Proceedings of the 27th IEEE International Symposium
on Robot and Human Interactive Communication,
Nanjing, China, August 27-31, 2018

Expressing Coherent Personality with Incremental Acquisition of Multimodal Behaviors

Pedro Mota, Maike Paetzel, Andrea Fox, Aida Amini, Siddarth Srinivasan,
James Kennedy and Jill Fain Lehman*

***Abstract*— As social robots increasingly enter people's lives, coherence of personality is an important challenge for long-term human-robot interactions. We extend an architecture that acquires dialog through crowdsourcing to author both verbal and non-verbal indicators of personality. We demonstrate the efficacy of the approach through a four-day study in which teams of participants interacted with a social robot expressing one of two personalities as the host of a competitive game. Results indicate that the system is able to elicit personality-driven language behaviors from the crowd in an incremental and ongoing way and produce a coherent expression of that personality during face-to-face interactions over time.**

I. Introduction

The use of crowdsourcing platforms to author agent behavior is a relatively new phenomenon, and attractive to system builders as a solution to the problem of scale [1], [2], [3]. A potential difficulty for the approach, however, is the variability introduced by multiple authors, particularly when those authors have little or no access to each other's contributions. For agents that interact via language, large variability in 'tone' – e.g., unmotivated swings from pleasant to unpleasant – can create a jarring incoherence in personality.

In long-term interaction, crowd authoring also offers the possibility of growing the agent's repertoire of behaviors incrementally and flexibly as a function of its experience. Unfortunately, repetition exacerbates the problems associated with distributed authorship. The more we have experienced a coherent personality in the past, the more we come to expect it in future interactions.

We are interested in autonomous, language-based characters that can interact with the same individuals repeatedly over time. The architecture we have developed for building a Persistent Interactive Personality (PIP) contains multiple learning mechanisms, including crowdsourcing, that incrementally acquire a character's verbal and non-verbal language behavior. Previous work has explored how such learning mechanisms support interactivity and persistence [4], [5]. We extend this work by focusing on personality, and whether the PIP architecture supports its expression.

This contribution details the techniques developed for incorporating personality into the architecture, as well as their deployment on a social robot, named 'William'. William is tasked with hosting a trivia competition between two teams, where hosting involves both managing the game play and engaging in social chit chat. Each team experiences a different personality as authored by the crowd. We intend that **IMP** (impatient) William be perceived as 'quick to overreact with little patience for life's imperfections' and **OPT** (optimistic) William be perceived as 'lighthearted, optimistic, and determined to find the fun in every situation'. We demonstrate that although crowd workers did not coordinate with each other or directly experience the interaction scenario, the character is able to combine and generalize their work into two coherent and distinct personalities.

*During the time of the project all authors were affiliated with Disney Research, Pittsburgh, PA 15213, USA

II. Related Work

The effect of agent personality on user experience has been studied in a number of domains, including personal assistance [6], television program recommendation [7], and rehabilitation therapy [8]. In some tasks, the particular personality may matter; whether an agent exhibits more introverted or extroverted traits influences users' perceived level of control [7], as well as how closely they attend to the interaction and take suggestions from the agent [7], [9]. In general, however, personality engages the user, and the more naturally an agent can express its personality, the better it will maintain user attention [9].

Though some research focuses on communicating personality through facial expression alone [10], [11], others combine facial expression, gesture, and language use in human-robot interactions [7], [8], [9], [12], [13]. The work in [12] assesses the emotive capacity of synthesized speech in coordination with facial expression via hand-adjustment of synthesizer settings. In contrast, [8] features a mobile, machine-like agent as 'therapist' that adapts its own behavior (e.g., speech rate, phrasing of language cues, activity level, and proximity to others) to that of the user as 'patient' along the lines of introversion/extroversion. [13] also explores introversion versus extroversion, studying users' interactions with small humanoid agents that vary the rate and size of their arm movements as well as the rate, volume, pitch, and verbosity of their speech while acting as cleaner or museum guide.

These researchers, along with [7], evaluate user reactions to agent personalities but do not explore how coherent personality can be developed. While the agent in [8] does learn in the sense that it adapts to the user within a single interaction, it does not carry that learning over multiple interactions. Furthermore, the previous agents typically have a language capacity limited to basic phrase variations. In contrast, [9] incorporates PERSONAGE [14], a natural language generator, that combines with posture, gesture, head movement, and

978-1-5386-7980-7/18/$31.00 ©2018 IEEE

facial expression to produce a variety of expressive phrasings in a full-body robot offering restaurant recommendations.

Despite their success in expressing personality, [9] and the other systems described above are largely hand-crafted approaches for agents to exhibit a single static personality. Because of this dependence on predefined traits, these techniques do not scale well and limit the range of expression an agent has in an interaction. [15] attempts to tackle this limitation by providing an 'infinite personality space' for a mobile, rover-like agent that explores an area by making behavioral decisions. By using a combination of factors in a continuous space to determine the robot's motor movement, unique and distinct personalities are created for each instantiation of the agent. Because the problem of conveying personality in a language-based interaction cannot be expressed on a continuous scale, this approach does not transfer to characters engaging in language-based interactions.

