Incremental Acquisition and Reuse of Multimodal Affective Behaviors in a Conversational Agent

Maike Paetzel*
Department of Information Technology
Uppsala University, Sweden
maike.paetzel@it.uu.se

James Kennedy
Disney Research
Los Angeles, USA
james.kennedy@disneyresearch.com

Ginevra Castellano
Department of Information Technology
Uppsala University, Sweden
ginevra.castellano@it.uu.se

Jill Fain Lehman
Disney Research
Los Angeles, USA
jill.lehman@disneyresearch.com

ABSTRACT

To feel novel and engaging over time it is critical for an autonomous agent to have a large corpus of potential responses. As the size and multi-domain nature of the corpus grows, however, traditional hand-authoring of dialogue content is no longer practical. While crowdsourcing can help to overcome the problem of scale, a diverse set of authors contributing independently to an agent's language can also introduce inconsistencies in expressed behavior. In terms of affect or mood, for example, incremental authoring can result in an agent who reacts calmly at one moment but impatiently moments later with no clear reason for the transition. In contrast, affect in natural conversation develops over time based on both the agent's personality and contextual triggers. To better achieve this dynamic, an autonomous agent needs to (a) have content and behavior available for different desired affective states and (b) be able to predict what affective state will be perceived by a person for a given behavior. In this proof-of-concept paper, we explore a way to elicit and evaluate affective behavior using crowdsourcing. We show that untrained crowd workers are able to author content for a broad variety of target affect states when given semi-situated narratives as prompts. We also demonstrate that it is possible to strategically combine multimodal affective behavior and voice content from the authored pieces using a predictive model of how the expressed behavior will be perceived.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Human-centered computing** → *Natural language interfaces*; *Collaborative content creation*; • **Computing methodologies** → Intelligent agents;

KEYWORDS

Affective behavior; multimodal behavior generation; crowdsourcing

ACM Reference Format:
Maike Paetzel, James Kennedy, Ginevra Castellano, and Jill Fain Lehman. 2018. Incremental Acquisition and Reuse of Multimodal Affective Behaviors in a Conversational Agent. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3284432.3284469

1 INTRODUCTION

Autonomous characters that can interact with the same individuals repeatedly over time presuppose extensive cross-domain knowledge and a varied corpus of responses to choose from in all situations. Producing and maintaining such a corpus using expert authors is infeasible given the time and resource constraints of most organizations. In response, crowdsourcing platforms to author agent behavior have become attractive to system builders as a potential solution to the problem of scale [9–11]. While the variability in responses authored by crowd workers ensures conversations feel natural and non-repetitive over time, it simultaneously introduces a different set of challenges: combining dialogue lines generated by hundreds or thousands of crowd workers with different personalities, vocabulary and writing style into an artificial agent so that it feels like interacting with one consistent personality.

In this paper, we investigate whether crowdsourcing can be used to effectively generate multimodal behaviors that create a consistent development of affect over the course of a conversation. In our sample scenario, a Pepper robot [2] is at a science fair in Los Angeles where it asks visitors to help pick a souvenir for its (robot) boyfriend at home. The robot has either an optimistic (**OPT**) or impatient (**IMP**) target personality: **OPT** is intended to be "lighthearted, optimistic, and determined to find the fun in every situation," while **IMP** should be perceived as "quick to overreact with little patience for life's imperfections." We hand-picked two affective states from Russel's affect model [23] that match our scenario narrative. The robot can get *excited* about the visitor's or its own suggestions, or *frustrated* if they cannot agree on a good souvenir to buy. While each personality can make use of both affective states, the affect intensity and development over time would naturally vary between them. Fig. 1 illustrates a potential emotional arc for each personality

*The research project was conducted while the first author was affiliated with Disney Research, Pittsburgh.

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284469

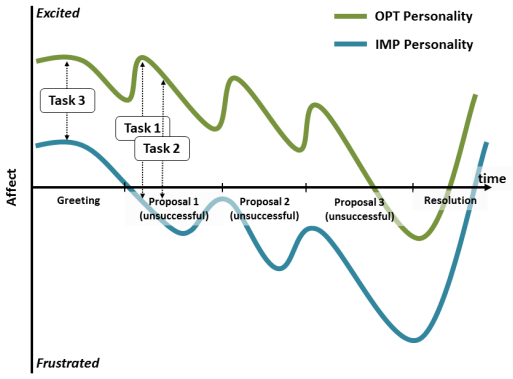


Figure 1: Sample of desired development of the agent's affective state in a task-related dialogue for an optimistic and impatient personality.

when exposed to the same visitor responses. While the **OPT** personality starts out quite excited and has a boost of excitement for every new souvenir suggestion, eventually the agent gets slightly frustrated about continuous failures. The **IMP** personality, on the other hand, begins with neutral affect and changes to a mildly frustrated state after the first bad suggestion by the visitor.

Although a satisfying narrative arc can be authored for the agent in advance, its unfolding is highly dependent on the context and conversational goals. The challenge is to create a corpus of multimodal behaviors that allows choice in the next conversational turn according to the hand-authored target affect, at any point in a given conversation and for any current affective state. Obviously, the size of the target corpus needs to be significantly larger compared to a simpler agent where the affective state is not a constraint. This makes crowdsourcing for corpus generation an attractive approach.

In Sec. 3, we start our investigation by examining whether untrained crowd workers are able to author lines for given target affects and show that crowdsourcing can, indeed, be used to collect such conversational content in different intensities. While the approach we use was successful, it was also expensive, causing us to consider ways we might reduce the size of the corpus without compromising its effectiveness. Thus in Sec. 4, we suggest reusing affective utterances across personalities and find this useful for handling affective states that occur seldom in one personality but more frequently in another. Then, in Sec. 5, we explore a second technique, demonstrating how re-combining authored content with multimodal behavior can be used to broaden the perceived affective intensity of a given line.

