The only consistent shift observed was for *neuroticism*, targets were perceived as less neurotic when controlling the robot. This suggests that the NAO robot gives strong appearance based cues of low *neuroticism*, and these overshadowed the targets' vocal cues, shifting the perceived personality of the operator. Though similar trends were observed for other traits, sample sizes were too small to evaluate this effectively. However, it can be clearly seen there are changes in personality perception when communication is robot mediated for around half the operators, even if these are not applied to different targets as consistently as we might expect. Different people produce different cues, and thus how these cues interact with those relating to the robot's appearance appears to be dependent not only on their unmediated perceived personality, but what cues they produce. Indeed, an important factor in thin slice personality analysis is how easy a person is to judge, which includes the quality of the cues they produce [23].

VII. Limitations and Future Work

While this paper provides evidence for how personality perception is affected for people tele-operating a humanoid robot avatar, it has a number of limitations we hope to address in future work. To more precisely identify how different cues are utilised in the aforementioned personality perception, we intend to analyse in-depth the behaviours of targets relative to their judged personality. To facilitate this we aim to apply automated personality classifiers to our data (such as [29], [37]), which can extract and identify useful cues automatically. This process is also likely to require more targets to ensure sufficient members of each personality type. Gathering additional data might also help us find more conclusive evidence to better address RQ4.

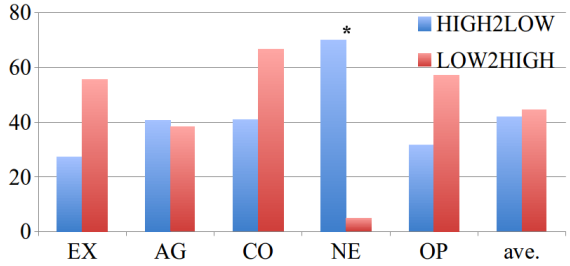

Fig. 3. Amount of shifts (%) from high to low (HIGH2LOW) and from low to high (LOW2HIGH) (* : $p < 0.05$) between AV and TO.

Another limitation in our study is that we have only used one robot platform. So while we can show that there is personality perception shifts, we can only make limited conclusions on implications for robot avatars in general.

Due to the design limitations in CrowdFlower at the time we used it, we could not enforce precautions of eliminating low-quality judges during task performance. As a result we felt it necessary to use a judge selection method based on the gathered responses. The procedure we used had a slight biassing effect on the between-judge consistency (ICC) result for *agreeableness* and *openness*. This bias means that where ICC values are not significant it is strong evidence that there is either a lack of cues or conflicting cues, as even amongst the most agreeing judges consensus of opinion was not possible. Where there is significant agreement, it indicates there are cues for that trait in the particular task and condition and some judges are able to pick up on these cues. Indeed, Funder points out that there exists good and bad judges of personality [23], and we suggest our selection method allowed us to bias toward good judges. This limits the generalisability of our results to judges more adept at picking up on personality cues.

VIII. Conclusion and Design Implications

In this paper we have shown that the appearance of a tele-operated robot avatar influences how the personality of its controller is perceived, i.e., robot appearance based personality cues are utilised along with cues in the speech of the operators. Using a large set of robot operators producing naturalistic communication behaviours we have shown that this interaction is highly operator-dependent. Behavioural personality cues differ from person to person, even for the same personality type, and this affects their interaction with robot appearance. In light of these findings we suggest that robot avatar appearance and behaviour be carefully considered relative to the person who will be controlling it, and this needs to be done on an individual basis. Training of operators to produce clear cues, or having some cues appropriate to the operator's personality autonomously generated, might allow some control of appearance effects.

Our findings also have implications for autonomous robot personality expression, as cues that work on one platform may not be transferable to another. It is important to consider what appearance cues for personality a robot has, and whether the planned behavioural cues might conflict with them.

In contrast to past work [20] we have shown that judges are able to make judgements that are as consistent with a

robot avatar as when the same people are viewed on video. Importantly, gestural cues, reproduced on a NAO robot using a motion capture based tele-operation system, improve the ease with which judges can assess personality, relative to audio-only communication. Hence, we suggest that it is important for tele-presence systems to be able to transmit gestural cues, whether this be actuation of physical systems, or large enough screens on remote presence devices.

Further, we suggest autonomous robots that wish to portray a range of personality traits can use gestural cues to do so. Our method of using a large set of naturalistic human behaviours, and more in-depth personality analysis, significantly adds to previous approaches to personality synthesis which have largely focused on introversion and extroversion, and a limited set of communications. This lends more weight to the growing body of literature on the utility of gestures in HRI (e.g. [25]), and in particular robot personality perception [38].

ACKNOWLEDGEMENTS

This work was funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref: EP/L00416X/1).

REFERENCES

[1] O. Daly-Jones, A. Monk, and L. Watts, "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus," *Int. Journal of Human-Computer Studies*, vol. 49, no. 1, pp. 21–58, 1998.

[2] B. O'Conaill, S. Whittaker, and S. Wilbur, "Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication," *Human-Computer Interaction*, vol. 8, no. 4, pp. 389–428, Dec. 1993.

[3] A. Tang, M. Boyle, and S. Greenberg, "Display and presence disparity in Mixed Presence Groupware," in *Proc. of Australasian User Interface Conf.* Australian Computer Society, Inc., Jan. 2004, pp. 73–82.

[4] 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 *IEEE Int. Conf. on Robotics and Automation*. IEEE, May 2013, pp. 184–190.

[5] S. O. Adalgeirsson and C. Breazeal, "MeBot: A robotic platform for socially embodied telepresence," in *Proc. of Int. Conf. Human Robot Interaction*. ACM/IEEE, 2010, pp. 15–22.