Developing an agent's personality incrementally over time is a relatively unexplored concept. There have been efforts to achieve continuity, but an agent that can 'remember' cached experiences when faced with similar interactions [16] or adapt game play based on a prior interaction with a human player [17] is not learning personality but recalling situational strategies. The approach to automatic content-authoring described here is not only scalable but also, like [15], allows the agent to develop its personality over time. In contrast to [15], the current method expands the definition of personality to multimodal, language-based interactions.

III. Research Questions

The purpose of the current work is to explore whether a coherent personality can emerge from the incremental acquisition of language behavior from crowd workers. Three related questions shape the inquiry:

RQ1 Given only simple personality descriptions as part of a semi-situated narrative, can crowd workers generate dialog lines that are perceived as *in character* by others?

This question relates to the usefulness of crowdsourcing as a technique for dialog generation with personality-driven features. To acquire dialog behavior, PIP characters provide crowd workers with a narrative that describes the situation for which a dialog line is needed, either to be used by the robot's character or to compare against a human reply. In previous characters, the task required only that crowd workers write a line that they would, themselves, produce in those circumstances. When authoring for William's turns, however, the task requires writing for someone else, and it is not clear if the crowd working platform can do that task well. We evaluate the efficacy of crowd authoring by asking whether lines written for a personality are judged to be acceptable less often than lines written without reference to personality.

RQ2 Can an agent using incrementally-generated dialog and accompanying facial expressions produce a *coherent* personality experience over time?

This question relates to how the individually authored pieces are perceived by Williams' conversational partners in the context of interaction. Even if each line and expression produced by the crowd has been judged to be in character, there is no guarantee that the real-time combination of the pieces produces a coherent multimodal signal over time. We evaluate coherence of the robot's behavior both by measuring how much personality was exposed to each player over time and by asking participants directly via a survey.

RQ3 Can an agent using incrementally-produced dialog and accompanying facial expressions produce a *distinct* personality experience over time?

This question also relates to the participant's perception, but is separated from RQ2 because even if each line and expression has been judged to be adequately in character and their combined use has been coherent, there is no guarantee that the resulting experience creates the intended impression. In other words, each of the two versions of William could be coherent without the two versions being distinct. We evaluate distinctness with participants' ratings of the robot along a number of personality dimensions.

IV. Designing for Personality

The research questions motivate an interaction design in which there is opportunity to learn, portray, and contrast personalities. We created a competitive game where both the game play structure and social chat afford those possibilities. The competition is a trivia contest in which the user thinks of a character from an animated movie that William must guess by asking yes/no questions. Points are awarded based on how often the robot has guessed the character for that team before. Chit chat occurs before and after playing the trivia game and is centered around how the individual and team are doing in the competition. Personality-driven behavior is acquired for both game and chat stages, with social chat being the natural locus for ongoing dialog-learning after deployment.

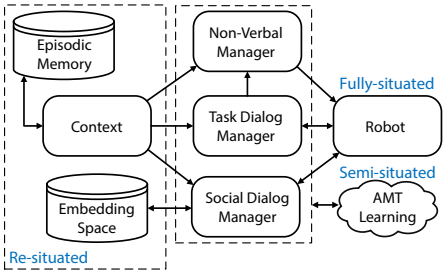
A. Platform

William is embodied in a back-projected Furhat robot head [18] with an adult male face (see Fig. 1). The head sits at about eye level on a stand that hides the controlling computer. Visual input from a Microsoft Kinect V2 on a tripod behind the robot tracks the body of the player and controls William's eyes and 2-DOF head movements. The stand also supports a touch screen, which players use to identify themselves, and a Logitech C920 camera whose output is sent to cloud-based Microsoft Bing ASR. William computes its response as described in Sec. IV-B. The reply is then synthesized using the male CereProc voice *William* and delivered via Furhat's speakers. Corresponding lip movements are controlled automatically by the Furhat platform, while facial expressions are computed as described in Sec. IV-D.

B. Agent Architecture

Previous characters built in the PIP architecture have explored persistent interaction either solely through task [4] or chit chat [5] applications. William represents an extension of the architecture to accommodate both kinds of dialog (Fig. 2a). An interaction with William is composed of a sequence of

Fig. 1: A user and William during interaction.



(a) System diagram.

(b) Interaction stages.

Fig. 2: The generic PIP architecture (Fig. 2a) and William's pre-determined sequence of interaction stages (Fig. 2b).

stages encompassing both types of dialog, interleaved in a fixed manner (Fig. 2b). As most of PIP's mechanisms have been described elsewhere, we review them only briefly here and then focus on their particular use in learning and expressing personality in William's task.

The main data structure in a PIP character is a dialog graph for interactions learned incrementally over time. In William's case, there is a dialog graph for each interaction stage. Each node in the graph corresponds to a collection of semantically similar utterances, where similarity is defined by the angle between the vector representations of the utterances in an Embedding Space (Fig. 2a: bottom-left). The Embedding Space was produced using doc2vec [19], trained with soap opera scripts (to reflect the social chat portion of our task) and documents related to the animated characters in the game (to cover situations where users might discuss the characters).

