Measuring Physiological Markers of Stress During Conversational Agent Interactions

Shreya Datar, Libby Ferland, Esther Foo, Michael Kotlyar, Brad Holschuh, Maria Gini, Martin Michalowski and Serguei Pakhomov

Abstract Conversational agent (CA) technology is rapidly becoming ubiquitous. Understanding how CAs impact users on multiple levels, including physiology, thus becomes increasingly important. In this study, we examined the effects of a CA interaction on naive users' physiological markers of stress i.e. heart rate (HR) and electrodermal activity (EDA). Participants (n = 21) prepared and executed a speech as part of a stressful interview, followed by a "Wizard-of-Oz" CA interaction. We expected the CA interactions to be mildly stressful. For a subset of participants with an initial resting period (n = 10), HR was elevated by 4.06 beats per minute (bpm) on average during the speech task, relative to the resting baseline. During the CA interaction however, HR was found to be 1.16 bpm lower on average. Moreover, HR and EDA values during the CA interaction were highly correlated with those during the resting period (Spearman's rho: HR = 0.97, EDA = 0.96) with small differences (mean diff: HR = 1.16, EDA = 1.82). Contrary to initial expectations, users do not seem to exhibit a stress response during the CA interaction. We additionally performed similar analyses and compared our results with the Wearable Stress and Affect Detection (WESAD) dataset [1].

1 Introduction

Conversational agents have amassed multitudes of uses and users [2–5]. However, because this technology is relatively new, its impact on users on multiple levels including physiology, has not yet been extensively investigated. Understanding the impact of CA technology on users is critical to its success and can help in guiding its design and the most effective and appropriate use.

Shreya Datar
University of Minnesota, Minneapolis, MN, USA e-mail: datar010@umn.edu

Serguei Pakhomov
University of Minnesota, Minneapolis, MN, USA e-mail: pakh0002@umn.edu

Various aspects of CA technology such as user experience [6], spoken dialogue system design [7] and modes of interaction [8,9] have been studied. Although physiological responses to technology have been extensively researched in the human-computer interaction literature [10], fewer studies have examined the physiological effects, particularly stress effects, of interacting with CAs. For instance, Lee et al. [11] used physiological signal data to explore users' experience with assistive CA technology while watching television. Prendinger et al. [12] proposed a method to measure responses to an affective and empathetic animated interface agent using skin conductance response. Similarly, Mori et al. [13] investigated the difference in the effect of an affective and non-affective embodied CA (ECA) on users and found that an ECA expressing empathy may offset the frustration or stress caused by shortcomings of the interface. A majority of these studies examined differences in physiological response for alternative versions of a CA. These studies demonstrate that various aspects of CA interactions can elicit physiological responses that can be analysed further to advance our understanding of user responses to CA technology.

Considering the pervasiveness of CA technology, studying the effect of CAs on users on a physiological level becomes all the more relevant. Speaking in front of even a very small audience is known to induce a physiological stress response [14]. If interacting with a CA is similarly stressful, it may result in limited use of such a system. Additionally, use of a CA that consistently elicits a physiological stress response could have negative health consequences. Frequent exposure to psychosocial stressors has been shown to negatively impact a variety of health measures including those indicative of poor sleep quality or those associated with progression of cardiovascular disease [15,16]. To our knowledge, only one prior study [12] specifically investigated users' stress response while interacting with a CA system. Better understanding of such stress responses could be valuable in guiding CA system design especially in populations of vulnerable users (e.g. cognitively impaired or older adults).

Advances in CA technologies provide a wealth of new opportunities for interaction, including applications in personalised care and targeted interventions. Since ELIZA [17], CA technology has improved dramatically and has increasingly been applied in health care to develop systems for clinical decision and triage support [18], screening and diagnosis [19], physical and mental health [20], and patient monitoring [21] to state a few, with interventions also targeted at older adults [22,23].

To contextualize the presented research, this study is part of a Grand Challenges project at the University of Minnesota that includes assistive technology interventions to advance the health and well-being of older adults. The project is aimed at integrating information relating to anticipated stressful everyday life events elicited via natural conversations with a CA in order to inform the interpretation of physiological signals collected using wearable sensors (e.g. heart rate, motion and electrodermal activity) for the purpose of activating just-in-time interventions designed to attenuate a user's stress response (e.g., upper-body compression via a specially designed garment [24,25]). The role of the CA system is to provide information supplemental to physiological signal data to identify time periods in which stressful events are more likely to occur, in order to minimize potential false positive activations of the

intervention. The ultimate overarching goal of this approach is to investigate the development of CA assisted personalized technology to deliver real-time interventions in response to stress detection in users, with a particular focus on vulnerable populations. This study focuses on just one of the multiple aspects of the larger project - whether the CA system itself induces stress.

The purpose of the current study was to examine and quantify the effects that a CA interaction has on naive users' physiological parameters associated with stress i.e. heart rate (HR) and electrodermal activity (EDA). In order to establish that participants in this study do in fact produce a physiological stress response that is measurable with the wearable sensor device we chose to use (Empatica E4), we used a standard mild stressor task widely used to study stress response - the speech portion of the Trier Social Stress Test [14]. At the outset, we expected to find that interacting with unfamiliar technology such as a CA system, would be at least somewhat stressful and result in a measurable physiological stress response comparable to one induced by a standard stressor task. Contrary to this expectation, we found preliminary evidence that while study participants had a measurable stress response to the stressful speech task, they did not show any stress response while interacting with the CA system. In fact, heart rate was lower on average during the CA system interaction as compared to baseline.

