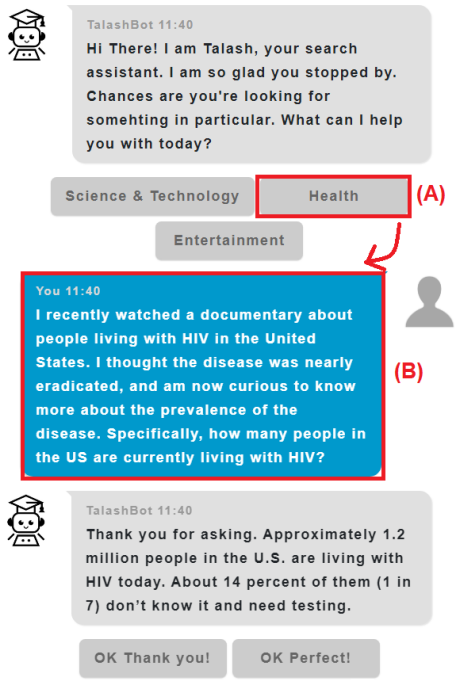

Figure 1: User Interface of Talash bot. It is designed to answer IR queries in three domains – Health, Science & Technology and Entertainment. (A): Talash bot initiates the dialogue by displaying some predefined domains. (B): After selecting a domain, Talash then replaces quick replies with a predefined search task. Finally, it displays the predefined answer after a set response time is elapsed.

How Do User Moods Affect Perceived Delays in Crowd-Powered Conversational Interactions?

Tahir Abbas
Eindhoven University of Technology
Eindhoven 5600 MB, Netherlands
Mirpur University of Science & Technology
Mirpur AJK, Pakistan
t.abbas@tue.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

Panos Markopoulos
Eindhoven University of Technology
Eindhoven 5600 MB, Netherlands
p.markopoulos@tue.nl

ABSTRACT

Crowd-powered conversational systems (CPCS) are gaining considerable attention due to the ease with which they can be deployed for a range of domains without any substantial training costs. On the downside, CPCSs currently suffer from long response delays, which hampers their potential as conversational partners for real users. Furthermore, time perception theories are equivocal regarding the impact of user's mood states on the perceived delays. Thus, our current research examines the combined influence of response delays and mood upon the perceived latency of CPCS in an information retrieval context.

KEYWORDS

Crowd-powered conversational system, time perception, system response time, emotions, chatbots

INTRODUCTION & BACKGROUND

Due to the shortcomings of current artificial intelligence (AI) techniques and natural language understanding, automated methods are not yet capable of dealing with the complexity of conversational

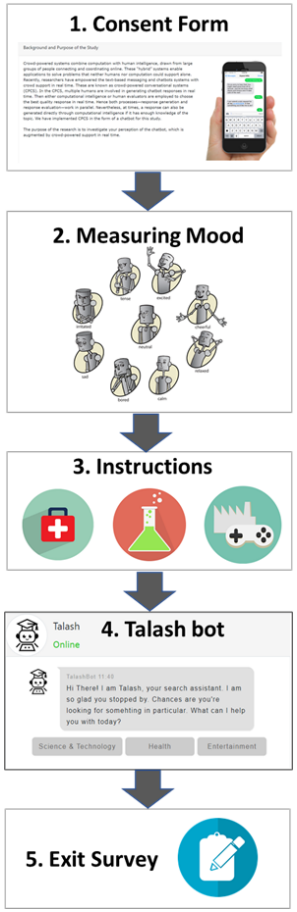

Figure 2: Steps involved in the Procedure

interactions, often resulting in conversation breakdowns [2]. Crowd-powered conversational systems (CPCS) [10, 14] have been proposed as a remedy to these shortcomings of AI. CPCSs proposed recently make use of sophisticated recruiting, rewarding and user interface techniques to reduce latency of crowd input from hours to few seconds [3]. A pioneering example is Chorus [14], which is a text-based conversational agent that assists end-users with information retrieval tasks by conversing with an online group of workers in real time. Since such CPCS rely (fully [14] or partially [10]) on the cohort of humans, significant response delays are expected that might induce frustration in end-users. The vast majority of work in the CPCS area has focused mainly on reporting the "actual" or system response latency [See pioneer works in CPCS: 4, 11, 12, 14], but relatively little is understood concerning the "perceived" latency from the users' point of view. The advantages of understanding the perceived waiting times in CPCS are twofold: 1) we can understand the upper limits acceptable for waiting time in CPCS, which is unknown as of yet. 2) Based on acceptable waiting time, we can design interventions, such as time fillers to counter the negative effects of waiting in CPCS. This apparent gap in our current understanding leads naturally to our first research question (**RQ1**): *How do end users perceive the response latency of varying length while interacting with CPCS?*.

