Personality-Adapted Language Generation for Social Robots

Alessio Galatolo[1,*], Iolanda Leite[2] and Katie Winkle[1,*]

***Abstract*— Previous works in Human-Robot Interaction have demonstrated the positive potential benefit of designing social robots which express specific personalities. In this work, we focus specifically on the adaptation of language (as the choice of words, their order, etc.) following the extraversion trait. We look to investigate whether current language models could support more autonomous generations of such personality-expressive robot output. We examine the performance of two models with user studies evaluating (i) raw text output and (ii) text output when used within multi-modal speech from the Furhat robot. We find that the ability to successfully manipulate perceived extraversion sometimes varies across different dialogue topics. We were able to achieve correct manipulation of robot personality via our language adaptation, but our results suggest further work is necessary to improve the automation and generalisation abilities of these models.**

I. Introduction

Personality can be defined as the set of characteristics that influences many aspects of humans' lives: friendship development [1], work performance [2], etc. Previous works in Human-Robot Interaction (HRI) have demonstrated the positive potential benefit of designing social robots which express a specific personality [3], [4], [5], [6], [7], [8], [9]. This has resulted in numerous attempts to identify mechanisms, typically grounded in psychology literature, that can support HRI designers in making robot behaviour personality-expressive. These range from using hand-written 'rules' or heuristics [10], [11], [12], [13], [14], [15], [16] to more autonomous extraction and generation methods [17], [18]. Many such works focus only on body language [6], [19], but there is evidence in the literature that personality has a significant influence on other aspects of interaction, including language [12], [13], [14], [15], [16]. By language, we specifically refer to the choice of words, their order and how they are used to formulate sentences to express certain or multiple ideas. For example, a very extraverted person is generally more talkative and uses less complex sentences than an introverted one [12], [13]. Relying on expert/hand-written personality adaptations of robot language, as often used in HRI experiments examining the impacts of robot personality expression [7], [9], is resource expensive, and not long-term sustainable for large-scale, real-world robot deployment. For HRI researchers with a non-psychology background, this can also pose a significant challenge during Wizard-Of-Oz experiments in which the researcher is supposed to generate personality-congruent speech for a particular interaction, potentially even doing so on the spot/in real-time. A method for automatically adapting robot language would therefore provide a useful tool for HRI researchers conducting interaction studies involving unplanned, dynamic robot-participant interactions, and move us closer towards robots which can generate specific, personality-expressive behaviours in the wild.

We adopt the Big Five personality framework [20], [21] which aims to describe personality based on the five traits of Openness (to experience), Conscientiousness, Extraversion, Agreeableness and Emotional (in)stability[1]. This is opposed to other frameworks such as the Myers–Briggs Type Indicator [22] and other non-trait-based personality frameworks or even frameworks specifically designed for conversational agents [23]. The choice of the framework was mainly dictated by the proven reliability of the Big Five framework, whereas MBTI is heavily criticised [24], [25] and the work of Vökel et al. [23] is aimed towards agents with the role of *assistants*.

Furthermore, similar to previous studies [4], [8], [19], we focus our attention only on the extraversion trait with the intention of extending and re-evaluating our methods in future works, noting how the choice of our methods does not hinder in any way an extension to the whole spectrum of personality.

A. Related Works

Mairesse et al. [17], [18] propose a method called PERSONAGE to generate language according to different personality traits. In order to achieve this, they use linguistic parameters such as verbosity and length of the sentence, well-grounded in psychology literature, to manipulate a text that can be attributed different personality traits. The authors use this method to generate utterances about restaurant recommendations and comparisons in New York City modelled following different personality traits selected on a scale from 1 to 7. This method relies on the collection of a large context-specific dataset which hence does not support the longer-term aim of developing a more universal language adaptation system that can be used across different contexts. Aly et al. [6] combined the PERSONAGE language generator with the Behavior Expression Animation Toolkit (BEAT) [26] to animate the movements of the humanoid robot NAO. They show that incorporating gestures increases the overall engagement and effectiveness of the robot. Furthermore,

[1]Alessio Galatolo and Katie Winkle are with the Department of Information Technology, Uppsala University, Uppsala, Sweden [alessio.galatolo / katie.winkle]@it.uu.se

[2]Iolanda Leite is with the Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden iolanda@kth.se

*Work done while at KTH Royal Institute of Technology.

[1]Historically referred to as Neuroticism.

they identified a preference for personality matching (in the extraversion trait) with the user. Their results further motivate the manipulation of robot personality but also share those same constraints as the original PERSONAGE work.