In the remainder of this paper, we describe the study design, data collection and analysis procedures, present results for two types of physiological measures of stress (HR and EDA) and discuss strengths and limitations of this study, as well as implications for future studies. We also present analyses and comparisons with the Wearable Stress and Affect Detection (WESAD) dataset [1], a publicly available dataset for wearable stress and affect detection.

2 Methods

Our analyses use data from three distinct sets of experiments. Data collected as part of the Grand Challenges project conducted at the University of Minnesota focused on measuring physiological responses as participants performed a series of tasks, which included the CA system interaction (CA Study). We also use another dataset from a substudy within the Grand Challenges project. As mentioned previously, a broader goal of the project is to incorporate personalized, real-time interventions through wearable technology. The substudy focused on designing and developing wearable haptic garments aimed at promoting relaxation. Experiential effects of various aspects of compression actuation with the garment were investigated and physiological signals were simultaneously recorded (Garment Study). Baseline HR and EDA data obtained in the Garment Study were used to externally validate, in an independent sample, the measurements obtained during the restful periods in the CA Study. The WESAD dataset [1] was used as an additional source of data (WESAD Study) to validate our study findings.

The Empatica E4 wristband was used to record physiological signals in all three studies mentioned above. It contains an optical HR sensor, a 3-D accelerometer, an EDA sensor, and a body temperature sensor. This device has been validated for HR and EDA measures, showing acceptable accuracy as compared to reference devices [26, 36].

2.1 Study Designs

2.1.1 CA Study

University of Minnesota students were recruited to be part of the initial participant cohort for this study. Each participant attended four laboratory sessions over a period of two days, with two sessions on each day. The first laboratory session consisted of four major tasks: resting period (for a subset of 10 participants), walking task, modified Trier Social Stress Test and the Wizard-of-Oz CA system interaction. This enabled a comparison of physiological responses during the CA interaction with a known standard stressor task. Hence for this analysis we used data collected during the first session only. The second, third and fourth lab sessions included only the WOZ CA interaction and data from these were not used in the current analysis. Participants also completed a post session survey at the end of each session.

The Wizard-of-Oz CA interaction was carried out through Amazon's Polly service. The voice chosen was a low to mid-range female voice as similar to the Alexa voice as possible (AWS Polly service's "Joanna" voice). There is some evidence that people react more positively to a female voice [40]. There are however other studies that have reported no significant differences in outcome based on gender characteristics [41]. The "system" was also highly consistent in interactions.

2.1.2 Garment Study

Physiological signal data was collected for participants as they completed a baseline survey and subsequently evaluated various facets of the wearable haptic garment.

2.1.3 WESAD Study

WESAD is a multimodal dataset consisting of physiological signal data recorded using a wrist-worn device (Empatica E4) and a chest-worn device (RespiBAN Professional) as part of a lab study. Physiological signal data were collected during three different affective states: neutral (during the resting period), stress (during the Trier Social Stress Test), amusement (where participants were shown eleven clips of funny videos). The stressful and amusement states were both immediately followed by a

meditation period. In addition to this, ground truth was obtained through subjective self-assessment questionnaires on five occasions during the lab session.

2.2 Participants

2.2.1 CA Study

Participants in this study were University of Minnesota students with minimal or no prior experience with a smart home system. There were 21 participants (17 females, 4 males). Three participants refused to disclose their age. All other participants were between 18 and 23 years of age (mean = 19.4, sd = 1.5). All participants completed a pre-screening survey in which information pertinent to studying physiological responses e.g. smoking behaviors, pregnancy and current medications, was noted. Previous experience with a smart home system was negative for 13 participants, positive for 4 (less than a year of use) and unknown for 4. The study was performed with IRB approval, and written informed consent from all participants. Participants received monetary compensation and additional course credits upon completion of the study.

2.2.2 Garment Study

This study involved 17 University of Minnesota students (9 females, 8 males) between 18-29 years old (mean = 22.1) who self-identified as healthy individuals with no existing cardiovascular/circulatory health concerns. Data had to be discarded for 4 participants because data collection was paused due to hardware issues with the garment.

2.2.3 WESAD Study

The WESAD study included 15 graduate students at the University of Siegen, Germany (3 females, 12 males) who were between 24 and 35 years of age (mean = 27.5, sd = 2.4). See [1] for further details.

2.3 Procedures

2.3.1 CA Study

Participants were fitted with the Empatica E4 at the beginning of the first experimental session on both days. The first session of the day was generally conducted in the

morning. Out of the four lab sessions for a participant, only the first session contained a standardized stressor task in addition to the CA interaction and was thus used for the current analysis. This session consisted of four tasks: a rest period, slow and fast walks (collectively, the walking task), a stressful speech task and a Wizard-of-Oz CA interaction. A session flow is illustrated in Fig. 1

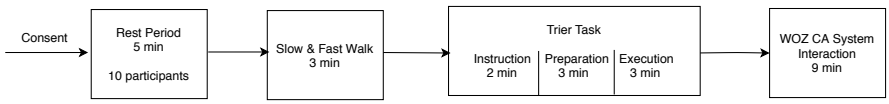

Fig. 1 Participant task flow for the first laboratory session during the CA Study.