Prior research has produced contradictory results concerning the influence of emotions on time perception. Research in time perception and emotions has supported the hypothesis that being in the unpleasant mood states can lead to an extended sense of elapsed time as compared to being in neutral states [9]. In another study, Gil and Droit-Volet [8] argued that unpleasant mood states cause the time to pass more quickly. We hypothesize that in Information Retrieval (IR) conversational tasks, workers who are in an unpleasant mood state will perceive the response time lengthier than those who are in the pleasant state. We formed our hypothesis based on the prior works in microtask crowdsourcing, which argued that workers in pleasant mood perceived greater engagement [18], produced high quality results, and reported lower cognitive load [16]. This leads us to second research question (**RQ2**): *How does user's mood influence the perceived latency of CPCS?*

METHOD

We conducted a between-subjects experiment on the Prolific crowdsourcing platform to study the combined influence of delay and mood on the waiting experience of users while they interacted with the CPCS. We had four conditions based on the response delays: 2, 4, 8 and 16s. For each condition, we hired 60 unique workers. Our total sample includes 242 workers. We restricted the experiment to only US and UK workers. Each worker was paid £1.00 fixed amount (£8.57/h).

Fig. 2 represents steps involved in the procedure. **(1)** First of all, participants who accepted the task were asked to read and sign the consent form. This form explained that the chatbot is powered by hybrid intelligence explaining that answers are generated by a combination of computational and human input. Specifically, participants interacted with a simulated chatbot that replays human

Table 1: Examples of Search Tasks used in the study, adopted from Kelly et al. [13]

Domain	Task
Health	I recently watched a documentary about people living with HIV in the United States. I thought the disease was nearly eradicated, and am now curious to know more about the prevalence of the disease. Specifically, how many people in the US are currently living with HIV?
Science & Technology	I recently watched a show on the Discovery Channel, about fish that can live so deep in the ocean that they're in darkness most or all of the time. This made me more curious about the deepest point in the ocean. What is the name of the deepest point in the ocean?
Entertainment	I recently attended an outdoor music festival and heard a band called Wolf Parade. I really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album?

generated statements in a programmed sequence, while controlling the response rate. **(2)** Before proceeding further, we asked participants to express their current mood using the Pick-A-Mood scale [6]. Pick-A-Mood is a cartoon-based pictorial instrument for reporting moods. It measures eight distinct mood states, which can be divided into two groups: pleasant (excited, cheerful, relaxed, calm), and unpleasant (tense, irritated, bored, sad) with the addition of an optional neutral state. **(3)** After that, participants were redirected to the instructions page which contained explanations on how to interact with the bot, and which presented them with IR tasks in three domains (Health, Science & Technology, Entertainment).

(4) After indicating their mood, participants could interact with the conversational task. To imitate IR tasks with CPCS, we built the Talash bot (Fig. 1), which is simply a chatbot designed to answer IR queries in three domains – Health, Science & Technology and Entertainment. For the design of IR tasks, we relied on the framework developed by Kelly et al. [13], which is based on the principles of Bloom's taxonomy. They have provided a list of search tasks that can be reused by others (see sample tasks in Table 1). We varied the response time of the Talash bot based on the following geometric sequence: 2, 4, 8, and 16s. We adopted this sequence from prior research done by Butler [5] where he studied the relationship between computer response times (2, 4, 8, 16 and 32s) and user performance with simple data entry tasks. Nevertheless, we removed the excessive 32s delay level for the sake of current study. The Talash bot initiates the discussion by greeting the user. The user is then provided with some predefined domains in the form of quick replies (Fig. 1.A). After the user selects any domain of interest, Talash replaces quick replies with a predefined search task based on the dataset provided by Kelly et al. [13], and displays it in a chat bubble (Fig. 1.B). After the set time has elapsed, a predefined response is shown to the participants. Finally, users are shown quick replies (*"OK Thank You!"* or *"OK Perfect!"*) which they need to click to exit the IR task. We did not provide the option for users to type whatever they want or to ask follow-up questions because the priority for this study was to examine the combined influence of delays and mood on waiting experience. Thus, we were not interested in the variation of user input and response quality and wish to avoid confounding influences upon the perceived time delay.

