Ponzoni Carvalho Chanel, Caroline and Wilson, Matthew D. and Scannella, Sébastien Online ECG-based Features for Cognitive Load Assessment. (In Press: 2019) In: IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), 6 October 2019 - 9 October 2019 (Bari, Italy).

Online ECG-based Features for Cognitive Load Assessment

Caroline P. C. Chanel
ISAE-SUPAERO,
Université de Toulouse, France
caroline.chanel@isae-supaero.fr

Matthew D. Wilson
ISAE-SUPAERO,
Université de Toulouse, France
matthew.wilson.engineer@gmail.com

Sébastien Scannella
ISAE-SUPAERO,
Université de Toulouse, France
sebastien.scannella@isae-supaero.fr

***Abstract*—This study was concerned with the development and testing of online cognitive-load monitoring methods by means of a working-memory experiment using electrocardiogram (ECG) analyses for future applications in mixed-initiative human-machine interaction (HMI). To this end, we first identified potentially reliable cognitive-workload-related cardiac metrics and algorithms for online processing. We then compared our online results to those conventionally obtained with state-of-the-art offline methods. Finally, we evaluated the possibility of classifying low *versus* high working-memory load using different classification algorithms. Our results show that both offline and online methods reliably estimate the workload associated with a multi-level working-memory task at the group level, whether it is with the heart rhythm or the heart rate variation (standard deviation of the RR interval). Moreover, we found significant working-memory load classification accuracy using both two-dimensional linear discriminant analyses (LDA) or a support vector machine (SVM). We hence argue that our online algorithm is reliable enough to provide online electrocardiographic metrics as a tool for real-life workload evaluation and can be a valuable feature for mixed-initiative systems.**

1. Introduction

Mixed-initiative systems are those wherein humans and automated systems share decision-making authority [1], [2]. This is a rapidly growing field in today's world, with everything from airliners to kitchen appliances being partially automated but relying, at some level, on high-level commands from humans. One can easily envision a situation, however, where the human operator makes the wrong decision due to poor user interface design, the complexity of the automation, or high operational pressure, in particular when mental workload exceeds human mental resources [3]. One potential line of improvement for such human-machine interaction (HMI) could be to carry out online monitoring of operators' cognitive state (mental resource) to decide when automated systems should or should not have authority. For instance, in the work of Régis and colleagues [4], ocular activity and Heart Rate were successfully used *attentionnal-tunneling* to determine a degraded cognitive state, corresponding to overly-long fixations on visual interfaces or objects to the exclusion of other important cues or alarms. More recently, Scannella et al. [5] successfully used the same metrics to evaluate pilots' flying-related mental workload in real, light aircraft. Building on the work of Régis and colleagues [4], de Souza et al. [6] implemented a successful method to develop control policies that could take such human state estimations into account to improve a mixed-initiative mission's success rate. Success was achieved despite various system failures (human or machine), providing that both human and machine state estimations could be reliably inferred *online*.

However, estimation of a human operator's cognitive state is a hard and challenging task [7] and is currently a hot research topic in neuroergonomics and human-factors communities [8], [9]. The greatest challenge is to find psychophysiological markers that are *relevant* given the mental state being considered in a biocybernetic system, such as the mixed-initiative HMI framework.

A recent survey has indicated promising workload algorithms to access cognitive load online [10]. In particular, metrics computed based on ocular and cardiac activities were demonstrated to be particularly useful. Ocular activity features, such as the number and duration of fixations and blinks, are already rendered online by acquisition systems. However, there is a lack of online metrics, such as ECG-based features, that are less dependent on the visual interface. In light of this, the present work reports details on the development and testing of algorithms and tools to compute ECG features online, potentially useful for mixed-initiative HMI systems.

It is known that the electrocardiographic signal is decomposed into successive positive and negative peaks forming a complex know as the "QRS complex". Heart Rate (HR) is often defined in terms of intervals between two successive periods rather than the rate itself, known technically as the Inter-Beat Interval (IBI). For IBI analysis, among all features that can be used, the R-peak is the easiest to extract and identify (see Fig. 1). Hence, "RR interval" is used interchangeably with "IBI".