Following informed consent, participants were fitted with the Empatica E4. For a subset of 10 participants, this was followed by a 5 min stationary rest period during which baseline measurements were recorded. All participants (n = 21) were then instructed to perform a slow and fast walking task. The Trier Social Stress Test task was administered following the walking tasks.

The Trier Social Stress Test is a widely used standardized stressor task [14]. The speech portion of the Trier test served as the stress-inducing component of the study. In this task, participants were asked to prepare for an interview (instruction phase) for a hypothetical job of their choice. They were given 3 min to prepare for the interview (preparation phase) and 3 min to execute the speech in front of 1 interviewer (execution phase). The interviewer prompted participants with other common interview questions if they were unable to speak for the entire 3 min duration.

The Wizard-of-Oz CA interaction took place immediately after the Trier task. The WOZ CA system asked participants questions to elicit information about their schedules and whether they anticipated any upcoming event to be stressful. The interactions lasted 9.2 min on average. The CA interaction was the last task of the laboratory session. Participants continued to wear the E4 device as they went about their day, until the end of the second session which was conducted towards the end of the day.

2.3.2 Garment Study

The wearable technology used in the Garment study involved the application of compressive and warmth sensations to a wearer while investigating how various haptic parameters such as location, duration, intensity, and pattern change a user's experience. The developed technology was a first step towards a long-term goal of employing haptic wearables to promote relaxation [24].

After the consent process, participants completed a paper survey in regards to their baseline clothing comfort and demographics. Physiological signal data corresponding to this stationary period was used as a resting baseline that could be used to

externally validate baseline measurements from the CA study. After completion of the baseline survey, participants evaluated the wearable haptic garment on varying forms of warm compression stimuli (data not presented here since it is not within the scope of this paper) [24].

2.3.3 WESAD Study

The WESAD study involved a 20 min stationary period where baselines were recorded, the Trier Social Stress Test, an amusement period where participants watched clips of funny videos, and 7 min guided meditation periods immediately after the stress and amusement states. In this study, both the speech and arithmetic portion of the Trier test were included, each lasting 5 min. Participants executed the Trier test in front of a three-person panel. The guided meditations were introduced specifically to de-excite participants and bring them back to close to a neutral affective state (cool-off period). See [1] for further details.

3 Analysis

3.1 Data Preprocessing

In the CA study, each lab session was audio-recorded. Event timestamps (i.e. the start time and end time for each experimental task) were obtained manually from the voice recording after it was aligned with the Empatica E4 signal streams as follows. The Empatica E4 records the Unix epoch timestamp of the time when the device was switched on, with the timing of each subsequent HR and EDA sample identified at a predefined frequency (HR: 1 Hz, EDA: 4 Hz) relative to the initial timestamp. The study facilitator was instructed to say out loud when they finished fitting the participants with the device. Using this event in the audio stream as an anchor point, the relative offsets in the E4 data streams were converted to absolute timestamps synchronized with the start and end of each task.

To examine measures of heart rate variability (HRV), an accepted indicator of stress [27], in the CA study, successive inter beat intervals (IBIs) were isolated from the IBI data stream recorded by the Empatica E4. For every participant, the total durations of IBI segments were calculated for each task.

In the Garment study, there was no categorically defined resting period; however, participants filled out a demographics and baseline comfort questionnaire. In order to obtain baseline measurements, we isolated HR and EDA values from stationary periods during the baseline survey period using E4's accelerometer data. This resulted in 1 min intervals of HR and EDA signal values on average per participant. Specifically, these periods were obtained by identifying a set of continuous G-values (sum of squared acceleration values along the three accelerometer dimensions) that had a standard deviation less than or equal to 5 along each dimension. Stationary pe-

riods could not be isolated for 3 participants. Baselines obtained for the 10 remaining participants were used in the analysis.

From the WESAD study, physiological signals collected using only the wrist-worn Empatica E4 were considered. The RespiBAN chest device records measurements at a frequency of 700Hz. Data synchronised between both devices were available in the dataset, however event labels were available only at the sampling frequency of the RespiBAN device. These were downsampled to get corresponding event labels for signals measured by the E4.

3.2 Statistical Analyses

To distinguish between participants associated with the different studies included in the analysis, we define the following:

- P1: set of all 21 participants in the CA Study
- P2: subset of 10 participants from the CA study with resting baseline measurements
- P3: subset of 10 participants in the Garment Study with resting baseline measurements
- P4: set of 15 participants from the WESAD study

A reference standard for physiological measures was not available for the CA study. Hence to externally validate the baseline measurements, HR and EDA rest period measures from the Garment Study (P3) and WESAD study (P4) were compared with those in the CA study (P2) using an unpaired t-test. Stationary period measurements from the Garment Study (P3) were also compared with the system interaction task of the CA Study (P2).

To compare the physiological response between tasks, we calculated the Spearman's correlation coefficient for both, HR and EDA for all pairs of tasks in the CA study (P2). For every participant, a mean HR value was calculated for each task by taking the average of all HR values recorded during the task. However for EDA, we considered the maximum value during the task. The EDA signal has two components: a tonic component (slower changing component) and a phasic component (faster changing component). The tonic component makes up the majority of the EDA signal. The EDA signal thus exhibits a lagged stress response. Hence, instead of calculating means as in the case of HR, maximum EDA values during tasks were used for analysis.