Social conversations emerge from graph traversal by first matching the user's utterance with a node that is a continuation of the current path, then selecting an utterance from one of its successor nodes. An utterance matches if its similarity value falls above an empirically-defined threshold with respect to the embedding space. Each node also has an associated context, and possible responses are ranked based on context overlap. For William, the context were pre-defined to be *familiarity*, *agent personality*, *agent gender*, *user gender*, *user performance*, *team performance*, and *last game result*. Task dialog emerges in the same way as way as social dialog, except that the task graph is static and needs to be specified *a priori* in order to trigger task related actions according to the traversed nodes.

Like other PIP characters, William's behavior changes over time due to three types of learning: fully-situated (language from users), semi-situated (language from crowd workers), and re-situated (language from generalization). Fully-situated learning is driven by conversational failures in face-to-face interactions and is predicated on access to all available dialog history and corresponding context. Semi-situated learning occurs through crowdsourcing in which crowd workers only have access to a narrative description of the context and a partial history. Re-situated learning is the generalization of language, acquired by the other mechanisms, to contexts other than that in which it was first learned. In this study, trivia game language was hand-authored and personality-independent for consistency across conditions. Social language was initialized pre-deployment using semi-situated learning and acquired dynamically during the study via all three learning mechanisms. The dynamic learning procedure took place at the end of each day of the study. An expert designed an initial set of possible non-verbal behaviors during pre-deployment based on the target personalities and possible contexts. For each context the appropriateness of all expressions was learned through crowd-sourcing and general rules for dynamically enriching the language content with non-verbal behavior were derived from this initial evaluation.

C. Personality through Language

Much of William's language behavior comes from semi-situated learning either pre-deployment or after the contest has begun. This method of dialog acquisition occurs through an autonomous pipeline of Human Intelligence Tasks (HITs) on Amazon Mechanical Turk (AMT). There are two stages in the pipeline: authoring and editing. In the authoring stage, workers are asked to write a single line of dialog based on a story narrative and up to five lines of previous conversation. The narrative provides constraining context to the author and the initial variable values that should hold when the line is considered for use in the future. William's personality is one such variable, and, thus, workers need to consider it when authoring lines. In the editing stage, workers are asked to judge a set of authored lines given exactly the same narrative. Three different workers provide responses per HIT. Judgments are made on a 0-5 scale and dialog lines are added to the graph only if they receive a sufficiently high average rating. Thus, a single HIT can expand the graph with multiple lines.

To ensure chat will be experienced even in the first interaction players have, William uses the AMT pipeline to generate initial graphs for all social interaction stages prior to deployment. HITs are generated for each possible combination of context variable values relevant to an interaction stage. An accepted dialog line is randomly selected for further expansion and included in the dialog history to acquire the next response set. This loop continues until a depth of four is reached, with alternating turns representing William or the player. During this graph-initialization process, lines that anticipate player dialog simply require crowd workers to

respond based on what their own behavior would be in the situation described. But when the output of the HIT is to be used by William, the narrative includes a description of either the **OPT** or **IMP** personality, and workers are told to author/judge dialog lines relative to that characterization. From a system design point of view, the personality variable is no different from other context variables. From the worker point of view, however, it adds complexity to the HIT because workers need to produce or judge lines based on personality traits that may differ from their own. It is this distinction between who is expected to say the dialog line (user or William) that allows us to evaluate **RQ1**.

The AMT pipeline is also used as part of fully-situated and re-situated language learning when there is no response for a player utterance, the player's utterance is not similar to anything seen before, or an utterance has been used after generalizing its context. In the case of no response, the full pipeline is used. In the remaining cases, editing HITs are used, allowing the character to filter out errors made by the Automatic Speech Recognition (ASR) and bad language generalizations over contexts (for example, an answer given to a female user may contain pronouns that are inappropriate for a male user). If an acceptable ASR utterance is found, a full HIT for the next turn is generated, allowing William to have an answer if this situation occurs again. As above, lines on William's turns include narrative text related to personality, while lines that reflect player turns do not. If these HITs are approved, the corresponding lines are added as nodes to the dialog graph. This mechanism allows William to incrementally learn by expanding its initial graph in a personality-driven manner.

D. Multimodal Personality

The robot platform allows two non-verbal modalities to be controlled: voice tone and facial expression. Both were used to create a lifelike impression of the character and to enhance the personality expressed by the robot.

Voice Tone

CereProc voices allow several modifications to the synthesis of an utterance, including labeling emotion, changing speech rate, and adding pitch contours. The *cross* emotion tag was used for **IMP** William to impart impatience. Following [20], **OPT** William was given a faster rhythm and accent on phrase-final words to create a happy voice. These hand-crafted settings were validated pre-deployment in an on-line study in which participants rated the virtual robot performances of dialog lines by neutral, cross, and happy voices given either an **IMP** or **OPT** personality description for the character. During game play, William switches between neutral and personality-based voice to avoid overacting. The choice of which voice to use is tied to the selection of the facial expression, as explained in the following subsection.