2 RELATED WORK

Approaches to generating content for conversational agents have turned from pure rule-based to data-driven in recent years. While all data-driven approaches rely on huge corpora for training, they make use of that content via one of two different paths: either they use full sentences that were authored by crowd workers directly [4, 9–11], or they train stochastic models to produce new language content based on the underlying characteristics of the corpus. Wen et al. [27], for example, use Long Short-Term Memory (LSTM) neural

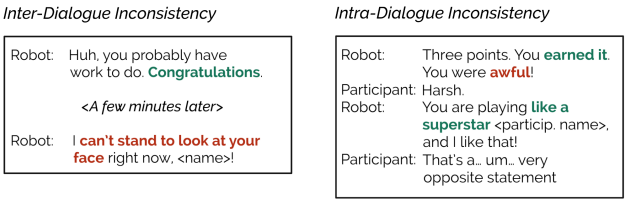


Figure 2: Personality-driven inconsistencies seen in conversations from an earlier data collection [18].

networks to generate naturally varied responses for an information system. Li et al. [14] go further, proposing the use of Maximum Mutual Information (MMI) as the objective function to improve the rather generic responses generated using traditional LSTMs.

Both data-driven approaches have in common that they occasionally introduce inconsistencies when using the data in real interactions. Li et al. [15] mainly discuss inconsistencies in the elicited content. They introduce the idea of two Persona-Models, one for the speaker and one for the addressee to which the speaker can adapt. Breazeal et al. [4] encountered inconsistencies when directly using crowd authored content as well, for example when the robot asked a dialogue partner to execute a certain task in a collaborative scenario and then moments later expressed lack of interest in the outcome of that task. Such inconsistencies produced lower ratings for clear communication and predictability of the agent.

Another type of problem is inconsistency in the agent's personality because "People attribute personalities to conversational agents, strongly influenced by social expectations, whether or not a particular personality was designed deliberately" [5]. To accommodate this reality, in previous research we have explored the use of explicit descriptions of personality when crowd-authoring an agent's conversational content [18]. It was demonstrated that giving crowd workers semi-situated narratives with a personality description could elicit content that was perceived as one of two distinct personalities over repeated face-to-face interactions. Although those results indicated that building more consistent personalities using crowd-sourcing is possible, it was also found that crowd workers had a tendency to overreach personality markers in the authored lines. For example, requests to author a greeting for an **IMP** personality elicited several harsh opening lines. Even though the **IMP** personality was intended to give the overall impression of impatience, an overreaction still requires some trigger to appear natural in a conversation. The use of harsh language without a trigger event led to *inter-dialogue inconsistencies* as shown in Fig. 2 (*left*). The problem is not exclusively a distance effect across turns. Because the agent had no measurement for the strength of personality markers in the given lines or the affective state that would be communicated, recombination of their efforts *in situ* also led to *intra-dialogue inconsistency*, with the agent's affect changing rapidly within turns, as in Fig. 2 (*right*).

The reason for those inconsistencies seems to lie in the nature of the narratives given to the crowd workers; in order to use elicited lines in multiple, albeit similar contexts, the authors are presented with only very limited history and thus cannot ensure a consistent development of affect within the personality. Hence, the agent

needs an explicit model of affect it can present to the authors in order to resolve inconsistencies in the affective state.

Prior studies in human-robot interaction (HRI) have looked at signaling affect through the combination of conversational content, non-linguistic utterances, and full-body behaviors [6, 7, 22, 24, 25]. However, the integration has predominantly been hand-authored by experts, which does not scale to large domains. Crowdsourcing affective behavior has been used before (see [17] for a discussion), mostly in video data collection, but also in written content creation and by assigning affective labels to given stimuli. Ravenet et al. [21] developed an online system to allow crowd authoring of non-verbal behavior for a virtual agent's attitude. While they could find patterns in the facial expressions and gestures used by crowd authors, they did not focus on the perception of the created behavior.

In the studies that follow, our goal is to efficiently extend a conversational agent with an explicit model of affective state by combining concepts from the literature on signaling affect in HRI and crowdsourcing conversational content. We thereby contribute to solving the problem of inconsistencies in data-driven approaches for dialogue content generation and provide useful insights for both developers designing traditional agents and those already using crowdsourcing techniques.

3 AUTHORING AFFECTIVE LANGUAGE

The overarching aim of our work is to develop and evaluate a method of eliciting multimodal affective behavior using crowdsourcing. In this paper we focus in particular on the question of consistency of expression. While it has been shown that untrained crowd workers are generally able to author dialogue content [9–12], and can, further, author personality-specific content [18], it is unclear if we can elicit enough variety in the affect to follow a logical set of affective transitions, for example as in Fig. 1. Thus, we start our investigation with the following research question:

RQ1 Given personality descriptions with specific affective states as part of a narrative, can crowd workers generate dialogue lines covering a broad spectrum of affect intensity?

3.1 Method

The context for the study is a collaborative scenario in which the communication of the robot's satisfaction level is core to solving the task. The narrative places the robot in a booth at a science fair with the goal of engaging visitors in a conversation. After a brief phase of chit-chat, the robot introduces the task to be solved, which is picking a souvenir to bring home to the robot's boyfriend. To create a satisfying experience, the dialogue should follow the affect graph as in Fig. 1. A three-stage crowdsourcing pipeline similar to [12] was used to automatically generate the affective content:

1. Dialogue Authoring: Crowd workers were given a narrative as shown in Fig. 3 and asked to author a line to continue the conversation for a given personality (**IMP** or **OPT**) and affective state. The affective state consisted of an affect type (excited or frustrated) and a modifier ("not at all", "slightly", "extremely"). In some narratives, the affective state was omitted to serve as a baseline for comparison (the personality was still shown, so authors provided their own interpretation of the current affect). In [Task1] crowd workers were explicitly given directions for the robot's response (*accept* or *reject*), while in [Task2] and [Task3] they could invent content as long as they satisfied the personality and affect constraints.