In the CA study, only the rest period (resting baseline or rest), Trier task preparation (preparation), Trier task execution (execution) and the WOZ CA system interaction (interaction) were considered for further analysis. Since the Trier instruction phase immediately followed the walking task, an elevation in HR during this task could either be attributed to stress or the preceding physical activity, and was thus left out. The walking task was not considered since it is not directly relevant for the purposes of the analysis. One-way ANOVA and pairwise t-tests were conducted to

analyze mean differences between the selected tasks. Pairwise t-tests were adjusted for multiple comparisons using the Bonferroni procedure. Similar analyses were performed using the WESAD dataset and the results were compared to those from the CA study. All statistical tests assumed the probability of Type 1 error <0.05 threshold for determining statistical significance. Statistical tests were carried out using the R statistical software package (Version:4.0.2; [28]).

4 Results

Fig. 2 shows raw HR and EDA values for a participant (Subject ID: 504) during the first laboratory session. Since EDA is sampled at a frequency of 4Hz, it was downsampled by taking rolling means, using a sliding window. The Trier instruction phase took place immediately after the walking task; and after Trier execution, participants were briefed on interacting with the CA system.

We observed individual variability in how people respond to stress. Mean HR response during Trier task execution was 6.32 beats per minute (bpm) with a standard deviation of 12.45 bpm. Mean EDA response was 0.32 microsiemens (μS) with a standard deviation of 0.67 μS.

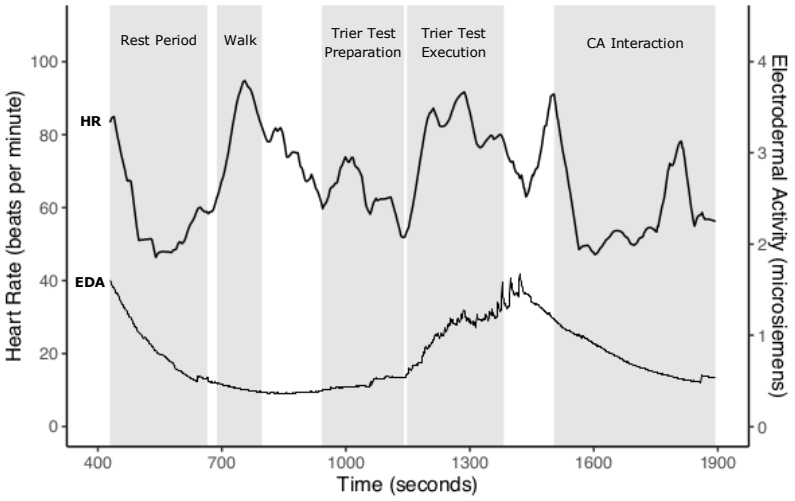

Fig. 2 Raw HR and EDA signals for participant 504 during the laboratory session.

4.1 External Validation of the CA Study Resting Period

No significant differences were found between the rest period of the CA study (P2) and the estimated stationary periods of the Garment study (P3) for both HR and EDA (HR: mean diff = 1.18 bpm, p = 0.84; EDA: mean diff = 1.53 μS, p = 0.19). Comparing HR values during the rest periods of the CA study (P2) and WESAD study (P4) resulted in a mean difference of 10.27 bpm (p = 0.09), however this difference was not statistically significant. No significant differences were found in EDA values between the two studies (mean diff = 0.01 μS, p = 0.99).

Additionally, no significant differences were found between the CA system interaction (P1) and the Garment study stationary periods (P3) (HR: mean diff = 0.12, p = 0.97; EDA: mean diff = 2.17 μS, p = 0.07).

4.2 Signal Comparisons Between Tasks

Fig. 3 shows correlations of mean HR and maximum EDA values between tasks in the CA study. There is a strong correlation between HR values during the rest period and CA system interaction (rho = 0.97, p<0.001) with small differences (mean diff = 1.16 bpm). Maximum EDA values during the rest period and system interaction were also strongly correlated (rho = 0.96, p<0.001) with small differences (mean diff = 1.82 μs). Overall, larger correlations between tasks were observed in the EDA signal compared to HR. Since the EDA signal exhibits a lagged response, tasks closer to one another temporally are expected to have large correlations. Relative to baseline, EDA was lower on average during the instruction phase (mean diff = -0.43 ms). This difference increased progressively during the Trier task (preparation mean diff = 0.31ms, execution mean diff = 1.87ms) but decreased during the system interaction (mean diff = 1.82ms).

Fig. 3 Correlations of mean HR values (Panel A) and maximum EDA values (Panel B) between tasks in the CA study.

We also compared the inter-task signal correlations of the CA study with the WESAD study for selected tasks. We define the following correlations, depicted in Fig. 4:

- C1: Correlation between Baseline and Trier Execution
- C2: Correlation between Trier Execution and CA Interaction (CA) / Meditation (WESAD)
- C3: Correlation between Baseline and CA Interaction (CA) / Meditation (WESAD)

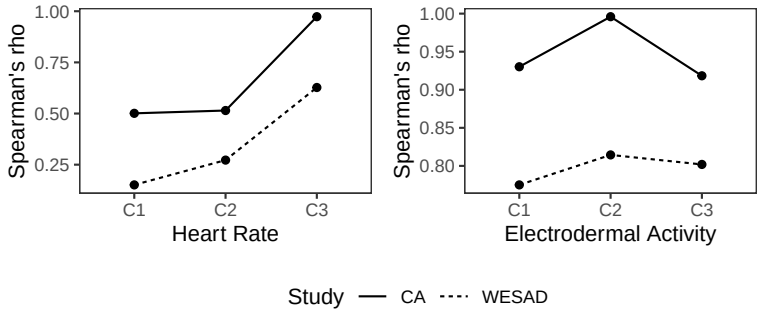

Fig. 4 Trends in correlations between tasks for the CA and WESAD studies.