Facial Expression

The robot's facial expressions were created by an expert using the Furhat SDK [18] and transferred to the physical robot as accurately as possible. Expressions were used in two different scenarios: in-game responses to participants' yes/no answers to trivia questions, and social chat. For the in-game responses, different expressions were evaluated on AMT with more than 120 crowd workers. Then, we selected the expressions that were rated significantly higher in each personality. As a result, the **IMP** character gives a brief neutral acknowledgement to positive responses but reacts impatiently if participants respond to a question with "no". In contrast, the **OPT** personality reacts joyfully if the user responds with "yes" and surprised otherwise. The expressions may be accompanied by a short nonlinguistic utterance.

For social chat, a hand-crafted set of possible expressions is tied to the interaction stages based on the language learned pre-deployment. The utterances from the first expansion of the initial dialog graph were analyzed in order to manually create six different expressions per personality. These expressions were evaluated on AMT with 40 crowd workers, and the 4 expressions best rated for each personality were selected. All expressions performed during social chat were accompanied by the happy or cross voice for the respective personality. During the game, the voice was always neutral except for **OPT** William's responses to positive replies, which used the happy voice. The cross voice was not applied to the game questions, because questions follow a different voice contour in the synthesis, which is difficult to understand in combination with the high speech rate of the cross voice.

In the previous step we obtained the expressions to be used in social chat. But it should not be assumed that we can use them in every situation without breaking coherence in personality, as context needs to be taken into account. For example, showing a happy smile might not be appropriate if the conversational partner just lost a game. Given the number of possible context instantiations, having to author such personality coherence rules is a considerable effort. We alleviate the developers' effort by having the robot learn the appropriateness of each expression in a semi-situated context from AMT by choosing sentences from the initial graph expansion. Two sentences were randomly chosen for each possible context variable assignment. Then, we automatically generated one video per dialog turn and possible expression, and acquired judgments about how likely it would be for the character to convey the corresponding utterance shown in the video. The previous dialog turns were also provided to give context. Five crowd workers per task were asked to judge the expression and voice tone given the description of the character's personality. The ratings used a six point Likert scale ranging from 'very unlikely' to 'very likely'. As a result, William learned the appropriateness of each expression in every context. In addition, it learned that the initial connotation of the context (e.g., joyful for winning the game) becomes even stronger for later turns in that same context. Initially, we hypothesized that expressions become less appropriate as conversations unfold, since we have less knowledge about the direction in which the conversation evolved, and, consequently, we also would not know what is the current relevant context. However, the random examples

show that topic switches rarely occur, and that the original emotion often becomes more intense as the dialog continues. Therefore, we kept separate ratings for the first and later turns to make a situated decision for showing an expression.

During all interactions, the decision to perform an expression was taken in two steps. First, the robot decided which expression would be used by drawing from a categorical distribution where the size of the probability vector is equal to the number of available expressions for the current context. The weight w for each expression e in the probability vector was based on the ranking of the expressions according to the appropriateness rating and corresponding standard deviation in the current context. Expressions with higher rating and lower standard deviation were ranked higher. The cumulative distribution function factors in the conversation history to prevent repetitiveness when generating expressions.

In the second step, the robot decides if the chosen expression should be performed or not. The goal was to perform an expression for about every other line, which avoids being too static or overacting the personality. Different decision mechanisms are used depending on whether William is in a social chat or game stage. During the game, the decision was based on the conversation history by taking a weighted random choice using a normal cumulative distribution function. In social chat, if the specific expression fits the context particularly well, or if the expression is rated much higher than the unimodal (neutral) delivery of the line, William might decide to perform an expression for two utterances in a row. If, however, the expression fits the context poorly, or if the neutral delivery is more appropriate, William might not perform the expression. In addition to the context-dependent decisions, the robot adds subtle in-personality expressions like eyebrow raises or frowns to give a more lifelike impression.

V. Evaluation

To evaluate the coherence (RQ2) and the distinctness (RQ3) of the two personalities generated using the architecture proposed in Sec. IV, we designed a between-subject experiment with *personality* as the independent variable. The dialog graph produced by all subjects also provides the data for evaluating the efficacy of the semi-situated pipeline (RQ1).

A. Participants

In the evaluation, 25 employees were recruited for the experiment. Two subjects withdrew, leading to a slight subject imbalance; 23 participants (13 female; age *M*=29.87) were assigned to the two conditions with 13 subjects in the **OPT** condition (7 female; age *M*=31.15) and 10 in the **IMP** condition (6 female; age *M*=28.20). 69.57% of participants had interacted with the physical robot platform (not the William character) at least once before this study (76.92% in **OPT** condition; 60% in the **IMP** condition). All participants are currently living and working in an English-speaking environment, 60.87% of them are native English speakers (**OPT**: 61.54%; **IMP**: 60%), and 52.17% have a native-born American accent (**OPT**: 61.54%; **IMP** 40%), which has the best recognition rate for the ASR settings used. The study was IRB-approved, and participants received monetary compensation for taking part in the study.