2. Dialogue Editing: Because the content is authored by untrained crowd workers, quality must be checked before using the work in an interactive agent. Thus, a second set of crowd workers saw the same narrative, personality and affect descriptions and were asked to rate how *ordinary and typical* each line was for the given personality. The rating was performed on a five point Likert scale, with an additional possibility to mark the line as *nonsensical*. Crowd workers were given 10 lines to rate per task and every line was rated by 30 different workers. Lines that were rated as nonsensical by more than 5 raters (20%) or that received an average rating of lower than 2.66 were discarded.

3. Affective Rating: The first two stages of the pipeline produce lines that make sense in context and are intended to express a desired level of the named affect for the specified personality. The final stage evaluates how the level of affect will actually be perceived. In particular, crowd workers read the narrative and description of personality, but were not given the original target affect. They were asked first to classify the dialogue line as excited or frustrated, and then to rate the degree of the selected emotion on a scale from *not at all* to *extremely*.

3.2 Evaluation

We select three distinct points at which to evaluate the robot's potential affective responses in the chosen conversational scenario (Fig. 1, Fig. 3):

- **[Task1]**: The visitor makes the first proposal of a potential gift, which the robot might decide to reject or accept.
- **[Task2]**: The robot makes a counter-proposal, which is in turn either liked or disliked by the visitor.
- **[Task3]**: A short episode of chit-chat about the convention hall and the weather outside, which would be mostly used in the beginning of the conversation.

[Task1] and [Task3] are designed to explore the full range of affective responses, while in [Task2] the content of the visitor's response implies the robot's logical reaction to it. For every condition, we also generated narratives without an explicit target affect. Those narratives were designed to enable comparison of the range of affect elicited with and without an explicitly stated target affect. In total, the design leads to 58 unique task descriptions. For each task, 30 crowd workers authored a line for the robot to continue the conversation, resulting in 1740 lines for our evaluation.

To understand if the 1740 lines elicited in Stage 1 are appropriate for the given personality and successfully convey the target affect state, all lines were evaluated by 30 crowd workers each in Stage 2 (editing) and Stage 3 of the pipeline (affective rating). Every assignment contained 10 content lines to rate, leading to 174 unique Human Intelligent Tasks (HITs). This results in 10440 assignments (174 HITs · 30 assignments each · 2 stages) in total to answer RQ1.

3.3 Pre-Processing

Frustration and excitement are not two opposing affects on a continuous scale and were not presented to crowd workers as such. Even though every affective state and intensity was mapped from

<table>
<tr><th></th><th colspan="6">[Task1] In-task Phase 1</th><th colspan="4">[Task2] In-task Phase 2</th><th colspan="3">[Task3] Chit-Chat</th></tr>
<tr><td>Narrative</td><td colspan="6">A visitor at a technology convention in Los Angeles, California has been chatting with this robot in one of the booths.
They have been talking about the robot's boyfriend (also a robot).
Robot: I want to bring a souvenir from my trip here back home to my boyfriend, but I'm not sure what to get him.
Visitor: Hm, what about a Laker's T-shirt?</td><td colspan="2">Robot: I don't think he is into basketball that much, but how about an "I love LA" shirt?
Visitor: Oh boy, that's such a cliché.</td><td colspan="2">Robot: I don't think he is into basketball that much, but how about an "I love LA" shirt?
Visitor: Yeah, that's a really good idea!</td><td colspan="3">They have been talking about how busy the convention hall is.
Robot: There are certainly a lot of people here today.
Visitor: I know, it's getting pretty crowded. Must be because it's such a nice day outside!</td></tr>
<tr><td>Task</td><td colspan="3">How does the robot accept this recommendation and discuss it some more?</td><td colspan="3">How does the robot reject this offer and suggest a different gift?</td><td colspan="2">How does the robot continue with the topic?</td><td colspan="2">How does the robot continue with the topic?</td><td colspan="3">How does the robot continue talking about the weather?</td></tr>
<tr><td>Affect (& Modifier)</td><td>Excited extremely, slightly, not at all</td><td>Frustrated extremely, slightly, not at all</td><td>None</td><td>Excited extremely, slightly, not at all</td><td>Frustrated extremely, slightly, not at all</td><td>None</td><td>Frustrated extremely, slightly, not at all</td><td>None</td><td>Excited extremely, slightly, not at all</td><td>None</td><td>Excited extremely, slightly, not at all</td><td>Frustrated extremely, slightly, not at all</td><td>None</td></tr>
<tr><td>Personality</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td><td>OPT, IMP</td></tr>
<tr><td>Total Count</td><td colspan="6">28</td><td colspan="4">16</td><td colspan="3">14</td></tr>
</table>

Figure 3: Narratives, descriptions, target personality and affect for all three tasks used in our evaluation. In total, this leads to 58 unique combinations. Note that "Laker's" remains uncorrected, as collected from the crowd workers.

the crowd workers' qualitative scale to a unique number – frustration ranging from 1 (not at all) to 5 (extremely) and excitement ranging from 6 (not at all) to 10 (extremely) – we do not summarize the data across the two types of affect. Instead, we first compute the dominant affect judgment for a line, then compute the average and variance only over the ratings for that affect type. Thus, if a line was dominantly rated as excited/frustrated, the average and variance is only computed over those ratings that were performed on the respective scale. Those lines with an agreement of less than 60% were labeled as "undecided". This preprocessing stage reflects our intended purpose: a general method by which we can generate a large corpus of utterances to be deployed effectively across any of a set of given emotional arcs. The ability to select the subset of the data that best serves that purpose is one of the strong motivations for using an overly generative approach like crowdsourcing.