In both studies, C3 for HR values is the largest correlation (CA study: rho = 0.97; WESAD study: rho = 0.63). The largest EDA correlation is C2. In the CA study, this could be explained by the temporal proximity of Trier execution to the CA interaction or meditation periods.

4.3 Comparing the Rest Period and CA Interaction

One-way ANOVA and pairwise t-tests were conducted to compare the differences in mean HR values between selected tasks during the CA study (P2). No significant differences were found between any pairs of tasks. On average, mean HR was elevated by 4.06 bpm (sd = 14.89 bpm) during the Trier task execution but was lower by 1.16 bpm (sd = 4.31 bpm) during the system interaction, relative to baseline (Fig. 5). With the same set of participants, one-way ANOVA and pairwise t-tests were also conducted using maximum EDA values during tasks however no significant differences were found between any pairs of tasks (Fig. 6). Log values were used for these calculations to meet the normality condition.

Since HR and EDA were similar in the rest period and system interaction (for P2), we used the CA interaction as a proxy for the baseline for all participants (P1), and repeated the above analysis. We found that mean HR was significantly elevated during Trier task execution (Fig. 7). EDA was significantly elevated during the Trier task execution as compared to both, baseline and Trier task preparation (Fig. 8).

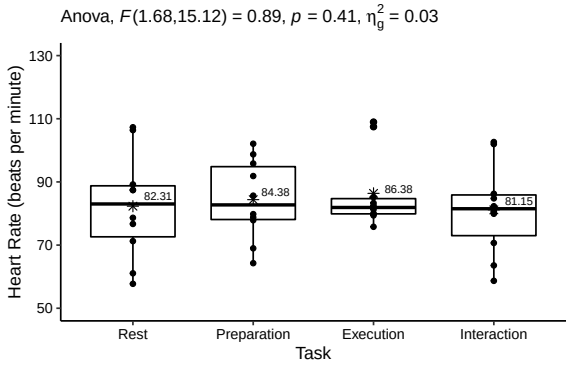

Fig. 5 Comparison of mean HR values between tasks in the CA study (P2).

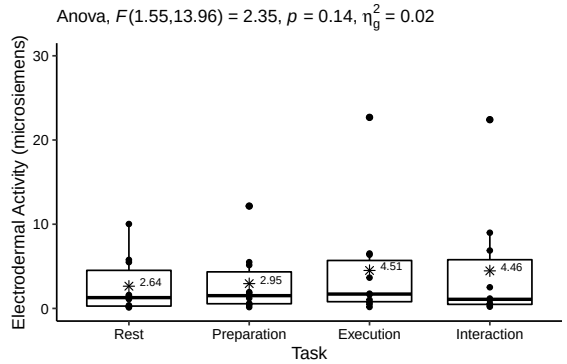

Fig. 6 Comparison of maximum EDA values between tasks in the CA study (P2).

4.4 Comparing the CA and WESAD studies

Using data from the WESAD study, one-way ANOVA between tasks for HR mean yielded significant differences between the baseline rest period and Trier test ($p < 0.001$) and between the Trier test and meditation period ($p < 0.001$). No significant differences were found between the baseline and meditation tasks. For EDA, significant differences were only found between the Trier test and meditation period ($p = p < 0.001$).

Fig. 9 and Fig. 10 show a comparison of the CA (P1) and WESAD (P4) studies considering their rest periods, Trier test and CA system interaction in case of the CA study, or meditation period in case of the WESAD study. The rest period is the baseline in both studies.

Increases and decreases in average HR means and maximum EDA values for the tasks follow the same pattern for both studies, although their magnitudes differ.

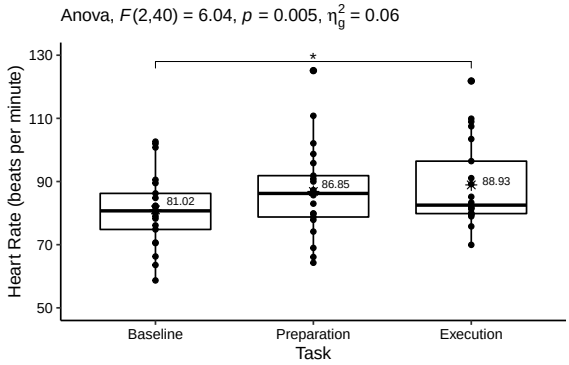

Fig. 7 Comparison of mean HR values between tasks in the CA study (P1) with CA system interaction as baseline. '*' indicates significance (p<0.05)