B. Procedure and Measures

To engage participants in repeated interactions with William, the experiment was designed as a trivia competition about animated movie characters. Participants were divided into two teams by balancing gender, native language, and prior experience with the robot platform; participants on the same team were assigned the same personality condition. Each interaction corresponded to playing a round of the trivia game. Team members were assigned to play with the same personality, although they were unaware of the existence of different personalities or of any other purpose to interacting with the robot other than playing the game. The competitive setup meant team members had a disincentive for exchanging information about their interactions across conditions.

The competition was held in an office environment for four consecutive days. The robot was placed in a private space without direct human supervision. Prior to their first interaction, participants provided informed consent and received written game instructions. In the beginning of each session, players identified themselves via touch screen. After greeting and revealing the current score, William initiated a short social chat before moving on to the trivia game. The game started with William asking players to think about an animated movie character. The rest of the game play consisted of a variable length sequence of yes/no questions about that character, during which William had two attempts to guess it. The game was followed by another brief social chat before the robot said goodbye. Participants could then either play another round immediately or come back at a later time.

Conversations were logged for later analysis, including the computation of how much and in what modality personality was expressed, as necessary for RQ1 (comparison of personality-driven versus non-personality-driven learning) and RQ2 (coherence of the multimodal personality).

After their last game session, participants were given a questionnaire that covered both demographics and their perception of William. Participants' general subjective experience of the character over time was measured by five questions based on the Intrinsic Motivation Inventory (IMI) [21]. With respect to RQ2, seven questions adapted from [22] and [23] were used to evaluate the quality and coherence of the multimodal expressions. Regarding RQ3 (distinctness), participants were asked to rate the robot on the ten-item personality inventory from [24]. We hypothesized that a significant difference would be observed in at least one personality dimension if we successfully created two distinct robot personalities. Finally, different personalities potentially influence the overall perception of the robot as well. This was assessed by three scales from the Godspeed questionnaire [25] which were selected based on the factor analysis by [26].

VI. Results

On average, participants had a total of 10.87 sessions with the robot (10.85 in the **OPT** condition; 10.90 in **IMP**) with

an average length of 2.86 minutes (2.92 in **OPT**; 2.78 in **IMP**) per session. Overall, 124 different characters were guessed correctly, leading to 328 points (**OPT**: 162, **IMP**: 166). Neither the difference in outcome for the competition nor in success rate when guessing characters influenced people's self-assessed success in the game (**OPT**: *Mdn*=3.0; **IMP**: *Mdn*=3.5), $Z = 0.459$, $p = .647$.

Below we report statistics from users' conversations and their responses to the questionnaire. Shapiro-Wilk normality tests revealed that the survey responses and most conversation-related measures were not normally distributed, so we use the two-sided Wilcoxon Rank Sum statistic to test for the significant influence of personality as a grouping factor.

A. RQ1: Can crowd workers author in character?

To answer this research question, we computed the percentage of dialog lines that were added to the graph through semi-situated learning for personality-driven versus non-personality-driven narratives. The comparison reveals whether it is harder to generate dialog lines when specific personality traits are required. Across all semi-situated dialog lines acquired, the observed approval rates were 98.2% (**OPT**), 96.3% (**IMP**), and 98.2% (no-personality). Given these results, we argue that the burden of accommodating personality does not make PIP's reliance on crowdsourcing as a learning mechanism impractical. The difference between **OPT** and **IMP** suggests some loss of productivity under some personality descriptions, a point we return to in the discussion.

The average rating for the generated lines is another way in which the success of the authoring can be judged. Across all lines, the average ratings were: *M*=4.33, *SD*=1.07 (**OPT**; *n*=995), *M*=4.24, *SD*=1.17 (**IMP**; *n*=941), and *M*=4.29, *SD*=1.01 (no-personality; *n*=524). It can be seen that these lines have similar overall ratings, and similar variance, with all averages above a 4 on the 5-point scale. Thus, from these numbers it appears that completing the task with personality poses no greater challenge to the crowd workers than authoring without personality descriptors.

B. RQ2: Are verbal and non-verbal expressions coherent?

As described in Sec. IV, crowdsourcing is employed separately for creating the verbal and non-verbal behaviors of the robot, but a combination of both conveys the personality during the interaction. Survey questions were used to evaluate if the result was perceived as coherent. Overall, on a 5-point scale, the quality of the multimodal behavior was rated to be high: participants felt that William's speech and expressions matched the personality well, that the expressions matched the verbal content of the utterances, that speech and expressions were well synchronized, and that the behavior was expressive (**OPT**: *Mdn*=4.0; **IMP**: *Mdn*=4.0, for all of the aforementioned). According to participants' self-assessment, they barely felt distracted by the multimodal nature of the behavior (**OPT**: *Mdn*=1.0; **IMP**: *Mdn*=2.00).

Despite the generally promising rating of the multimodality, people rated the combination of speech and expressions to be only somewhat natural (**OPT**: *Mdn*=3.0; **IMP**: *Mdn*=3.50), and the conversations to be only moderately engaging (**OPT**: *Mdn*=4.0; **IMP**: *Mdn*=3.0). Nevertheless, personality had no significant influence on any of the survey items related to the multimodality of the interaction ($p > .2$, for each).