3.4 Results

We evaluate the success of crowd authoring affective content using three metrics. First, we analyze how many utterances are rejected in the editing stage. While including the affective state explicitly alongside the personality enables more efficient authoring, it may also make the task too difficult for untrained crowd workers. As a consequence, a high rejection rate would make the pipeline inefficient and impractical for an actual agent interaction. We find that *adding a target affect to the narrative does not lead to a significantly higher rejection rate and thus does not decrease the efficiency of the crowd authoring pipeline.* 5.0% of the lines without a target affect, 6.5% authored for frustration, and 6.9% for excitement were rejected. Moreover, a 3x2x3 ANOVA analysis with task ([Task1], [Task2], [Task3]), personality (**IMP**, **OPT**) and affect (frustrated, excited, none) as factors showed no effect of the affect-driven (N = 1440) versus non-affect-driven narratives (N = 300), $F(2, 1722) = 0.581$, $p = .506$, and the personality, $F(1, 1722) = 1.391$, $p = .238$.

Next, we consider whether crowd authors were able to convey the intended target affect in the utterances they produced. In spite of some problems with the interpretation of the modifier "not at all", *the perception of the majority of content lines matched the original task description.* We performed the same 3x2x3 ANOVA analysis on the results from the affective rating stage after filtering out the lines that were rejected in the editing stage. There was no significant influence of the type of affect on the recognition of the intended affect, $F(1,1331) = 0.352$, $p = .553$. However, only two thirds of all utterances (63.4% for excited, 64.9% for frustrated) got the intended affect assigned and only a third of all lines that were authored using the modifier "not at all" were recognized as intended. This suggests that crowd workers interpreted "not at all excited" to be rather frustrated and vice versa. If we only consider the other two modifiers, 80% of the lines were ascribed the intended affect. While the affect alone does not have a significant influence on the recognition rate, the interaction between personality and affect is significant, $F(1, 1331) = 73.128$, $p < .001$. A Tukey Post-Hoc analysis suggests that those lines that were authored with less natural combinations of personality and affect (frustration for **OPT** and excitement for **IMP**) were recognized correctly less often. The ANOVA analysis showed an additional significant influence of the task on the recognition rate, $F(2, 1331) = 12.018$, $p < .001$: the intended affect was significantly less often ascribed in [Task1] compared to both [Task2], $p < .001$, and [Task3], $p < .001$, which is likely due to specifically low recognition rates for lines with very unnatural task descriptions, like "reject extremely excited".

Finally, we show that *the modifiers manipulate the perceived affect as intended.* An ANOVA analysis with the affect modifier as a factor was performed separately for the lines predominantly perceived as excited and frustrated. For both affects, the modifiers had a significant influence on the average affect rating. For excitement, $F(2, 583) = 110.1$, $p < .001$, the modifier "extremely" (M= 8.4, SD= 0.57) was perceived as significantly more excited than "slightly" (M= 7.68, SD= 0.56), $p < .001$, and "not at all" (M= 7.61, SD= 0.57),

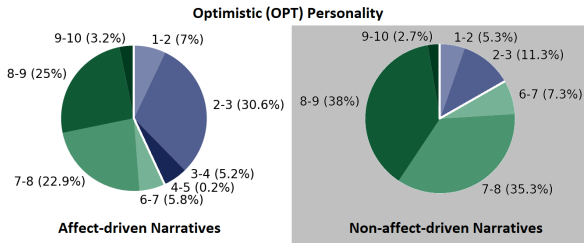


Figure 4: Distribution of the content lines according to their perceived affect type and intensity for the OPT personality.

$p < .001$. The difference between "slightly" and "not at all" is not significant, $p = .414$. Results for frustration are similar, $F(2, 602) = 67.12$, $p < .001$: lines authored with the modifier "extremely" were rated significantly more frustrated (M= 2.83, SD = 0.55) than both the lines authored with "slightly" (M= 2.46, SD= 0.42), $p < .001$, and "not at all" (M= 2.32, SD= 0.39), $p < .001$. The difference between "slightly" and "not at all" is also significant, $p = .006$.

The results suggest that crowd workers are indeed able to successfully author agent responses for a given target affect. However, our approach certainly puts additional workload on the authors. In order to justify this, we need to be able to generate a broader variety of target affect intensities with affect-driven narratives in comparison to traditional, non-affect-driven narratives. While the overall distribution across both personalities and all tasks is comparable, our approach is favorable when it comes to authoring affective states that are counterintuitive for a given personality or task. Fig. 4 demonstrates the distribution of perceived affect for lines authored for the **OPT** personality using affect-driven and non-affect-driven narratives over all three task conditions. With non-affect-driven narratives, we elicited few frustrated lines on the lower end of the intensity scale. Only with the affect-driven narrative were we able to collect the full spectrum of affect intensities. In other words, if we do not give crowd workers an affect we get the default for the personality given the context. If we do give them an affect we can get the full range, but the further we get from the default, the more difficult the task and higher the rejection rate.