II. Methods

Our ultimate aim would be to have a system which can take any starting text and autonomously adapt it to express a particular personality ready for use in robot speech. We propose two methods, both based on text style transfer, an approach with the goal of changing the 'style' of a given sentence, often by changing the language (as per our previous definition) while keeping the semantics of the sentence intact. Each method takes an initial dialogue and generates two versions[2], one that is more introverted and the other that is more extraverted. In order to examine how our language adaptation methods might support (autonomous) multi-modal behaviour generation, we feed our adapted texts into an emotion recognition system, which identifies emotions in the dialogue. The different versions of the text and the associated emotion labels can then be input into any robot which has in-built/pre-designed categorical emotional expression capabilities (e.g. Furhat, Pepper, Nao, Cozmo). In this work, we specifically utilise the Furhat robot as an exemplar social robot platform. Furhat is rated to be one of the most anthropomorphic robots commercially available to researchers (cf. the ABOT Database[3] [28]) thus we expect it might require high personality-expressiveness *and* congruent and appropriate emotional expression in order meet user expectations and minimise the risk of uncanniness [29], [30], [31]. Given our choice of Furhat, we specifically target the generation of congruent facial expressions that should accompany the robot dialogue speech, but this could also be applied to other non-verbal cues such as gesture, pitch and tone.

A. Style Transfer Method 1: STRAP

The first model we tested for personality adaptation is Style Transfer via Paraphrasing (STRAP) from [32], where the authors tackle the problem of text style transfer as a paraphrasing task. This method expects the input to be first paraphrased by a style-neutral model in order to get a different formulation of the sentence. Then, the output is fed into a style-specific model that has been trained on only one particular style. The output of the second model should follow the style it was trained on. The starting model used for all paraphrasing, both style-neutral and style-specific, is a pre-trained version of GPT-2 [33]. We chose to use this method, which is based on the old GPT-2 large [33] rather than newer models such as BART [34], T5 [35], and others, according to (i) its proven success on different style transfer tasks and (ii) the ease of training.

[2]As text style transfer generally concerns binary style e.g. polite vs impolite, formal vs informal, etc. [27] we do not aim at having different degrees of personality expression.

[3]http://www.abotdatabase.info/collection

1) Training and dataset: We trained the model using the code provided by the original authors[4] with no meaningful changes. The dataset we used [12] is made up of 2467 stream-of-consciousness essays written by psychology students that were classified into binary Big-Five personality traits through a self-report questionnaire. We split each essay into single sentences following punctuation. Each sentence was labelled with the original extraversion binary trait of the essay's author and we excluded sentences with more the 50 words (constraint given by the method). The processed dataset comprehends 91359 sentences where 45295 of which are from extraverted people ($\sim$50%). During our testing, we trained only the style-specific model and not the style-neutral one; relying here instead on the one trained by the original authors.

2) Performance and parameters: The inference time is high in both time and resource consumption. Our tests suggest an average inference time for an example dialogue ($\sim$180 words) of $\sim$6s ($<$ 1s per average sentence) on a Nvidia GeForce RTX 2080 Ti GPU. Both the style-neutral and style-specific models have a parameter 'top-p' that can be tweaked during inference to achieve different results. The top-p value acts as a restriction to the low-probability tokens (words). A high top-p value would create more diverse responses but with an increased risk of grammatical, syntactical or semantic errors. When using the style-neutral model, we kept a top-p value of 0 (as advised by the original authors) whilst exploring the whole range of values from 0 to 1 in the style-specific model. In initial testing, we found the model to perform best when given single sentences rather than whole dialogues or groups of sentences. We also observed that differences between the two generated versions of the sentence only start to emerge once the top-p value is 0.5 or higher. As we approached the very high values of top-p (> 0.8) we observed the output starting to significantly deviate from the original input, in cases appearing as a succession of random words. Please refer to the online supplementary material[5] for a detailed rundown of the model performance and outputs.

B. Style Transfer Method 2: GPT-3

The second method we tested for text style transfer was GPT-3 [36] (text-davinci-002[6]) using few-shot learning. Models such as GPT-3 have generally proved to be easily adaptable to different domains, often requiring only a broad description of the task (zero-shot learning) or a few examples (few-shot learning) in order to be able to achieve very good results. As such, they have also been shown effective in the problem of text style transfer through zero-shot learning [**?**]. We did not focus our efforts on prompt engineering but rather

[4]https://github.com/martiansideofthemoon/style-transfer-paraphrase

[5]https://github.com/alessioGalatolo/PersonalityLanguageGeneration/blob/main/Appendix.pdf

[6]This work has been conducted before the release of text-davinci-003, gpt-3.5-turbo (ChatGPT) and GPT-4.

experimented with simple prompts for zero-shot and one-shot learning while also varying whether we provided single sentences, whole dialogues or topic-grouped sentences as input. Through this (ad-hoc/exploratory) testing we achieved the best results by using one-shot learning, splitting the input into short sentences and using the PERSONAGE dataset [17], [18] as the source of examples given. The prompts used can be found in full online.

1) Implementation and performance: Unlike STRAP, GPT-3 is not open-source and is accessible only through the public API. The API grants access to the prompt-based inference of the model. The model runs directly on internal servers, such that processing is not transparent to the user. The API also offers a way of fine-tuning the model, but this was not explored in this work.The resource consumption is also quite difficult to estimate given that processing is done on an external server, but it's a reasonable assumption that the inference time would be quite higher compared to the previous model.