A coherent personality does not only depend on a good interaction between modalities; the robot should also be coherent in the level of personality it expresses during conversations. Players should experience the personality similarly every time they interact with William. Ideally, this expression level is similar across the personalities, to ensure we did not build one 'strong' and one 'weak' personality for the robot. The level of personality exposure is determined from the average rating each line or expression received in the AMT pipeline and the number of times it was used. On average, William engaged in social chat for 10.73 turns per interaction. People had slightly fewer social chat turns with the **IMP** personality (*M*=9.85, *SD*=1.94) than with **OPT** (*M*=11.41, *SD*=1.51), $Z = -1.954$, $p = .051$, possibly because the personality conveyed impatience. The robot verbally exposed a stronger **OPT** personality (*M*=4.2, *SD*=0.06) compared to the **IMP** robot (*M*=4.05, *SD*=0.1), $Z = -3.039$, $p = .002$. A similar effect is seen for expressions: although the ratio between robot turns that are delivered neutrally vs. multimodally is similar in both conditions (**OPT**: 34.60% multimodal, **IMP**: 33.27%), $Z = -1.426$, $p = .154$, the average strength of expressions people were exposed was significantly higher for the **OPT** personality (*M*=4.95, *SD*=0.07) than for **IMP** (*M*=4.28, *SD*=0.05), $Z = -4.031$, $p < .001$.

Because the robot learns, we also examine how exposure to the personality changed over time. The length of conversation grew from an average of 6.18 turns on the first day (**OPT**: *M*=6.23; **IMP**: *M*=6.12) to 14.7 on the last day (**OPT**: *M*=15.69; **IMP**: *M*=13.2). Despite the lengthening of the conversations, each person experienced the same amount of personality in the utterance content over the four days of interaction. On average, the strength of personality in verbal output differs *M*=0.046 for people playing with the **OPT** personality and *M*=0.079 when playing with the **IMP** one between repeated interactions with William. The personality does not have a significant influence on that variance, $Z = 1.799$, $p = .072$. In addition, the ratio of unimodal to multimodal output does not vary over time, suggesting that people played with a robot of the same level of expressiveness for every game in both conditions (**OPT**: *M*=0.007; *SD*=0.008; **IMP**: *M*=0.013, *SD*=0.023; $Z = 0.434$, $p = .664$). However, the strength of the multimodal expressions varies more in the **IMP** personality (*M*=0.045, *SD*=0.036) than in the **OPT** one (*M*=0.023, *SD*=0.012), suggesting that people in the **IMP** condition experienced a wider spectrum from a 'mildly impatient' to a 'very impatient' multimodal reaction, while users in the **OPT** condition experienced less variance between repeated interactions. Even though this effect is significant, $Z = 2.109$, $p = .035$, the variance is so small that whether this effect is actually perceivable by the user is questionable.

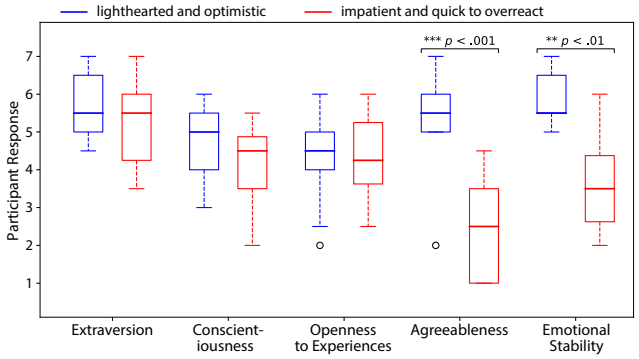


Fig. 3: Box plots for the Big 5 dimensions for William's personalities.The **OPT** personality is deemed significantly more agreeable and emotionally stable than **IMP**.

C. RQ3: Are William's personalities distinct?

Participants generally found William to be quite extroverted (**OPT**: *Mdn*=5.5; **IMP**: *Mdn*=5.5), reasonably conscientious (**OPT**: *Mdn*=5.0; **IMP**: *Mdn*=4.5), and open to experiences (**OPT**: *Mdn*=4.5; **IMP**: *Mdn*=4.25). Personality had no significant influence on these dimensions ($p > .2$, for each). However, participants rated the **IMP** robot to be significantly less agreeable (*Mdn*=2.5) than the friendly version (*Mdn*=5.5), $Z = -3.716$, $p < .001$, and less emotionally stable (**OPT**: *Mdn*=5.5; **IMP**: *Mdn*=3.5), $Z = -3.067$, $p = .002$ (Fig. 3).