(5) After the conversational task, participants were asked to fill out the exit survey. In the survey, *Perceived Waiting Time (PWT)* was assessed by asking the participants to give an estimate of total time (in seconds) they spent waiting between the user's response and the bot's response. We also measured the *cognitive* component of waiting time; it measures the perception of the time spent in terms of long or short judgement [15]. It is measured on a five-point scale (1: 'very short' , 5: 'very long'). We also asked participants to share any additional thoughts, remarks, or feedback that they may have regarding their experience interacting with Talash. In summary, we had two independent variables *mood* and *delay* and two dependent variables *PWT* and *cognitive*.

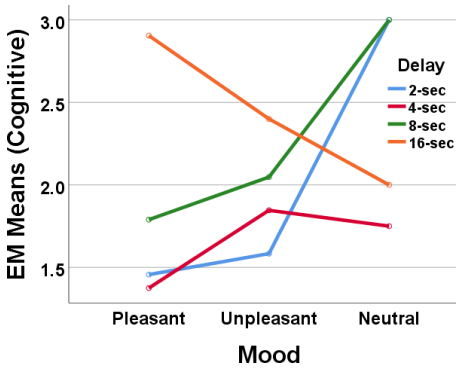

Figure 3: Interaction effects between *Mood* and *Delay*: Results indicate that participants who were in the *Neutral* group perceived the response delays longer than those who were in the *Pleasant* group except the 16-sec condition that showed inverse behaviour.

RESULTS

Since there were fewer participants in the unpleasant group (tense, irritated, bored, sad), we split the participants into three groups: pleasant, unpleasant, and neutral group. A two-way ANOVA was conducted that examined the effect of *mood* (3 levels) and *delay* (4 levels) on the *cognitive* variable. The assumption of homogeneity of variances was not violated ($p = 0.418$). We found a statistically significant interaction between the effects of Mood and Delay on cognitive, $F(6, 226) = 2.831, p = .011$ (Fig. 3). We further investigated the effect of *mood* on the mean cognitive scores at every level of delay through one-way ANOVAs. We only found a significant main effect of *mood* on the *cognitive* variable at the delay level of 2*s* ($F(2, 57) = 3.822, p = .028$) – Fig. 3. Pairwise comparisons with Bonferroni adjustment indicates that cognitive score was higher for the participants who had neutral mood ($M = 3.0$) than those who had pleasant mood ($M = 1.46$) – $p = .024$. Please note that a higher mean value of the cognitive scores indicates that participants perceived the response delays as longer. We expand on this further in the discussion. We did not find any difference in the mean cognitive scores at other levels of delays.

We also ran two-way ANOVA for the dependent *PWT*. We did not find interaction effect between *mood* and *delay* levels. However, the simple main effect of *delay* on mean *PWT* was statistically significant $F(3, 226) = 5.243, p < .002$. Pairwise comparisons indicates that the difference in *PWT* was significant between 2-sec ($M = 2.89$) and 16-sec ($M = 9.42$) – $p < .001$, 4-sec ($M = 2.44$) and 16-sec ($M = 9.42$)–$p < .001$, and 8-sec ($M = 5.78$) and 16-sec ($M = 9.42$)–$p < .001$– Fig. 4

DISCUSSION

From Fig. 3, it is evident that participants who reported being in a pleasant mood perceived the response delays smaller for 2s ($N = 60$), 4s ($N = 57$) and 8s ($N = 61$) delays. The difference in cognitive score was not significant among 2s ($M = 1.4$), 4s ($M = 1.3$) and 8s ($M = 1.7$) conditions for the pleasant group. This finding can be compared with an earlier study [7] that argues that while interacting with the conversational application, users can only tolerate a silence that does not exceed 8 seconds. On the other hand, participants who reported being in a pleasant mood perceived the response delays longer for the 16s ($M = 2.9, N = 60$) condition. Thus, we conclude that with the IR-based CPCS, users who are in pleasant state can tolerate the delays up to 8s and after this limit, their perception of response time will increase. This is also clear form the Fig. 4 where participants perceived 16s delay as 9.4s (> 8s). One possible interpretation of this finding is that longer response times can change the mood of the users from pleasant to unpleasant state due to stress and frustration [17]. Furthermore, it is not clear whether participants who reported 'neutral' mood were actually sad or calm because in a prior study [1], 5% of their participants failed to differentiate between the mood-states neutral, sad, and calm. We will investigate this further in future study. Thus, we conclude that for IR-based