4 REUSING CONTENT ACROSS PERSONALITIES

In the previous section, we found affect-driven narratives to be promising for collecting a corpus of dialogue lines with a broad variety of affect states and intensities. While we were able to show that we can elicit dialogue content even for less obvious combinations of personality and affect, like frustration for the **OPT** personality, we also need substantially more resources to elicit such an agent's responses to cover the full range of affect. To make our approach more attractive given limited resources, we want to decrease the number of required content lines by not exhaustively generating content for the full space of affect and situation combinations. One solution to keep a conversation going in a new context is to reuse content from similar, known situations [8]. Introducing personality and affect potentially restrict which content can be borrowed: if a frustrated affect is to be conveyed, it is unlikely that a line elicited in an excited context can be used. However, as shown in Fig. 1, certain affective states are more common for some personalities than others. Thus, we aim to reuse content across personalities and predict how that influences the perceived affective state.

Table 1: Affect perception for reused conversational content.

	original affect	perceived when reused: excited	frustrated	undecided
IMP	excited	257	6	8
	frustrated	14	340	33
OPT	excited	398	8	36
	frustrated	5	222	14

RQ2 Can affective conversation content be reused across personalities and if so, how does reuse influence the perceived affect of the content?

A 3x2x3 ANOVA analysis of the results from the affective rating stage on the data used in Sec. 3 shows that the *personality has a significant influence on the perceived affect of a content line.* For frustration, $F(1,593) = 41.647$, $p < .001$, lines authored for the **IMP** personality were on average perceived as an intensity of 2.62 (SD = 0.52), while lines authored for the **OPT** personality only received a 2.39 (SD= 0.46). Similarly for excited lines, $F(1, 574) = 8.760$, $p = .003$, those authored for the **OPT** personality were on average perceived as an intensity of 7.95 (SD= 0.66) and lines for the **IMP** personality as 7.82 (SD= 0.67). If those differences are systematic, an agent could successfully reuse lines across personalities.

4.1 Evaluation

To evaluate the feasibility of reusing content across personalities we use the same 1740 Stage 1 lines elicited in Sec. 3.2. However, for Stages 2 and 3 in the pipeline, we exchanged the description of personality: lines that were originally authored for the **OPT** personality were ascribed to **IMP** and vice versa. All lines were evaluated by 30 crowd workers in both of the latter stages, leading to 10440 assignments each containing 10 lines (174 HITs · 30 assignments each · 2 stages) to answer RQ2.

4.2 Results

For lines that were neither rejected for the original personality nor for the one reusing the line, a Wilcoxon signed rank test showed *no significant difference in the recognition rate of the target affect*, $V = 2050$, $p = .088$. When reused, 69.24% of the lines that were authored for excitement (82.95% without "not at all" modifier) and 64% that were authored for frustration (85.07% without "not at all" modifier) were also recognized as the intended affect description in the rating stage. A Wilcoxon signed rank test also revealed that the *perception of affect intensity is not significantly different between the original personality and the one reusing the line*, $V = 580680$, $p = .128$. In addition, Table 1 demonstrates that there are only a few lines which switch the perceived affect when being reused.

When excluding lines that were rejected for the original personality description and those that switch the perceived affect, the original perceived affect is the closest approximation to predict the affect intensity when reused (Fig. 5). The overall error calculated using ten-fold cross-validation is 0.24, which is sufficiently high

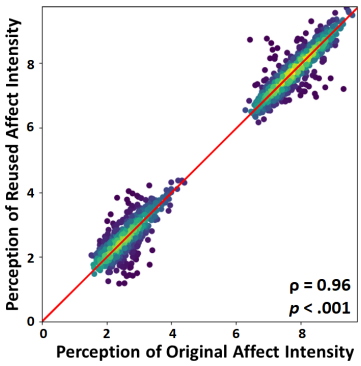


Figure 5: Spearman rank-order correlation between the perception of affect in the lines with the original and switched personality. The brighter the color of a data point the higher the density of observations at that point.

considering 99.7% of all content lines with the original personality have a higher variance between the individual ratings per line.

Even more important than conveying the intended affect is that an agent does not use lines that are inappropriate and thus rejected for a given personality and context. A paired Wilcoxon signed rank test showed that *there is a significantly higher reject rate when reusing lines across personalities*, $V = 3960$, $p < .001$. For the original personality description 6.4% of lines were rejected in the editing stage, while 10.3% are discarded if the other personality reuses the same line. If we assign the authored line to the opposite personality, both the affect, $F(2,1722) = 5.665$, $p = .004$, personality, $F(1,1722) = 28.305$, $p < .001$, and interaction between them, $F(2,1722) = 9.933$, $p < .001$, have a significant influence. Interestingly, 15.33% of the lines without any given affect are rejected, which is significantly more than both for the excited (10.1% rejected, $p = .03$) and the frustrated affect (8.5% rejected, $p = .002$). 14.14% of lines originally authored for **IMP** are rejected when used for the **OPT** personality, while only 6.55% originally authored for **OPT** are rejected when used by the **IMP** personality. Lines originally authored for the **IMP** personality with no affect description given were rejected significantly more often than all other combinations of affect and personality, $p < .001$ for all combinations. This is likely because those lines are quite harsh without an explanatory target affect given, so people find them inappropriate when assigned to the **OPT** personality.

5 MULTIMODAL COMMUNICATION OF AFFECT

In Sec. 3 we demonstrated that it is possible to use a crowdsourcing approach to elicit affective conversational content. By borrowing utterances across personalities (Sec. 4), we can decrease the burden on crowd authors by avoiding unintuitive combinations of personality and affect. This also lowers the resources necessary to effectively use this approach.

In a situated interaction, however, the perception of an utterance is not just influenced by its content. The voice tone and non-verbal behaviors used by an agent can potentially increase or decrease its perceived affective state. This is interesting for system builders, because even if borrowing across personalities, we cannot expect the agent to always have a novel content line specifically for the target affect level available in any situation. Thus, we are interested in understanding how enriching the dialogue content with multimodal behavior can influence the perception of the content and whether we can strategically use the multimodal delivery of a line to broaden our set of available affect intensities to meet the desired emotional arc. And, again, if we are able to predict the perceived affect of the re-situated multimodal content based on the perception of the separate parts involved, we may greatly decrease the required crowdsourcing resources.