C. Emotion Recognition from Text

We propose the use of automatic emotion classification to identify the emotion expressions within our personality-expressive dialogues in order to generate congruent emotional labels for input to our robot platform. We utilised an off-the-shelf model provided via the HuggingFace platform [37]. The model [38] is a fine-tuned version of RoBERTa-LARGE [39] and takes a sentence as input and outputs a score from 0 to 1 for the six Ekman basic emotions [40], plus one for emotion-neutrality. All the scores add up to 1. For Furhat, we took each emotion with its score and generated an appropriate gesture[7] (facial expression) with an intensity proportional to the score (halved to avoid unnaturally exaggerated expressions and uncanniness [31]). The 'gesture' lasted throughout the sentence from which it was extracted.

III. Evaluation

We undertook two user studies to evaluate the performance of our methods. Study 1 is concerned with evaluating the output of the two different style transfer models, looking for differences in personality expression through language by assessing their text-only outputs. Study 2 is concerned with evaluating the outputs when incorporated into the robot Furhat. This allows us to examine whether a spoken delivery of the same text may have an influence on its perception and whether the manipulation of personality or the variation in the model used have an impact on overall perception of the robot.

A. Study 1: Text-Only Model Evaluation

The first study we conducted was aimed at assessing the performance of the models presented, both in terms of their ability to convey the right personality and their fluency, a common practice when evaluating Natural Language Generation (NLG) models.

We designed three dialogues in three different contexts (available in full online) and asked participants recruited online to rate them on selected measures.

The first dialogue (d_1) is an introduction from a companion robot[8], the second one (d_2), originally from [41], represents a typical, health-related socially assistive robot (SAR) interaction and the third (d_3) represents a typical, education-related SAR presenting a specific topic[8] (humanities).

1) Study design: This study was carried out with a mixed design where each participant saw the three different dialogues (in random order), each generated either by STRAP, GPT-3 or hand-written by the authors (a control condition). This resulted in each participant seeing only one version of each dialogue, from one or more different models with the same or different personality manipulation.

We recruited through the platform Prolific[9] 30 participants[10] for each condition (dialogue$\times$model$\times$personality combination) for a total of 180 people aiming for men/-women balance per Prolific screening tools. Each participant saw the 3 different dialogues and, after each dialogue, they were asked to answer 14 questions posed in a randomised order each time. 10 of the 14 questions were aimed at assessing the perceived personality, 3 for the fluency and one was an attention check. 2 of the 10 personality questions were taken from PERSONAGE paper [18] and 8 from another HRI study on robot personality [4][11] (questions originally from [43]). The fluency questions were designed by us following a survey on the evaluation of NLG models [44]. All the questions were posed on a 7-point Likert scale and can be found in full online.

As an additional attention check to the 14th question, each participant was asked, at the end of the study, to select the topic of the dialogues they just read among 6 possibilities. The participants received a £1.20 compensation for completing the survey.

2) Generation of dialogues: For STRAP generations we used a top-p value of 0.7. Given that our primary goal is to test whether the model is actually capable of changing the personality in a dialogue, we chose to use a high top-p value that should improve the results of the personality transfer task. However, this choice worsens the performance in terms of fluency and closeness of the paraphrased sentence from the original one. To account for this, for each dialogue, we generated 10 outputs and discarded those that contained grammatical errors or that radically changed the content of the sentence. We are confident that, in the near-term future

[7]We used Furhat default facial expression for anger, disgust, fear, surprise and sadness in addition to a custom 'joy' option which combines a smile, cheek puff and openness of the jaw.

[8]The dialogue was designed with the aid of GPT-3 and then tweaked by hand.

[9]`https://www.prolific.co/`

[10]Study design, experimental protocol and data collection were conducted in line with local (Swedish) ethics regulations and guidelines.

[11]Originally only 3 out of the 8 questions measured the introversion while 5 measured the extraversion. To balance these numbers, we swapped the question for extraversion 'has an assertive personality' with a similar one for introversion 'holds back their opinion' taken from the extraversion part of The Items in the Big Five Aspects Scales, IPIP [42].

development, this selection could be done using automated metrics e.g. BLEU (BiLingual Evaluation Understudy) [45] for paraphrasing accuracy and a simple spell-checker for grammatical correctness.

For the dialogue generated with GPT-3, we generally utilised the first output of the model with the exception of some sentences in the second dialogue. The reason for this lies in the poor performance of the model when style transferring a question. For example, when asked to transfer the question "How do you feel about being here today?" the model would favour answering the question rather than changing its style. The only way we could find of making the dialogue progress as the others was to look for a generation where the model would answer the question and *then* ask what the other person thought e.g. "I don't really feel comfortable being here today. Do you?".

3) Results: Among the 180 participants, we excluded 39 due to failed attention checks. Among the remaining participants, 67 identified as women, 72 as men and 2 as non-binary. The average age was 29 ($M = 28.965, SD = 8.262$) with the majority reporting a medium knowledge of English (98 reported being "Comfortable enough to understand English in most cases"[12], the others reported being native speakers). All the statistical analysis that follows was done using the open-source software JASP v16.2 [46]. The data collected and the scripts used for the analysis can be found in our Repository[13] under text_study_data .