Online availability of the Heart Rate (HR) is not new [11]. However, it was suggested that additional IBI-dependent metrics, such as the standard deviation (SD) – also called temporal heart-rate variability (HRV) [10] – or the frequency-band power of the IBI series [12], [13], could

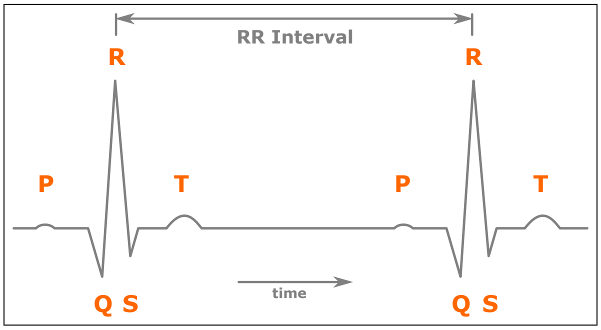

Figure 1. A simplified illustration of two heart beats on an ECG plot, showing QRS complexes and the RR interval.

be more reliable to infer cognitive states than the basic HR itself [3]. Andreoli et al. (2010) [14] have proposed the SPINE-HRV toolkit, an open-source framework supporting rapid prototyping of signal processing applications. In particular, they evaluate ECG-based time-domain metrics to detect participants' stress during everyday life activities. Yet, even though the inclusion of such metrics was proposed [10] to infer cognitive states, no existing ECG-based online tool for *cognitive workload evaluation* has been developed yet and experimentally evaluated, to the authors' best knowledge [12], [15], [16].

Particular attention is given here to online assessment of Heart Rate Variability (HRV), which has been shown to be correlated with cognitive workload [3], [10], [16], [17], [18]. More precisely, this metric has been shown to correlate inversely with cognitive workload, although that correlation diminishes with subject fatigue or "time on task" [16]. The temporal HRV metric can be defined as the standard deviation of the IBI series:

$$\sigma_{IBI} = \sqrt{\overline{\left(IBI - \overline{IBI}\right)^2}}. \tag{1}$$

In this context, the present work aims to develop and test a unified system to provide cardiac metrics (e.g. HR and HRV) online. To reach this goal, an experimental setting was designed in which participants' cognitive load was solicited by means of a working-memory span task. More specifically, online HR- and HRV-related metrics were assessed in a multilevel letter span task [19]. Such experimental design interleaves randomized-difficulty trials with blocks of trials of a given difficulty. This evaluated two things: sensitivity across time and, in particular, the relevance of each of these methods online cardiac-feature processing when estimating the cognitive load of human operators.

According to previous studies [20], [21] we expected a decrease of both offline mean RR interval (i.e. an increase of HR) and offline mean RR-SD (i.e. a decrease of the HRV) as a function of the working memory difficulty. Validating this hypothesis would provide reliable control metrics to compare with our online ones. Consequently we expected that our online metrics would reveal equivalent results to those obtained offline. Finally, we expected to be able to obtain significant working memory load classification using our online metrics.

2. Materials and Methods

2.1. Span task experiment

A working memory (WM) span task is a popular and well-documented type of cognitive psychology test [19] and has already been used for similar heart-rhythm-cognitive-load tests [16], producing noticeable effects in cardiac activity, albeit in offline block analysis. It involves presenting participants with a sequence of digits or letters that they are subsequently required to recall in the presented order.

One of the easiest ways to assess workload memory capacity (WMC) is a computer-based *counting span task* where the participant is visually presented with sequences of letters, one after the other, to be recalled on a keypad thereafter. Unlike typical WM span tasks, the goal in *this* research was not to evaluate participants' WMC *per se* but to induce varying levels of cognitive load to be assessed. Thus, *trials* could afford to vary in length and associated difficulty (a long trial is more difficulty to memorize) during the task.

The objective of this research was to find indicators that could at least distinguish zero load from full load. Hence, it was necessary to ensure that the hardest trials (the longest sequences) would fully load the participants – i.e. would reach the participant's WMC or come very close to it. In this sense, the task had to provide indications of the abilities of the various online metrics to identify changes in cognitive load (i.e. the measurement's difficulty resolution) and the time required for each of those metrics to do so (i.e. the measurement's time resolution).

2.2. Experimental Protocol

Participants were asked to reproduce sequences of alphabet characters (a span task). To avoid having phonically-memorable syllables or even whole words appearing in the sequence, the Latin alphabet was used minus vowels and 'y'. To indicate the metrics' abilities to differentiate task-difficulty, three levels were used: (i) low (0 letters ; called control); (ii) medium (3 letters); and (iii) high (7 letters). All trials presented the same number of characters and required the same number of key-presses in response (that is 7, as for the longest sequence). For 0- and 3-letter trials, '#' characters were used to fill the remaining spaces, with the participants pressing the space bar on the keyboard to enter them. That is, a 0-letter trial flashed 7 '#' characters and then the participant pressed the space bar 7 times. For a 3-letter trial, the 3 letters were flashed on the screen followed by 4 '#' characters, and the participant responded by typing the three letters and then pressing the space bar 4 times (see Fig. 2.b).