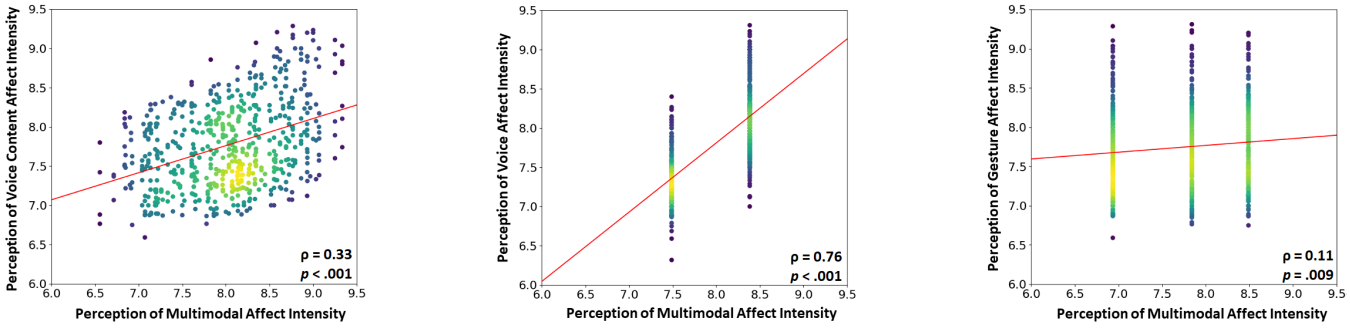
RQ3 How can the multimodal delivery of a dialogue line be used to broaden the number of available affective states of the agent?

5.1 Method

To explore RQ3, we use the humanoid robot Pepper [2] as a platform with a typical level of expressiveness. Apart from verbal content, Pepper's voice tone and full-body movements can be controlled to enhance the affect being expressed.

Voice tone: Pepper's default voice is rather high-pitched and fast, which according to [20] is characteristic for a happy voice. By making some modifications, we aimed to create three voices with different affective connotations. The default voice was taken as our neutral baseline, the *excited* voice made use of the 'joyful' voice tag, which creates an even higher pitch and different voice contour, while the *frustrated* voice was manually given a lower pitch. We evaluated the success of our voice tone design in a pilot study with 4 different HITs and 30 assignments each. The same HIT type described in Sec. 3.2 was used, but with a video instead of written stimulus. Videos were recorded in which Pepper synthesized one of 10 different sentences per personality in all three voice tones. The sentences were originally authored for [Task1] and [Task3] and selected so that a large variety of different affect intensities from Stage 3 are represented. A two-way ANOVA with personality and voice tone as independent variables showed a significant influence of the voice tone, $F(2, 54) = 10.38, p < 0.001$, and no influence of the personality. The excited (**IMP**: $M = 8.37$, **OPT**: $M = 8.38$) and frustrated voice (**IMP**: $M = 2.89$, **OPT**: $M = 2.72$) were predominantly perceived as intended. The neutral voice was within the "undecided" spectrum, with a slight tendency towards excitement (**IMP**: $M = 7.52$, **OPT**: $M = 7.49$).

Full-body movements: Affective full-body gestures were generated by an expert artist with professional experience in animated character design. The artist was instructed to create expressive behaviors that convey frustration and excitement that could be used with congruent conversational content. Behaviors were designed using the virtual embodiment in Choregraphe [1]. From the gestures created by the artist, five were selected in a second online pilot study: two excited and two frustrated gestures, each with different intensities, and one neutral behavior. Again, 4 HITs (one per personality and either one of [Task1] or [Task3]) with 30 assignments each were used to elicit ratings for the perceived affect communicated by the gestures. Using a two-way ANOVA, we found only an influence of the gesture, $F(7, 16) = 6.22, p = 0.001$, and no influence of personality.



(a) Conversational content **(b) Voice tone** **(c) Gesture**

Figure 6: Spearman rank-order correlation of the perceived multimodal affect and the conversational content, tone and gesture for the OPT personality and "excited" affect as a sample. Brighter colored data points represent higher density of observations.

5.2 Evaluation

We use the 433 dialogue lines elicited for [Task2] (480 original lines minus 47 rejected in Stage 2) to evaluate the multimodal delivery of dialogue content. For each of the lines we determine the perceived dominant affect in the lines (Sec. 3.3). We then generate the multimodal combinations which contain matching affective markers for the final evaluation: 214 lines of excited conversational content are combined with both the neutral and two excited gestures. Each of the three combinations is combined with the neutral and excited voice tone, leading to 6 combinations to be evaluated. Similarly, 187 lines of frustrated content are combined with the neutral and two frustrated gestures and the neutral and frustrated voice tone, leading to 6 combinations per line. 32 lines of content that are neither dominantly excited nor frustrated are combined with both excited and frustrated gestures and voice tone, leading to 11 combinations.

In total, we evaluated 2758 multimodal combinations of conversational content, tone, and gesture in the affective rating stage. Again, 30 ratings were received per combination. The stimuli were presented to crowd workers via video scenes with written transcripts of the content. The robot was filmed in front of a dark blue background, and professional lighting ensured good visibility with minimal reflections on the robot's body. A microphone invisibly attached above the robot's head accurately recorded its voice.