We evaluate our results on 2 key measures: fluency and extraversion. The former is the mean[14] score of the fluency questions ($\alpha = 0.81$) while the latter is given by the mean of the extraverted questions ($\alpha = 0.88$) and the introverted ($\alpha = 0.83$) ones (with the score reversed). Using this measure, an introverted dialogue should score low on this scale while an extraverted one should be high.

Figure 1 shows the extraversion scores plotted by model and divided by dialogue.

The figures indicate that the GPT-3-generated and our hand-written outputs were successful in differentiating all three dialogues based on personality. STRAP was only successful on the first two. To confirm this, we performed a series of independent samples t-tests or Mann-Whitney U test (if the normality assumption was violated) to check whether the difference in extraversion was significant between the two generated versions of each dialogue. We report the results in Table I, all the significant results have a small to medium effect size ($> 0.2, < 0.5$, Cohen's d or Rank Biserial r_B) except the conditions 'GPT-3$\times d_2$' and 'Expert$\times d_2$' that have $r_B = 0.758$ and $d = 0.796$ respectively and 'Expert$\times d_3$' that has $d = 1.127$.

On fluency, our Expert (hand-crafted) version ($M = 4.494, SD = 1.176$) and GPT-3 ($M = 4.557, SD = 1.009$) were rated best, with STRAP being rated much lower ($M = 3.497, SD = 1.469$). A Krukall-Wallis test (done in place of ANOVA due to violation of the equality of variances: Levene's, $F(17, 405) = 2.794, p < 0.001$) revealed a significant difference in fluency between models: $H(2) = 46.850, p < 0.001$. A post-hoc Dunn's test confirmed GPT-3 and Expert being more fluent than STRAP with $p < 0.001$ for both.

Condition	Normal	Test	p-value
STRAP $\times$ d_1	✓	$t(45) = 1.929$	0.03
STRAP $\times$ d_2	✓	$t(47) = 2.267$	0.014
STRAP $\times$ d_{all}	✓	$t(141) = 1.736$	0.042
GPT-3 $\times$ d_2	✗($p = 0.019$)	$W = 424.5$	< 0.001
GPT-3 $\times$ d_{all}	✗($p = 0.006$)	$W = 2805$	0.001
Expert $\times$ d_1	✓	$t(45) = 1.504$	0.07
Expert $\times$ d_2	✓	$t(46) = 2.755$	0.004
Expert $\times$ d_3	✓	$t(52) = 4.115$	< 0.001
Expert $\times$ d_{all}	✗($p = 0.008$)	$W = 3961.5$	< 0.001

TABLE I: Summary of the significant and close to significance results for difference in extraversion in the text-only study.

B. Study 2: Robot Output Evaluation

Having established that our language models were seemingly able to manipulate personality as desired, we designed this follow-up study to evaluate the perception of a robot utilising the output of our language adaptation and emotion expression generation process.

1) Choice of dialogue: For this study, we decided to use the first dialogue of those tested with the text-only study given its "success" in differentiating one personality from the other in all of our models. Further, we only used the default text-to-speech available in the Furhat SDK and we did not control for the prosody of the speech.

2) Study design and measures: We designed a between-subjects, video-based online study. We recorded 6 videos varying our experimental conditions: the model used for the personality adaptation (GPT-3, STRAP and Expert) and the personality (Int or Ext) for a 3x2 study design. The clips were shot using the Virtual Furhat SDK rather than recording a real Furhat to favour the correct viewing of all of the facial expressions (used to convey the emotions) of the robot[15]. The videos are available for watching under video_study_data / videos. We also show a screenshot of one of the clips in Figure 2.

We used Prolific to recruit 180 people, 30 participants per condition, aiming for an equal distribution of men and women using Prolific screening tools. Each participant in the study was initially asked to compile a short questionnaire to assess their personality (used to check for personality matching). The questionnaire has been extracted from the one used in Study 1 where we removed the questions on fluency and the questions explicitly asking the introversion/extraversion. Each participant then watches a video with the robot animating the first introduction dialogue from Study 1 and

[12]Independent samples t-tests were conducted to confirm that participants' fluency in English did not affect their perception of personality.

[13]https://github.com/alessioGalatolo/PersonalityLanguageGeneration/

[14]Before calculating the mean, we confirmed internal consistency with Cronbach's alpha test.

[15]Recording a real Furhat often worsens the quality of the face and differences in brightness/contrast in the scene can worsen the correct reception of its facial movements.