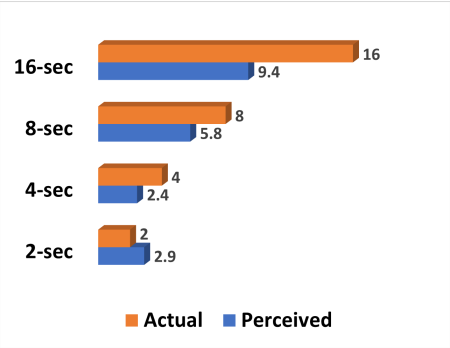

Figure 4: ANOVA indicates that there was significant difference between the PWT scores among different conditions. Also, participants perceived the PWT lower than the actual time

CPCS, response delays longer than 8s should be mitigated with time fillers, such as a holding message ("please bear with me while I think about that") or graphical typing indicators (three animated dots).

In summary, participants in our study perceived the response delays to be shorter than the system delays in reality (except when the delays were quite short; 2 seconds, cf. Fig. 4). Participants who indicated pleasant moods were only able to tolerate response delays up to 8s, after which their perception of waiting times reportedly increased.

The fact is that current CPCS are nowhere near the threshold of 8s; for instance, Chorus can only reduce delays up to 44.6 seconds by employing multiple workers, which may not always be affordable. Therefore, we argue that resolving response latency in CPCSs is a neglected area of research that deserves attention to effectively support a variety of applications.

POSITION STATEMENTS

We present an academic perspective about integrating CUIs and crowd computing to support wide variety of tasks, such as microtask crowdsourcing, crowd training and mental health. CPCS are currently more robust than AI to diverse domains and can competently hold a conversation with users in a more fluid, multi-turn conversation. Additionally, CPCS can be deployed quickly with no training cost and thereby can support a wide variety of application domains, such as tutoring, companionship, and search assistant, among others. On the downside, CPCSs currently suffer from long response delays, which hampers the feasibility of using CPCSs as conversational partners. From the user's point of view, we believe that different mood states can also shape user's perceptions about the perceived delays. This study is a part of a larger project that attempts to explore the connection between mood states and perceived latency for CPCS. This study can also initiate interesting discussion around various topics, such as the impact of response quality, task complexity, task type (e.g., mental health, education, information retrieval) and different waiting time fillers on the perception of delays.

REFERENCES

[1] Janna W Alberts, Martijn H Vastenburg, and Pieter MA Desmet. 2013. Mood expression by seniors in digital communication: evaluative comparison of four mood-reporting instruments with elderly users. (2013).

[2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[3] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in two seconds. *Proc. UIST'11* (2011), 33. https://doi.org/10.1145/2047196.2047201

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[5] Thomas W Butler. 1983. Computer response time and user performance.. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 58–62.

[6] Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. 2016. Mood measurement with Pick-A-Mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research* 14, 3 (2016), 241–279.

[7] Peter Fröhlich. 2005. Dealing with system response times in interactive speech applications. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1379–1382.

[8] Sandrine Gil and Sylvie Droit-Volet. 2009. Time perception, depression and sadness. *Behavioural processes* 80, 2 (2009), 169–176.

[9] Sandrine Gil and Sylvie Droit-Volet. 2012. Emotional time distortions: the fundamental role of arousal. *Cognition & emotion* 26, 5 (2012), 847–862.

[10] Ting-Hao Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[11] Ting-Hao Huang, Walter Lasecki, and Jeffrey Bigham. 2015. Guardian: A crowd-powered spoken dialog system for web apis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 3.

[12] Ting-Hao Kenneth Huang, Amos Azaria, and Jeffrey P Bigham. 2016. Instructablecrowd: Creating if-then rules via conversations with the crowd. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1555–1562.

[13] Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wan-ching Wu. 2015. Development and Evaluation of Search Tasks for IIR Experiments Using a Cognitive Complexity Framework. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval* (Northampton, Massachusetts, USA) *(ICTIR '15)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/2808194.2809465

[14] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 151–162.

[15] Ad Pruyn and Ale Smidts. 1998. Effects of waiting on the satisfaction with the service: Beyond objective time measures1Both authors contributed equally to this article.1. *International Journal of Research in Marketing* 15, 4 (1998), 321 – 334. https://doi.org/10.1016/S0167-8116(98)00008-1

[16] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Just the Right Mood for HIT!. In *International Conference on Web Engineering*. Springer, 381–396.

[17] Lawrence M Schleifer and Benjamin C Amick III. 1989. System response time and method of pay: Stress effects in computer-based tasks. *International Journal of Human-Computer Interaction* 1, 1 (1989), 23–39.

[18] Mengdie Zhuang and Ujwal Gadiraju. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*. 373–382.