5.3 Results

We first consider the multimodal combination of those dialogue lines where the content was originally perceived as dominantly excited or frustrated (not undecided). Since we combined the conversational content only with matching and neutral multimodal behavior, we expected that people would dominantly perceive the same affect type as the conversational content when judging the multimodal delivery of lines. Indeed, 92.56% of all lines matched the perceived affect type of the conversational content. Less than 1% (32/2758) of lines switched affective label; the remainder were not clearly rated as excited or frustrated. Considering the 32 neutral lines, 31 were indeed perceived as dominantly excited or frustrated depending on the multimodal delivery. In 29.5% of all combinations of content, voice tone and gesture, the multimodal delivery of an undecided line is undecided again.

The results suggest that *multimodal delivery is a practical approach to broaden the perceived intensity of affect for a single line of content.* The average difference between the lowest and highest perceived average intensity for the same content line is 0.77 for **IMP** and 0.93 for **OPT**. For those lines that did not switch from one affect type to the other, we examined the difference between the perception of the conversational content and the multimodal perception of the same utterance. In 26.4% of lines the multimodal delivery changes the perception only towards the less intense end of the scale. In 16.5% the perception is only changed towards the more intense end. In 57.1% the multimodal delivery can be perceived as both more and less intense, depending on the exact combination of voice tone and gesture. From those lines, the multimodal delivery ranges between -0.38 and +0.47 for **IMP** and -0.6 and +0.43 for **OPT** compared to the intensity of the conversational content alone.

For those lines that are dominantly perceived as excited or frustrated and that get the same affect type assigned in the multimodal delivery as in the conversational content, we use linear regression to predict the perception of the multimodal delivery. The linear regression receives the average rating, the variance and the percentage of raters picking the dominant scale for the specific conversational content, voice tone, and gesture as an input. Fig. 6 shows that the conversational content, voice tone, and gesture are all significantly correlated with the perception of the delivered multimodal line for the **OPT** personality and "excited" affect.

For excited conversational content, the prediction error determined using ten-fold cross-validation is 0.25. For frustrated content, the prediction error is 0.27. For both affect types, the prediction error is lower than the naive baseline which always predicts the affect intensity of the conversational content (excited: 0.58 average error, frustrated: 0.37). With the average variance of ratings for the multimodal stimuli being 0.79 and less than 1% of stimuli having a variance less than 0.27, our prediction error falls within the natural variance of almost all the stimuli in our training set.

6 DISCUSSION

Crowdsourcing conversational content for artificial agents is an attractive approach to fulfill the demand of novel responses over time and in different domains. However, a common problem in

such architectures are inconsistencies introduced by a variety of different authors with different imaginations of what it means for an agent to react appropriately in a given situation. In this paper, we propose a method to minimize inconsistencies when it comes to conveying affective state. Our results show that untrained crowd workers produce a variety of dialogue content with different affect types and intensities both for affect-driven and non-affect-driven narratives (RQ1). However, the spectrum of affect is broader with affect-driven narratives.

In this proof-of-concept, we picked three points in a sample conversation that differed in content and constraints given to the authors. Including pilot runs, a total of 2483 crowd workers participated in the studies presented in this paper. Our results suggest that the less people can relate to the context description given, the more difficulty they have in authoring for the given constraint. When talking about the weather, significantly fewer lines (2.86%) were rejected as inappropriate compared to the other contexts ([Task1]: 6.31%, [Task2]: 9.79%), $F(2,1722) = 9.699$, $p < .001$, presumably because being frustrated or excited about the weather is something most people can easily relate to. Indeed, [Task1], when workers were asked to reject a gift excitedly or accept a gift in frustration, was the case in which the most lines were rejected and most mismatches between target affect and perceived affect were observed. Again, we believe that people have less personal experience with such behavior and thus struggle to author these content lines in a meaningful way. This observation does not relate only to authoring for target affect, but speaks to an inherent violation of the assumptions underlying semi-situated learning: the ability of the crowd worker to substitute his or her own experience for the agent's.

Results obtained in the second and third study further suggest that both reusing lines across personalities and enriching dialogue content with multimodal affective behavior can decrease the necessary corpus size to implement our suggested model in an actual agent. If the affect type does not change when re-using content or re-combining multimodal content delivery, we are able to predict the perceived affect intensity with a sufficiently high precision. The main challenge remaining is to predict the infrequent occurrence of affective "flips" - those lines that change the perceived type of affect from frustration to excitement or vice versa when re-used in another personality or when being re-combined with voice tone and gesture. The fact that those flips occur very rarely is beneficial for the agent, because it means that it is unlikely to occur in a conversation and if it does, it likely will not happen more than once. However, an agent that uses a line with flipped affect potentially breaks the illusion of agency and intelligence, a fatal flaw in maintaining believability [16]. Unfortunately, the class imbalance makes it difficult to predict such flipped lines with high precision and recall and the same holds for predicting lines that are rejected when being reused by another personality. Future work is needed to explore more sophisticated approaches to this problem.

Before the approach proposed in this work can successfully be implemented in an agent, we also need to understand how semi-situated perception online differs from fully-situated perception face-to-face, especially when it comes to the multimodal delivery of lines. Only if we can predict the situated perception of affect with high precision can we follow a desired emotional arc such as that shown in Fig. 1. While Li [13] suggests that physical presence does generally influence peoples responses to an agent, more recent work suggests that social presence is more important than physical presence [26]. In a related study on recognizing emotions in Lego figures, Bartneck et al. [3] found some differences between the ratings given by crowd workers and those by on-site participants. The majority of raters still agreed on the dominant emotion across conditions, however, and Bartneck et al. argue that the observed differences are mostly practically irrelevant. This gives confidence that, even though perception *in situ* might differ, a mapping between semi-situated and situated perception can be learned.