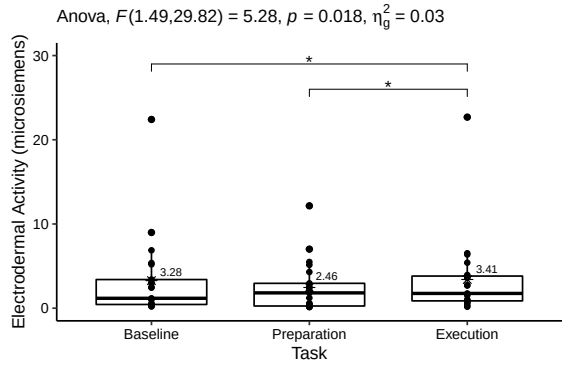

Fig. 8 Comparison of maximum EDA values between tasks in the CA study (P1) with CA system interaction as baseline.. '*' indicates significance (p<0.05)

4.5 HRV Analysis

We used the IBI signal recorded by the Empatica E4 to measure HRV in the CA study. A total of 822 IBI segments with average segment length of 7.1s were found across all participants (P1, n = 21) and tasks. The number of segments and average segment length for each task were as follows: rest 106 (27.6s), walking 41 (1.9s), Trier instruction 83 (7.3s), Trier preparation 133 (6.0s), Trier execution 67 (2.0s), CA system interaction 392 (8.5s). Note that rest period segments were only available for participant set P2 (n = 10). Since 3-5 min is the recommended length for measuring short-term HRV [29], we were unable to compute short-term HRV measures reliably. This is consistent with previous studies that used the Empatica E4 for HRV analysis [30]. Finally, we observed that a larger number of IBI segments were available in stationary, non-stressful periods such as the resting period and the CA system interaction.

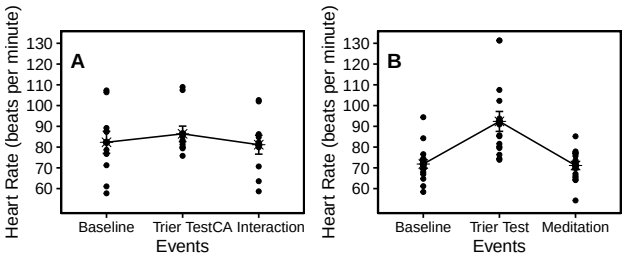

Fig. 9 Trends in mean HR during selected tasks for the CA (panel A) and WESAD studies (panel B).

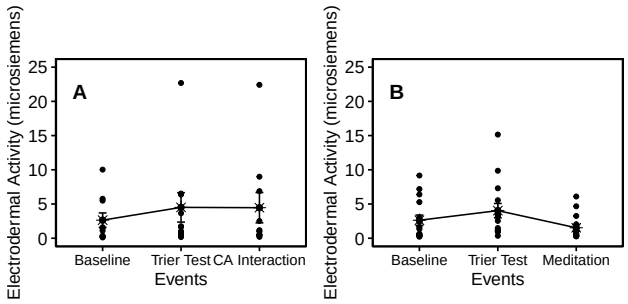

Fig. 10 Trends in maximum EDA during selected tasks for the CA (panel A) and WESAD studies (panel B).

5 Discussion

Prior literature shows that performing the Trier task results in a measurable increase in HR on the order of 5-10 bpm [31]. Our results are consistent with these prior findings and also show a similar elevation in HR during the Trier task. When the CA system interaction was used as baseline for all 21 participants, an elevation of 7.9 bpm was observed during Trier execution relative to baseline. This finding confirms that a) study participants were indeed stress-responders on average as they showed a response to a standard stressor, and b) the device used in the study was capable of measuring stress response when it was present. Contrary to our initial expectations, we found no elevation in HR during the CA interaction. This finding indicates that the CA system interaction may not be stress-inducing.

Similar trends were observed when selected tasks from the CA and WESAD studies were compared (Fig. 9, Fig. 10). The guided meditation period was specifically aimed at de-exciting participants after the stress and amusement affective states in the WESAD study. Relative to rest period baselines, mean HR and maximum EDA values were lower during the meditation task, as well as, during the CA interaction.

The consistently large correlations between tasks within the CA study, compared to the WESAD study (Fig. 4), could perhaps be attributed to the differences in duration and interval between tasks in the CA and WESAD studies. Tasks in the

WESAD study were longer and included a 5 min buffer between all tasks. This perhaps allowed participants to return to a rest-like phase before the beginning of the next task. The WESAD study incorporated a 20 min resting period where baseline measurements were taken, as compared to a 5 min resting period in the CA study. The WESAD study also included both the speech and arithmetic portions of the Trier Social Stress Test to be executed in front of a three-person panel whereas our study only used the speech portion, to be executed in front of one interviewer. This may explain the larger difference in mean HR between the resting period (baseline) and Trier test in the WESAD study, as shown in Fig. 9. Although we compared the CA and WESAD studies and found similar trends in physiological responses, a key difference between the two studies is the gender of participants. P1 consisted of 8 females and 2 males whereas P2 consisted of 3 females and 12 males. It is not clear what, if any, effect differences in gender composition of the two studies may have on the results. Some prior studies found that sex differences in the response to the Trier Social Stress Test were not significant [37, 38]; however, other studies found a significant effect [39]. Further investigation with larger samples is needed to answer this question.