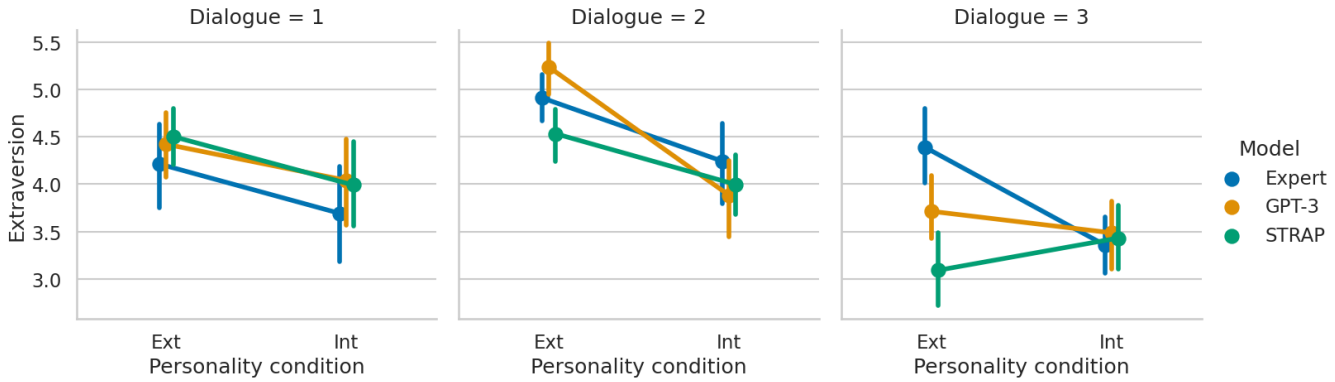

Fig. 1: Extraversion rating for each model×personality combination across each of our three different dialogues, shown on a scale from 0 to 6. Better as the Ext score gets higher than the Int score. Error bars are given by confidence interval.

Fig. 2: A screenshot of one of the videos used for the study.

manifesting congruent emotional expressions. The video was available to re-watch at every step of the questionnaire. For a direct comparison with Study 1, the participant is first asked to rate the same questions, again, on a 7-point Likert scale. We also integrated questions from the Godspeed questionnaire [47] for the measures of Anthropomorphism, Likeability and Perceived Safety and questions of Warmth and Discomfort from the RoSAS questionnaire [48], all presented on a 5-point Likert scale. We report these additional questions online.

We use Godspeed Anthropomorphism and Perceived Safety, RoSAS Warmth and Discomfort as a proxy for uncanniness, however, we analyse these separately as they aim to measure different aspects of uncanniness. We also use the questions from RoSAS Warmth as a measure of emotional expression.

We included in the survey different types of attention checks, 3 of them were among the questions. We also asked each participant at the end what was the name of the robot and the topic of the dialogue. The choice was among 6 possibilities for each. Participants received a £1.50 compensation for completing the survey.

3) Demographics: Among the 180 recruited participants, we excluded 23 due to failed attention checks. The remaining population was composed of 80 women, 84 men, 2 non-binary people and one that did not self-identified among our options (chose 'other'). The average age is 27 ($M = 27.234$, $SD = 8.02$).

35 participants reported being native speakers of English while 131 reported being "Comfortable enough to understand English in most cases"[12]. Also, following our self-report questionnaire on personality, 90 participants scored higher on extraversion than introversion, compared to 77 for whom it was the other way around.

The data collected and the analysis done are available under video_study_data .

4) Difference in extraversion and fluency: Our first analysis was aimed at analysing the fluency (Figure 3, centre) and verifying the correct manipulation of the personality (Figure 3, left) in the dialogues i.e. the extraverted dialogue is perceived as significantly more extraverted than the introverted one. Similarly to what was done in the first study, to compare the dialogues we use the extraversion score which is an average of the extraversion ($\alpha = 0.81$) and reversed introversion ($\alpha = 0.74$) questions. The fluency measure is, again, the mean of the three questions in the questionnaire ($\alpha = 0.82$).

Starting with the extraversion measure we run, for each of our models, independent samples t-tests or Mann-Whitney U tests if the assumption of normality (Shapiro-Wilk test) was violated. We also report Cohen's d or Rank Biserial r_B as the effect size. All of our models seemingly manipulated the personality as desired, resulting in significantly higher perceived extraversion in the Ext condition compared to the Int one: $W = 512.5, p = 0.045, r_B = 0.262$ (not normal, $p = 0.016$) for the Expert condition, $t(110) = 2.425, p = 0.008, d = 0.542$ for GPT-3 and $W = 1966, p = 0.01, r_B = 0.262$ (not normal, $p = 0.046$) for STRAP.

We run a Kruskall-Wallis test, in place of ANOVA due to the violation of the equality of variance (Levene's, $F(5, 161) = 2.613, p = 0.027$), to test the difference in perceived fluency across all of our conditions. The test revealed a significant difference in perceived fluency between models ($H(2) = 7.788, p = 0.02, \eta^2 = 0.041$), personality condition ($H(1) = 5.864, p = 0.015, \eta^2 = 0.044$) and model×personality ($H(5) = 14.089, p = 0.015, \eta^2 = 0.084$). A post-hoc Dunn's test revealed a difference between personality conditions where the Ext condition was perceived as significantly more fluent than the Int one ($p = 0.008$ with Holm's correction). Also, GPT-3 was significantly more