William's personality did not significantly influence the perception of his competence (**OPT**: *Mdn*=3; **IMP**: *Mdn*=4), $Z = 1.105$, $p = .269$ or human-likeness (**OPT**: *Mdn*=3; **IMP**: *Mdn*=2.5), $Z = 0.706$, $p = .48$. However, participants in the **IMP** version rated the robot significantly less pleasant (*Mdn*=2) compared to participants playing with **OPT** (*Mdn*=4), $Z = -3.621$, $p < .001$. Interestingly, the pleasantness of the robot did not significantly influence how much people enjoyed playing (**OPT**: *Mdn*=4; **IMP**: *Mdn*=4.5), $Z = 0.426$, $p = .67$, or chatting with the character (**OPT**: *Mdn*=3; **IMP**: *Mdn*=4.5), $Z = 0.902$, $p = .367$.

In the final questionnaire, participants were asked to describe William in three words. The descriptions assigned to the **OPT** and the **IMP** personality differ significantly in sentiment and valence: a sentiment analysis based on SentiWordNet [27] showed the words describing the **OPT** robot were more positive (*M*=0.43, *SD*=0.23) compared to the descriptions for the **IMP** personality (*M*=0.29, *SD*=0.26), $Z = -2.168$, $p = .030$, while the descriptions for **IMP** were more negative (**IMP**: *M*=0.37, *SD*=0.31; **OPT**: *M*=0.13, *SD*=0.17), $Z = 2.978$, $p = .003$. Similarly, based on [28], the valence of the words used to describe William with the **OPT** personality (*M*=6.78, *SD*=1.57) is significantly higher than for the **IMP** personality (*M*=4.96, *SD*=2.14), $Z = -3.025$, $p = .002$, suggesting that people indeed rated the optimistic personality to be more pleasant than the impatient one.

VII. Discussion and Future Work

The purpose of this work is to understand if extensions to the PIP architecture can support the expression of personality. The results show that the crowd workers, the backbone of the semi-situated learning mechanism, can author and edit dialog behaviors about as successfully when they are required to do so for a specific personality as when they are not (RQ1). Results also show that the personality expressed by the dynamic recombination of the pipeline's output is perceived as coherent (RQ2) and distinct (RQ3) over time. While coherent and distinct, the personalities show some differences. In particular, **OPT** was slightly easier to author and was delivered more strongly and more often, but with less variability in degree of expression. These differences suggest some caution in generalizing the results to a broader personality range.

Personality does not only influence how the content of an utterance is phrased or what expressions are more likely to accompany an utterance. It also affects the general nature of the conversation. The optimistic William, for example, often asks users how they are doing or if they have weekend plans, while the quick-to-overreact personality opens the conversation with impatient comments. Given such differences, even though the robot is programmed to attempt to continue the current chat, it is possible that the conversational partner was less likely to respond with a line that allowed continuation in the latter case. In essence, we observe that the type of personality may have a strong impact on the nature, content, and length of the conversation. It is in the nature of a lighthearted personality to engage in chit chat, while it makes sense for an impatient personality to keep conversations brief.

The type of personality additionally influences how often, and how strong personality markers are exposed in a conversation. It seems more reasonable, for example, for an annoyed voice tone to be used after turns that triggered such a response than for the start of a conversation. In contrast, an optimistic person might be likely to start the conversation in a very cheerful way and thus with a very strong personality marker. We observe this effect when comparing the two types of personality chosen in our experiment – the **OPT** character would produce voice content and multimodal behavior that was more highly rated in personality, while the **IMP** personality had a significantly higher variance in the multimodal content between the interactions. It is difficult to predict and an interesting focus for future work to analyze how these differences transfer to a broader range of personalities.

Further, we note that while the crowd workers could successfully author for these personalities, the adjectives used to describe them were both strong and strongly opposed. It is unknown whether crowd workers would have the same degree of success were the adjectives more subtle, and equally uncertain whether the distinctness would be compromised if the contrast was less pronounced.

In the conversational observations made here, we evaluated the personality exposed in the language and in the multimodal presentation of this language separately, as they were learned via different uses of the crowd-sourcing pipeline. Potentially, however, the combination of language-based and multimodal personality has a different impact on the strength of personality that is experienced, repeatedly, in the situation. In addition,

even if an utterance was not blended with an expression, the strength of personality depends more on the situation than we expose in the narrative for the HITs. For example, using a 'neutral' voice tone might diminish a cheerful line because it sounds less congruent. In the current computation of personality strength, this cannot be accounted for.

Finally, how a user perceives an utterance in a given situation depends on the entire history of the conversation. If the user only scored two points because the animated character had been chosen by a teammate before, the line "You are a superstar!" might be perceived as more ironic than enthusiastic. However, only a very limited history of the conversation was exposed to the crowd workers so it would be easier to generalize language across contexts. The cost of that decision is that it is more difficult to estimate how strong the content of an utterance in the current context actually is.

In the future, we would like to develop a deeper understanding of the similarities and differences between the semi-situated context the crowd workers are exposed to and the situated context of the on-site interaction with users when it comes to the expression of personality. This will allow us to further evaluate which, if any, of the proposed explanations holds, and will provide additional insights into how our findings could transfer to a broader set of personalities.