Our approach uses crowd authoring only in the generation of conversational content. The decision of how affect should change over the course of the conversation, as well as the design of voice tone and expressive gestures, was done by experts. Nevertheless, we believe that our approach scales to using crowdsourcing in other parts of the system. With sufficient tools similar to [19], crowd workers could alter voice tone and gestures to convey certain affect using the same narratives. In addition, it would be interesting to explore if we can automatically learn triggers for positive and negative affect changes using generic narratives without a personality or affect given. By asking crowd workers for their own natural response and then analyzing the affective rating, we could potentially automatically detect triggers to change the affective state. Further exploration is necessary to support these ideas.

In the future, we aim to fully expand the robot's knowledge graph using affect-driven narratives and test our approach in a long-term, face-to-face interaction to understand if the agent's affect is rated as consistent and if people find the changes in affect natural. It would also be interesting to evaluate how our findings translate to even more expressive embodiments and multiple affect types and how that might change the weight of gestures in the perception of multimodal lines.

7 CONCLUSION

In this paper, we show that untrained crowd workers can successfully author affective dialogue content given affect-driven narratives. This content can be combined with affective voice tone and full-body gestures designed by an expert to broaden the perceived affective intensity of the content. Using a simple linear regression, we can predict the perceived target intensity of the multimodal line, which drastically decreases the required crowdsourcing resources. The spectrum of intensity given those two methods is sufficiently large to react with a natural affect in every context in our sample scenario. While crowd workers author content for a target affect that is uncommon for a given personality less successfully, we introduced and evaluated reusing of content across personalities as a suitable solution to this problem. This work serves as a proof-of-concept, making crowdsourcing for autonomous conversational agents more attractive to designers by helping to overcome inconsistencies in the communicated affect.

ACKNOWLEDGMENTS

The authors would like to thank Sammie Ma for designing the affective behaviors, Sandy Fox and Pedro Mota for input on the initial project plan and Maarten Bos for significant infrastructure support in conducting the studies.

REFERENCES

[1] [n. d.]. Choregraphe. http://doc.aldebaran.com/2-5/software/choregraphe/index.html. Accessed: 2018-10-11.

[2] [n. d.]. Pepper. https://www.softbankrobotics.com/emea/en/pepper. Accessed: 2018-10-11.

[3] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. 2015. Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PloS one* 10, 4 (2015), e0121595.

[4] Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung. 2013. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction* 2, 1 (2013), 82–111.

[5] Andreea Danielescu and Gwen Christian. 2018. A Bot is Not a Polyglot: Designing Personalities for Multi-Lingual Conversational Agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, CS01:1–CS01:9.

[6] Rachel Gockley, Reid Simmons, and Jodi Forlizzi. 2006. Modeling affect in socially interactive robots. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 558–563.

[7] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* 4, 4 (2013), 341–359.

[8] James Kennedy, Iolanda Leite, André Pereira, Ming Sun, Boyang Li, Rishub Jain, Ricson Cheng, Eli Pincus, Elizabeth J Carter, and Jill Fain Lehman. 2017. Learning and Reusing Dialog for Repeated Interactions with a Situated Social Agent. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 192–204.

[9] Michael Kriegel. 2015. *Towards A Crowdsourced Solution For The Authoring Bottleneck In Interactive Narratives*. Ph.D. Dissertation. Heriot-Watt University.

[10] Walter S. Lasecki, Ece Kamar, and Dan Bohus. 2013. Conversations in the Crowd: Collecting Data for Task-Oriented Dialog Learning. In *First AAAI Conference on Human Computation and Crowdsourcing*. The AAAI Press, 2–5.

[11] Walter S. Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James Allen, and Jeffrey Bigham. 2013. Chorus: A Crowd-Powered Conversational Assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*. ACM, 151–162.

[12] Iolanda Leite, André Pereira, Allison Funkhouser, Boyang Li, and Jill Fain Lehman. 2016. Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 13–20.

[13] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[14] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 110–119.

[15] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. ACL, 994–1003.

[16] A. Bryan Loyall and Joseph Bates. 1997. Personality-Rich Believable Agents That Use Language. In *Proceedings of the First International Conference on Autonomous Agents*. ACM, 106–113.

[17] Robert R. Morris and Daniel McDuff. 2014. Crowdsourcing techniques for affective computing. In *The Oxford Handbook of Affective Computing*. Oxford Univ. Press, 384–394.

[18] Pedro Mota, Maike Paetzel, Andrea Fox, Aida Amini, Siddarth Srinivasan, James Kennedy, and Jill Fain Lehman. 2018. Expressing Coherent Personality with Incremental Acquisition of Multimodal Behaviors. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 396–403.

[19] Magalie Ochs, Brian Ravenet, and Catherine Pelachaud. 2013. A crowdsourcing toolbox for a user-perception based design of social virtual actors. In *Computers are Social Actors Workshop (CASA)*.

[20] Pierre-Yves Oudeyer. 2003. The Production and Recognition of Emotions in Speech: Features and Algorithms. *International Journal of Human-Computer Studies* 59, 1 (2003), 157–183.

[21] Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 263–274.

[22] Robin Read and Tony Belpaeme. 2016. People interpret robotic non-linguistic utterances categorically. *International Journal of Social Robotics* 8, 1 (2016), 31–50.

[23] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[24] Martin Saerbeck and Christoph Bartneck. 2010. Perception of affect elicited by robot motion. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 53–60.

[25] Matthias Scheutz, Paul Schermerhorn, and James Kramer. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 226–233.

[26] Sam Thellman, Annika Silvervarg, Agneta Gulz, and Tom Ziemke. 2016. Physical vs. Virtual agent embodiment and effects on social interaction. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 412–415.

[27] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL, 1711–1721.