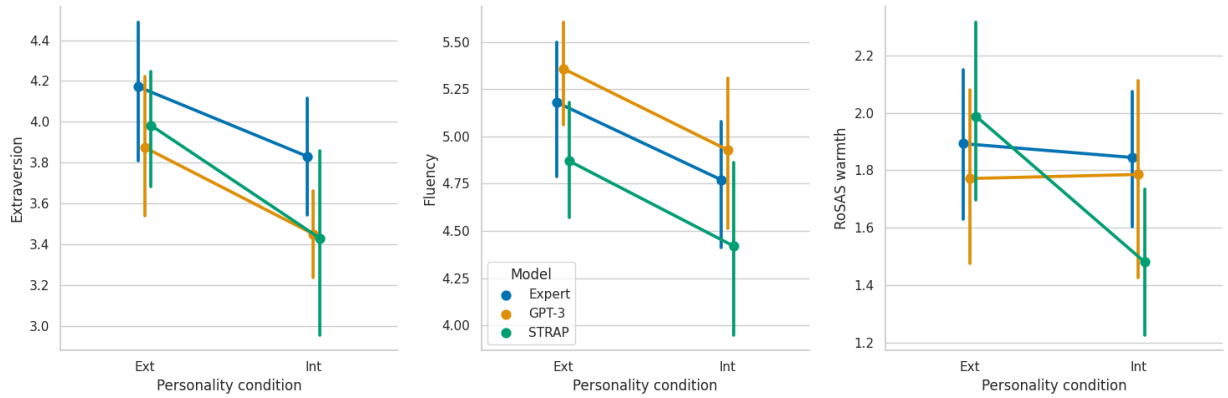

Fig. 3: Plots with the ascribed extraversion (left), fluency (centre) and RoSAS warmth (right). First two on a scale from 0 to 6, last one from 0 to 4.

fluent than STRAP ($p = 0.008$ with Holm's correction) with GPT-3 Ext being significantly more fluent than STRAP Int ($p = 0.011$ with Holm's correction).

5) Text-only vs multi-modal: We also compared the data collected in this study with that of the first one (considering only the first dialogue which is the one they share) in order to check for any difference in the perceived extraversion and fluency. We can see in Figure 4 (left) how both our

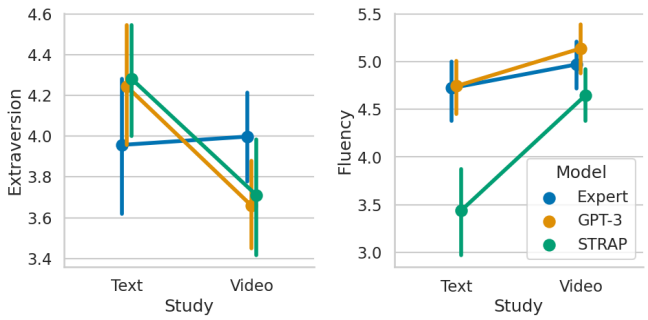

Fig. 4: Plots with the ascribed extraversion (left) and fluency (right) comparing text-only and video-based study.

automated models have a decrease in perceived extraversion (although the difference between the two personality adaptations is maintained) on being incorporated into a robot, while our Expert condition does not. We confirmed this with independent samples t-tests for STRAP ($t(85) = 2.260, p = 0.026, d = 0.565$) and a Mann-Whitney U test (not normal, $p = 0.006$) for GPT-3: $W = 1134.5, p = 0.001, r_B = 0.378$.

In Figure 4 (right) we also notice how all models experience a jump in perceived fluency, with STRAP gaining most of all. This increase in perceived fluency is significant for both STRAP and GPT-3, tested with Mann-Whitney U test: $W = 596, p = 0.026, r_B = 0.452$ (not normal, $p < 0.001$) for GPT-3 and $W = 513.5, p < 0.001, r_B = 0.285$ (not normal, $p = 0.007$) for STRAP.

6) Other measures: As explained earlier, in this study we also introduced measures from the Godspeed and RoSAS questionnaire to better evaluate the impact(s) of our personality adaptation approach. We did a series of ANOVAs (or Kruskall-Wallis if its assumptions were violated) to identify any differences among our conditions with, however, very few results reaching significance.

We also performed a series of independent samples t-tests within each model to compare between the two versions of the dialogue generated by it, finding only a significant difference in RoSAS warmth (our proxy for emotions) in STRAP ($t(110) = 1.943, p = 0.027, d = 0.669$) where the Ext condition was perceived as more 'emotional' than the Int one (see Figure 3, right). In STRAP, we also found a difference in Anthropomorphism with the Ext condition being perceived as significantly more anthropomorphic than the Int one ($t(53) = 1.885, p = 0.032, d = 0.508$).

Finally, given previous literature on user-robot personality (mis)matching, we run a series of tests (ANOVA/Kruskall-Wallis) to check for any significant effect of personality matching on our measures, finding no significant results.