[6] C. Bevan and D. Stanton Fraser, "Shaking Hands and Cooperation in Tele-present Human-Robot Negotiation," in *Proc. of Int. Conf. Human Robot Interaction*. ACM/IEEE, Mar. 2015, pp. 247–254.

[7] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance." *Personality & social psychology bulletin*, vol. 35, no. 12, pp. 1661–71, Dec. 2009.

[8] Y. Wang, J. Geigel, and A. Herbert, "Reading Personality: Avatar vs. Human Faces," in *Proc. of HAC Conf. on Affective Computing and Intelligent Interaction*. IEEE, Sep. 2013, pp. 479–484.

[9] K. Fong and R. A. Mar, "What Does My Avatar Say About Me? Inferring Personality From Avatars," *Personality and Social Psychology Bulletin*, vol. 41, no. 2, pp. 237–249, Jan. 2015.

[10] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin Slices of Behavior as Cues of Personality and Intelligence." *J. of Personality and Social Psychology*, vol. 86, no. 4, pp. 599–614, 2004.

[11] R. E. Riggio and H. S. Friedman, "Impression formation: The role of expressive behavior," *Journal of Personality and Social Psychology*, vol. 50, no. 2, pp. 421–427, 1986.

[12] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance." *J. of Personality and Social Psychology*, vol. 62, no. 4, pp. 645–657, 1992.

[13] R. Yamazaki, S. Nishio, K. Ogawa, and H. Ishigur, "Teleoperated android as an embodied communication medium: A case study with demented elderlies in a care facility," in *RO-MAN*. IEEE, Sep. 2012, pp. 1066–1071.

[14] H. Martins and R. Ventura, "Immersive 3-d teleoperation of a search and rescue robot using a head-mounted display," in *IEEE Conf. on Emerging Technologies Factory Automation (ETFA)*, Sept 2009, pp. 1–8.

[15] D. C. Funder and C. D. Sneed, "Behavioral manifestations of personality: An ecological approach to judgmental accuracy." *Journal of Personality and Social Psychology*, vol. 64, no. 3, pp. 479–490, 1993.

[16] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *Journal of Research in Personality*, vol. 41, no. 5, pp. 1054–1072, Oct. 2007.

[17] B. S. Connelly and D. S. Ones, "An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity." *Psychological Bulletin*, vol. 136, no. 6, pp. 1092–1122, 2010.

[18] I. Rae, L. Takayama, and B. Mutlu, "In-body experiences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. New York, New York, USA: ACM Press, Apr. 2013, pp. 1921–1930.

[19] M. K. Lee and L. Takayama, "Now, i have a body," in *Proc. of the conf. on Human factors in computing systems*. ACM Press, May 2011, p. 33.

[20] K. Kuwamura, T. Minato, S. Nishio, and H. Ishiguro, "Personality distortion in communication through teleoperated robots," in *Proc of IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, Sep. 2012, pp. 49–54.

[21] I. Straub, S. Nishio, and H. Ishiguro, "Incorporated identity in interaction with a teleoperated android robot: A case study," in *Proc of Int. Symp. in Robot and Human Interactive Communication*. IEEE, Sep. 2010, pp. 119–124.

[22] T. Behrend, S. Toaddy, L. F. Thompson, and D. J. Sharek, "The effects of avatar appearance on interviewer ratings in virtual employment interviews," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2128–2133, Nov. 2012.

[23] D. C. Funder, "On the accuracy of personality judgment: A realistic approach." *Psychological review*, vol. 102, no. 4, 1995.

[24] 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 *Proc of IEEE Int. Conf. on Robotics and Automation*. IEEE, May 2009, pp. 769–774.

[25] P. Bremner and U. Leonards, "Efficiency of speech and iconic gesture integration for robotic and human communicators - a direct comparison," in *Proc. of IEEE Int. Conf. on Robotics and Automation*. IEEE, May 2015, pp. 1999–2006.

[26] H. A. Murray, *Thematic Apperception Test*. Harvard University Press, 1943.

[27] A. Vinciarelli and G. Mohammadi, "A Survey of Personality Computing," *IEEE Trans. on Affective Computing*, 2014.

[28] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german," *J. of Res. in Personality*, vol. 41, no. 1, pp. 203 – 212, 2007.

[29] J. Biel and D. Gatica-Perez, "The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs," *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 41–55, 2013.

[30] CrowdFlower, "A data enrichment, data mining and crowdsourcing company," http://www.crowdflower.com/, accessed at September 2015.

[31] O. John, E. Donahue, and R. Kentle, "The big five inventory versions 4a and 54," Ins. of Personality and Social Research, Tech. Rep., 1991.

[32] M. Credé, P. Harms, S. Niehorster, and A. Gaye-Valentine, "An evaluation of the consequences of using short measures of the Big Five personality traits." *J. of personality and social psychology*, vol. 102, no. 4, pp. 874–88, Apr. 2012.

[33] L. S. Feldt, D. J. Woodruff, and F. A. Salih, "Statistical Inference for Coefficient Alpha," *Applied Psychological Measurement*, vol. 11, no. 1, pp. 93–103, Mar. 1987.

[34] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychology Bull.*, Jan. 1979.

[35] A. L.Edwards, "Note on the correction for continuity in testing the significance of the difference between correlated proportions," *Psychometrika*, vol. 13, no. 3, pp. 185–187, 1948.

[36] D. A. Kenny, L. Albright, T. E. Malloy, and D. A. Kashy, "Consensus in interpersonal perception: acquaintance and the big five." *Psychological bulletin*, vol. 116, no. 2, pp. 245–58, Sep. 1994.

[37] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE. Trans. on Affective Computing*, 2016.

[38] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *Proc. of Int. Conf. Human Robot Interaction*. ACM/IEEE, Mar. 2013, pp. 325–332.