Several limitations should be considered when interpreting the results of our study. The study relies on a small sample of participants. Our findings need to be replicated and confirmed in larger studies powered to determine that there truly is no difference between HR and EDA measures during resting or relaxation periods and the CA system interaction. Power analyses using recorded HR and EDA values indicate that task comparisons in the CA study would be significant at the 95% confidence level if the number of participants is at least 85 considering HR values, or 28 considering EDA values only. The participants in this study were students, thus limiting the generalizability of our findings to older individuals. In this study, we were only able to assess physiological parameters during interactions with a single WOZ task-based system-initiative CA. Further investigation is necessary to evaluate stress response to interactions with other CA system types designed for various purposes (e.g., CA systems designed as therapeutic agents for healthcare applications). Our study also did not include a cool-off period as in other Trier test studies. A comparison between such a phase and the CA interaction would further strengthen the argument that a CA interaction is in fact not stress inducing. Additionally, while we were not able to independently verify the accuracy of the Empatica E4 measurements against a reference device in this study, this device has been previously investigated in other studies and shown to produce reliable estimates of HR and EDA [26, 30, 36].

Our study is one of the first to explore approaches to objectively quantify effects of CA technology on users' physiological parameters in real time. Additionally, we were able to replicate findings of other laboratory [31] and naturalistic studies [32] showing that a wrist-worn optical HR sensor can detect elevation in HR in response to a standard mild stressor. Furthermore, the results of the current study also suggest that similarly to HR measures, EDA measures also respond to a stressful stimulus and can be measured with a wrist-worn sensor, which is consistent with prior studies showing that, despite lower correlation with galvanic skin response measures obtained from fingers, wrist-based EDA measures offered better discriminative power for stress

detection [33]. Finally, we also showed that consistent with prior studies [30], the Empatica E4 cannot be used to reliably measure HRV.

The pace of advancement in CA technology in the last decade has been very rapid. Reliability and adoption of technology however is a gradual process, and this becomes a particularly important consideration while designing systems for older adults. Yaghoubzadeh et al. [34] studied qualitatively, the acceptance of a virtual daily assistant by elderly or cognitively impaired users and the feasibility of a successful interaction in a Wizard-of-Oz setting. They found that although elderly cognitively impaired users were more reluctant to accept and recognise use of a daily assistant, focus groups, interviews and encounters with an actively engaging system prototype helped to advance acceptance. The CA system used in this study was simple and predictable, which meant that participants acclimatized to it fairly quickly. In fact participants, in post-study surveys, often rated the system as robotic, friendly, and/or polite. Directly examining older adults' interactions with the CA would provide valuable insights into their acceptance of this technology, in addition to inspecting the feasibility of real-time monitoring of physiological signals. The results presented here are preliminary, however they have important implications for integrating CA systems into a patient's care process.

References

1. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 400-408. (2018)
2. Porcheron, M., Fischer, J. E., Reeves, S., Sharples, S.: Voice interfaces in everyday life. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-12. (2018)
3. Car, L. T., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, YL., Atun, R.: Conversational agents in health care: Scoping review and conceptual analysis. J. Med. Internet Res. **22**, e17158 (2020)
4. Hobert, S., Meyer von Wolff, R.: Say hello to your new automated tutor–a structured literature review on pedagogical conversational agents. In: 14th International Conference on Wirtschaftsinformatik. Siegen (2019)
5. Mozer, T: Speech's Evolving Role in Consumer Electronics... From Toys to Mobile. In: Mobile Speech and Advanced Natural Language Solutions, p. 23-34., Springer (2013)
6. Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., Bandeira, N.: "What can I help you with?" Infrequent users' experiences of intelligent personal assistants. In: Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1-12. (2017)
7. McTear, M. F.: Spoken dialogue technology: enabling the conversational user interface. Assoc. Comput. Mach. Comput. Surv. (ACM-CSUR). **34**, 90–169 (2002)
8. Schroeder, J., Schroeder, M.: Trusting in machines: how mode of interaction affects willingness to share personal information with machines. In: Proceedings of the 51st Hawaii International Conference on System Sciences. (2018)
9. Ciechanowski, L., Przegalinska, A., Magnuski, M., Gloor, P.: In the shades of the uncanny valley: An experimental study of human–chatbot interaction. Future Gener. Comput. Syst. **92**, 539–548 (2019)