IV. Discussion

A. Automatic Generation of Personality-Adapted Text is Possible but...

Results from both the text-only and the video-based studies indicate our models were generally successful in the manipulation of personality through language, with the exception of STRAP performing particularly poorly in one of our three scenarios. Interestingly, whilst the capability to generate two differentiable versions of a dialogue *remains intact*, going from text-only to robot-embodied speech with associated emotional expressions, decreased the overall attribution of extraversion attributed to our automated model outputs. One explanation could be that, as we posited earlier in this article, observers might expect very human-like behaviour from Furhat based on its highly anthropomorphic design. We limited ourselves to language and facial expression manipulation, but in humans, both personality and emotion have been linked also with facial/body cues or through changes in voice pitch [10], [11], [49], [50].

B. STRAP does not generalise well... but should it?

In our text-only study, we were able to show that both GPT-3 and our hand-crafted dialogue were successful in conveying the right personality in all contexts. However, the same is not true for STRAP which underperformed in the third dialogue (the talk about humanities) where it failed to create a difference in perceived extraversion between the two versions of the dialogue.

Despite one possible explanation being a simple pitfall of the method, we would like to question whether it *should* work. In fact, this dialogue, compared to the other two, is an objective exposition of a topic that should not let out much of the inherit personality of a person (robot). Furthermore, if we look closely at the dialogue versions of both GPT-3 and Expert (see online) we can notice that some additions made may feel a bit out of context and unnatural.

C. How Important is Text Fluency for (Spoken) HRI?

In the text-only study, we saw how STRAP performs very poorly in terms of fluency, especially if compared to much bigger models such as GPT-3. However, we also saw in the video-based study that this difference almost vanishes when the text is actually spoken by the robot. Reasons for this could lie in a worse understanding of a spoken language where it is much easier to miss a misspelt word or misinterpret one for another, possibly preferring the one that better fits the context. This raises the question of how important fluency *is* for real-world HRI, something worthy of future study in the context of (increasingly) automatic dialogue generation for robots.

V. Conclusions

In this work we presented two automated methods to manipulate the language used in a dialogue with the goal of conveying different personalities, focusing on the extraversion trait. We then tested our methods on multiple dialogues in a text-only evaluation study, also comparing them to a hand-crafted equivalent based on relevant literature. We were able to show that all of our methods, including the hand-crafted approach, resulted in two versions of the same dialogue perceived significantly differently in the extraversion trait. We then used the output of our methods to power the speech of a Furhat robot. In doing so, we also manipulated the facial expression of the robot to reflect the emotional state that was (automatically) extracted from the text. When transitioning from text-only to multi-modal delivery of speech we observed a general decrease in extraversion (in all but the hand-crafted condition) while maintaining the same *significant* difference between the two versions of the same dialogue. We also observed that one of our methods which performed poorly in terms of fluency in the text-only study achieved noticeably higher perceived fluency when integrated into a robot.

In the end, we were able to successfully develop and evaluate two automated methods that, from a given text, are able to produce personality-adapted speech for an emotional social robot. We observe that not all of our methods were able to fully generalise, with one method particularly underperforming in one scenario evaluated in the text-only study. We believe this work can represent a good first attempt in aiding HRI research on personality adaptation for robots towards less hand-crafted and more automated methods of generation.

A. Limitations and Future Work

The first limitation is given by our focus on only one trait, extraversion. However, all our models could be extended to the whole Big Five framework: GPT-3 relied on text from the PERSONAGE dataset that also contains sentences for the other traits and STRAP was trained on a dataset where each essay's author was also rated on the other traits.

Also, when implementing the emotion recognition system and its incorporation into Furhat using its gestures, we relied on existing off-the-shelf methods which we did not internally (re-)validate. Future work could also explore the manipulation of prosody as part of personality/emotion expression.

Finally, our work assumes having already some text that the robot can turn into speech. This assumption was intentional and contributes to the ability of the method to be applicable to multiple domains. The very beginning of our pipeline, where we provided our models with a pre-scripted full dialogue could be replaced by, for example, a set of possible sentences to be used in an interaction or by another language model trained for e.g. Q&A task, etc. All these additions would seamlessly integrate with our work for which we advocate having good adaptability to any context.

Acknowledgements

This work was partially funded by grants from the Swedish Research Council (2017-05189), the Swedish Foundation for Strategic Research (SSF FFL18-0199), the S-FACTOR project from NordForsk, the Digital Futures research Center, the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and the Wallenberg Al, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

References

[1] K. Harris and S. Vazire, "On friendship development and the big five personality traits," *Social and Personality Psychology Compass*, vol. 10, no. 11, pp. 647–667, 2016.

[2] A. Neal, G. Yeo, A. Koy, and T. Xiao, "Predicting the form and direction of work role performance from the big 5 model of personality traits," *Journal of Organizational Behavior*, vol. 33, no. 2, pp. 175–192, 2012.

[3] S. Andrist, B. Mutlu, and A. Tapus, "Look like me: matching robot personality via gaze to increase motivation," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 3603–3612.

[4] A. Andriella, H. Siqueira, D. Fu, S. Magg, P. Barros, S. Wermter, C. Torras, and G. Alenya, "Do i have a personality? endowing care robots with context-dependent personality traits," *International Journal of Social Robotics*, pp. 1–22, 2020.