VIII. Conclusions

In this paper we presented a system capable of learning personality-driven dialog for use in multimodal robot behavior. The study demonstrated that crowd workers are equally capable of successfully completing this task, whether they must write for a specific personality or are not given personality requirements. Users who took part in the study reported that the **OPT** and **IMP** personalities were significantly different in the agreeableness and emotional stability dimensions of the Big 5 questionnaire. The verbal and non-verbal aspects were largely reported to be coherent in delivering the personality. This validates the approach adopted here, which can incrementally learn personality-driven language at scale.

References

[1] M. Kriegel, "Towards a crowdsourced solution for the authoring bottleneck in interactive narratives," Ph.D. dissertation, Heriot-Watt University, 2015.

[2] W. Lasecki, R. Wesley, J. Nichols, A. Kulkarni, J. Allen, and J. Bigham, "Chorus: A crowd-powered conversational assistant," in *Proc. of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2013, pp. 151–162.

[3] W. S. Lasecki, E. Kamar, and D. Bohus, "Conversations in the crowd: Collecting data for task-oriented dialog learning," in *AAAI Conf. on Human Computation and Crowdsourcing*, 2013, pp. 2–5.

[4] I. Leite, A. Pereira, A. Funkhouser, B. Li, and J. F. Lehman, "Semi-situated Learning of Verbal and Nonverbal Content for Repeated Human-robot Interaction," in *Proc. of the Intern. Conf. on Multimodal Interaction (ICMI)*. ACM, 2016, pp. 13–20.

[5] J. Kennedy, I. Leite, A. Pereira, M. Sun, B. Li, R. Jain, R. Cheng, E. Pincus, E. Carter, and J. F. Lehman, "Learning and Reusing Dialog for Repeated Interactions with a Situated Social Agent," in *Proc. of the Intern. Conf. on Intelligent Virtual Agents (IVA)*. Springer, 2017, pp. 192–204.

[6] G. Ball and J. Breese, "Emotion and Personality in a Conversational Character," *Embodied Conversational Agents*, pp. 189–219, 2000.

[7] B. Meerbeek, J. Hoonhout, P. Bingley, and J. M. B. Terken, "The influence of robot personality on perceived and preferred level of user control," *Interaction Studies*, vol. 9, no. 2, pp. 204–229, 2008.

[8] A. Tapus, C. Ţăpuş, and M. J. Matarić, "User-Robot Personality Matching and Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy," *Intelligent Service Robotics*, vol. 1, no. 2, pp. 169–183, 2008.

[9] 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 the 8th ACM/IEEE Intern. Conf. on Human-Robot Interaction (HRI)*. IEEE Press, 2013, pp. 325–332.

[10] E. Bevacqua, M. Mancini, and C. Pelachaud, "Speaking with emotions," in *Proc. of the AISB Symposium on Motion, Emotion and Cognition*, 2004, pp. 197–214.

[11] H. Miwa, T. Umetsu, A. Takanishi, and H. Takanobu, "Robot Personality based on the Equations of Emotion defined in the 3D Mental Space," in *Proc. of the IEEE Intern. Conf. on Robotics & Automation (ICRA)*, vol. 3. IEEE, 2001, pp. 2602–2607.

[12] C. Breazeal, "Emotive qualities in lip-synchronized robot speech," *Advanced Robotics*, vol. 17, no. 2, pp. 97–113, 2003.

[13] M. Joosse, M. Lohse, J. G. Pérez, and V. Evers, "What you do is who you are: The role of task context in perceived social robot personality," in *Proc. of the IEEE Intern. Conf. on Robotics and Automation (ICRA)*. IEEE, 2013, pp. 2134–2139.

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

[15] D. H. Grollman, "Infinite personality space for non-fungible robots," in *Intern. Conf. on Social Robotics (ICSR)*. Springer, 2016, pp. 94–103.

[16] E. Oliveira and L. Sarmento, "Emotional valence-based mechanisms and agent personality," *Advances in Artificial Intelligence*, pp. 771–780, 2002.

[17] C. T. Tan and H.-l. Cheng, "Personality-based adaptation for teamwork in game agents," in *AAAI Conf. on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2007, pp. 37–42.

[18] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," *Cognitive Behavioural Systems*, pp. 114–130, 2012.

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. of the 31st Intern. Conf. on Machine Learning (ICML)*, 2014, pp. 1188–1196.

[20] P.-Y. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," *Intern. Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 157–183, 2003.

[21] R. M. Ryan, "Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory," *Journal of Personality and Social Psychology*, vol. 43, no. 3, pp. 450–461, 1982.

[22] A. Csapo, E. Gilmartin, J. Grizou, J. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock, "Multimodal conversational interaction with a humanoid robot," in *Proc. of the IEEE Intern. Conf. on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2012, pp. 667–672.

[23] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *Intern. Journal of Social Robotics*, vol. 4, no. 2, pp. 201–217, 2012.

[24] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the big-five personality domains," *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, 2003.

[25] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Intern. Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[26] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," in *Proc. of the 2017 ACM/IEEE Intern. Conf. on Human-Robot Interaction (HRI)*. ACM, 2017, pp. 254–262.

[27] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. of the Intern. Conf. on Language Resources and Evaluation (LREC)*, vol. 10, 2010, pp. 2200–2204.

[28] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.