10. Barreto, A., Zhai, J., Adjouadi, M.: Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In: International Workshop on Human-Computer Interaction, pp. 29-38. Springer (2007)
11. Lee, S., Ryu, H., Park, B., Yun, M. H.: Using Physiological Recordings for Studying User Experience: Case of Conversational Agent-Equipped TV. Int. J. Hum. Comput. Interact. **36**, 815–827 (2020)
12. Prendinger, H., Becker, C., Ishizuka, M.: A STUDY IN USERS'PHYSIOLOGICAL RESPONSE TO AN EMPATHIC INTERFACE AGENT. Int. J. Hum. Robot. **3**, 371–391 (2006)
13. Mori, J., Prendinger, H., Ishizuka, M.: Evaluation of an embodied conversational agent with affective behavior. In: Proceedings of the AAMAS03 Workshop on Embodied Conversational Characters as Individuals. (2003)
14. Kirschbaum, C., Pirke, KM., Hellhammer, D. H.: The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiol. **28**, 76–81 (1993)
15. Han, K. S., Kim, L., Shim, I.: Stress and sleep disorder. Exp. Neurobiol. **21**, 141–150 (2012)
16. Kivim¨aki, M., Steptoe, A.: Effects of stress on the development and progression of cardiovascular disease. Natl. Rev. Cardiol. **15**, 215–229 (2018)
17. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. Assoc. Comput. Mach. (ACM) **9**, 36–45 (1966)
18. Sp¨anig, S., Emberger-Klein, A., Sowa, JP., Canbay, A., Menrad, K., Heider, D.: The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artif. Intell. Med. **100**, 101706 (2019)
19. Ghosh, S., Bhatia, S., Bhatia, A.: Quro: Facilitating user symptom check using a personalised chatbot-oriented dialogue system. Stud. Health Technol. Inform. **252**, 51–56 (2018)
20. Fitzpatrick, K. K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. J. Med. Internet Res. Ment. Health **4**, e29 (2017)
21. Galescu, L., Allen, J., Ferguson, G., Quinn, J., Swift, M.: Speech recognition in a dialog system for patient health monitoring. In: 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop, pp. 302-307. IEEE (2009)
22. Fadhil, A.: Beyond patient monitoring: Conversational agents role in telemedicine & healthcare support for home-living elderly individuals. arXiv preprint arXiv:1803.06000 (2018)
23. Ferland, L., Li, Z., Sukhani, S., Zheng, J., Zhao, L., Gini, M. L.: Assistive AI for Coping with Memory Loss. In: AAAI Workshops, pp. 431-434. (2018)
24. Foo, E. W., Lee, J. W., Compton, C., Ozbek, S., Holschuh, B.: User experiences of garment-based dynamic compression for novel haptic applications. In: Proceedings of the 23rd International Symposium on Wearable Computers, pp. 54-59. (2019)
25. Foo, E., Baker, J., Compton, C., Holschuh, B.: Soft Robotic Compression Garment to Assist Novice Meditators. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-8. (2020)
26. Bent, B., Goldstein, B. A., Kibbe, W. A., Dunn, J. P.: Investigating sources of inaccuracy in wearable optical heart rate sensors. Nat. Partn. J. Digit. Med. **3**, 1–9 (2020)
27. Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., Wager, T. D.: A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci. Biobehav. Rev. **36**, 747–756 (2012)
28. R Core Team.: R: A language and environment for statistical computing. R Foundation for Statistical Computing (2020)
29. Richardson, P., McKenna, W., Bristow, M., Maisch, B., Mautner, B., O'Connell, J., Olsen, E., Thiene, G., Goodwin, J., Gyarfas, I., Martin, I., Nordet, P.: Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the Definition and Classification of cardiomyopathies (1996) doi: 10.1161/01.cir.93.5.841
30. Barrios, L., Oldrati, P., Santini, S., Lutterotti, A.: Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, pp. 251-261. (2019)

31. Kotlyar, M., Brauer, L. H., al'Absi, M., Adson, D. E., Robiner, W., Thuras, P., Harris, J., Finocchi, M. E., Bronars, C. A., Candell, S., Hatsukami, D. K.: Effect of bupropion on physiological measures of stress in smokers during nicotine withdrawal. Pharmacol. Biochem. Behav. **83**, 370–379 (2006)
32. Pakhomov, S. V. S., Thuras, P. D., Finzel, R., Eppel, J., Kotlyar, M.: Using consumer-wearable technology for remote assessment of physiological response to stress in the naturalistic environment. Public Libr. Sci. One **15**, e0229942 (2020)
33. Ollander, S., Godin, C., Campagne, A., Charbonnier, S.: A comparison of wearable and stationary sensors for stress detection. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 004362–004366. (2016)
34. Yaghoubzadeh, R., Kramer, M., Pitsch, K., Kopp, S.: Virtual agents as daily assistants for elderly or cognitively impaired people. In: International workshop on intelligent virtual agents, pp. 79-91. (2013)
35. Iovine, A.: Conversational Agents for Recommender Systems. In: Fourteenth ACM Conference on Recommender Systems, p. 758-763. (2020)
36. van Lier, H. G., Pieterse, M. E., Garde, A., Postel, M. G., de Haan, H. A., Vollenbroek-Hutten, M. MR., Schraagen, J. M., Noordzij, M. L.: A standardized validity assessment protocol for physiological signals from wearable technology: Methodological underpinnings and an application to the E4 biosensor. Behav. Res. Methods, 1–23 (2019)
37. Kelly, M. M., Tyrka, A. R., Anderson, G. M., Price, L. H., Carpenter, L. L.: Sex differences in emotional and physiological responses to the Trier Social Stress Test. J. Behav. Ther. Exp. Psychiatr. **39**, 87–98 (2008)
38. Sgoifo, A., Braglia, F., Costoli, T., Musso, E., Meerlo, P., Ceresini, G., Troisi, A.: Cardiac autonomic reactivity and salivary cortisol in men and women exposed to social stressors: relationship with individual ethological profile. Neurosci. & Biobehav. Rev. **27**, 179–188 (2003)
39. Kudielka, B. M. and Buske-Kirschbaum, A. and Hellhammer, D. H. and Kirschbaum, C.: Differential heart rate reactivity and recovery after psychosocial stress (TSST) in healthy children, younger adults, and elderly adults: the impact of age and gender. Int. J. Behav. Med. **11**, 116–121 (2004)
40. Mitchell, W. J., Ho, C. C., Patel, H., MacDorman, K. F.: Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. Comput. Hum. Behav. **27**, 402–412 (2011)
41. Habler, F., Schwind, V., Henze, N.: Effects of Smart Virtual Assistants' Gender and Language. In: Proceedings of Mensch und Computer 2019, pp. 469-473. (2019)