[5] C. Esterwood, K. Essenmacher, H. Yang, F. Zeng, and L. P. Robert, "Birds of a feather flock together: But do humans and robots? a meta-analysis of human and robot personality matching," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 343–348.

[6] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 325–332.

[7] D. Ullrich, "Robot personality insights. designing suitable robot personalities for different domains," *i-com*, vol. 16, no. 1, pp. 57–67, 2017.

[8] T.-H.-H. Dang, A. Aly, and A. Tapus, "Robot Personality Design for an Appropriate Response to the Human Partner," in *Feedback Readability for Robots Workshop (in Conjunction with the 21st IEEE International Symposium on Robot and Human Interactive Communication "Ro-Man")*, 2012.

[9] M. Y. Lim, J. D. A. Lopes, D. A. Robb, B. W. Wilson, M. Moujahid, E. De Pellegrin, and H. Hastie, "We are all individuals: The role of robot personality and human traits in trustworthy interaction," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[10] K. J. Tusing and J. P. Dillard, "The sounds of dominance. vocal precursors of perceived dominance during interpersonal influence," *Human Communication Research*, vol. 26, no. 1, pp. 148–171, 2000.

[11] J. Pittam, *Voice in social interaction*. Sage, 1994, vol. 5.

[12] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[13] A. Furnham, *Language and personality.* John Wiley & Sons, 1990.

[14] A. Thorne, "The press of personality: A study of conversations between introverts and extraverts." *Journal of Personality and Social Psychology*, vol. 53, no. 4, p. 718, 1987.

[15] F. Heylighen and J.-M. Dewaele, "Variation in the contextuality of language: An empirical measure," *Foundations of science*, vol. 7, no. 3, pp. 293–340, 2002.

[16] J.-M. Dewaele and A. Furnham, "Extraversion: The unloved variable in applied linguistic research," *Language Learning*, vol. 49, no. 3, pp. 509–544, 1999.

[17] F. Mairesse and M. A. Walker, "Personage: Personality generation for dialogue," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 496–503.

[18] F. Mairesse and M. Walker, "Controlling user perceptions of linguistic style: Trainable generation of personality traits," *Computational Linguistics*, vol. 37, no. 3, pp. 455–488, 2011.

[19] L. Moshkina, S. Park, R. C. Arkin, J. K. Lee, and H. Jung, "Tame: Time-varying affective response for humanoid robots," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 207–221, 2011.

[20] E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," *Journal of personality*, vol. 60, no. 2, pp. 225–251, 1992.

[21] L. R. Goldberg, "The structure of phenotypic personality traits." *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[22] I. B. Myers, "The myers-briggs type indicator: Manual (1962)." 1962.

[23] S. T. Völkel, R. Schödel, D. Buschek, C. Stachl, V. Winterhalter, M. Bühner, and H. Hussmann, "Developing a personality model for speech-based conversational agents using the psycholexical approach," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[24] R. J. Harvey, "Reliability and validity," *MBTI applications*, 1996.

[25] G. J. Sipps, R. A. Alexander, and L. Friedt, "Item analysis of the myers-briggs type indicator," *Educational and Psychological Measurement*, vol. 45, no. 4, pp. 789–796, 1985.

[26] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[27] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Computational Linguistics*, vol. 48, no. 1, pp. 155–205, 2022.

[28] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, "What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database," in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2018, pp. 105–113.

[29] M. Paetzel-Prüsmann, G. Perugia, and G. Castellano, "The influence of robot personality on the development of uncanny feelings," *Computers in Human Behavior*, vol. 120, p. 106756, 2021.

[30] M. L. Walters, D. S. Syrdal, K. Dautenhahn, R. Te Boekhorst, and K. L. Koay, "Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion," *Autonomous Robots*, vol. 24, no. 2, pp. 159–178, 2008.

[31] M. Mäkäräinen, J. Kätsyri, and T. Takala, "Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley?" *Cognitive Computation*, vol. 6, no. 4, pp. 708–721, 2014.

[32] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating unsupervised style transfer as paraphrase generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 737–762.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[34] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[35] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[37] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[38] J. Hartmann, "Emotion english roberta-large," https://huggingface.co/j-hartmann/emotion-english-roberta-large/, 2022.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[40] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[41] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner, "Effective persuasion strategies for socially assistive robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 277–285.

[42] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in personality*, vol. 40, no. 1, pp. 84–96, 2006.

[43] O. P. John, S. Srivastava *et al.*, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[44] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," *arXiv preprint arXiv:2006.14799*, 2020.

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[46] JASP Team, "JASP (Version 0.16.2)[Computer software]," 2022. [Online]. Available: https://jasp-stats.org/

[47] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[48] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas) development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, 2017, pp. 254–262.

[49] A. Tapus, C. Ţăpuş, and M. J. Matarić, "User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intelligent Service Robotics*, vol. 1, no. 2, pp. 169–183, 2008.

[50] M. Schmitz, A. Krüger, and S. Schmidt, "Modelling personality in voices of talking products through prosodic parameters," in *Proceedings of the 12th international conference on Intelligent user interfaces*, 2007, pp. 313–316.