In order to test the time resolution of our online metrics, we also used two types of trials:

- Block : Consisting of a long time window (72 s) within which a block of 5 same-difficulty trials were presented. This was repeated for half of the task duration;

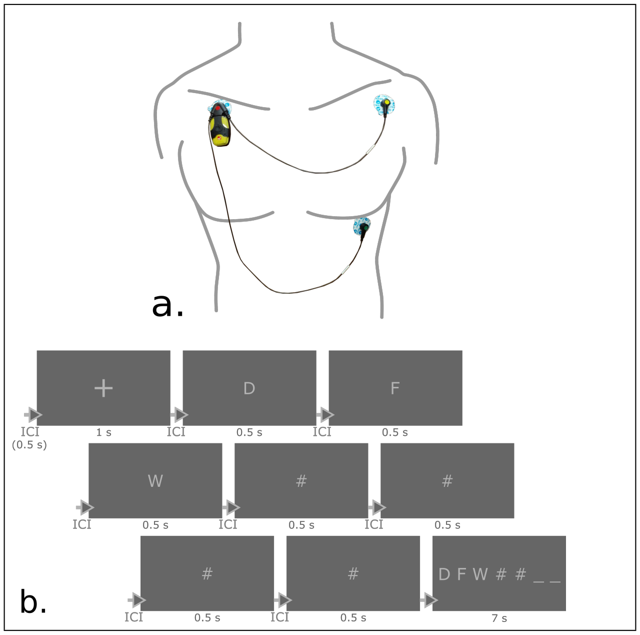

Figure 2. a. ECG device setup (Faros eMotion 360°). b. A representative sequence of a letter span trial.

- Event-related: Consisting of short time windows (14 s) within which a single trial of randomized difficulty was presented. This was repeated over the other half of the task duration.

Long- and short-sequence order was counterbalanced across participants.

The complete experiment lasted approximately 25 minutes, and can be summarised as follows:

- 150s of complete rest;
- Instructions presented, in simple English, on the computer screen;
- A training phase where the participant was given up to 3 attempts at each difficulty level to get comfortable with the concept of letters and particularly with the '#' characters;
- Long and short time-window sequences:
  - Blocks (long time-window) of 5 trials each for the three difficulty levels (i.e. 15 trials), repeated two times for a total of 30 trials;
  - Random-difficulty (short time-window) individual trials, also with 10 trials per difficulty, giving a total of 30 trials, with 5s pauses every 5 trials as with the blocks;
- 60 s of complete rest;
- Mental arithmetic problem;
- 30 s of complete rest.

The reference rest period (150 s) was introduced at the very beginning of the task with the aim to provide a reference level, for each measurement, for minimal physical and mental effort. In this rest period the participants were required to remain as still as possible, and were asked to think of as little as possible. Participants were allowed to close their eyes for this period; a bell sound indicated the end of the rest period. After the training phase of the task, the timing for each trial and block was automated and fixed to ensure consistency. The two halves contained the same total number of trials, and extra pauses were introduced between each 5 of the 'random' trials so as to match the timing of the block trials. Consequently, overall, each half had nearly identical duration and cadence.

Having noticed with some particular participants during the pilot tests that some measurements were showing minimal change over the difficulty levels, it was decided to add a "high contrast" phase at the end of the task, in case the span trial difficulty levels did not result in any demonstrable effect. This end phase consisted of a complete rest period of 60 s, under the same instructions as for the reference rest at the beginning, followed by a short challenging mental arithmetic problem participants had to solve, and ending with 30 s of complete rest.

2.2.1. Participants. The experiment was conducted among academics. 11 participants, all volunteers, completed the task (7 male, 4 female, mean age 24, s.d. 2.16). This research meets all American Psychological Association ethical standards [22]. Considering the participants, the research has been performed in accordance with the French Law on Bioethics that meets all the ethical standards laid down in the 1964 World Medical Association Declaration of Helsinki [23]. All volunteers gave their informed consent.

2.2.2. Task implementation and ECG data acquisition. In order to run the span task in a controlled manner, the experiment was automated using the Simulation and Neuroscience Application Platform (SNAP)[1] – a Python-based general purpose neuroscience testing framework designed to facilitate neuroscience-related studies with human participants. SNAP has built-in *Lab Streaming Layer* (LSL)[2] [24] support wherein it sends automatic and user-defined event markers (or triggers) to an LSL stream as either integers or strings. The ECG acquisition device (Faros eMotion 360°, see Fig. 2.a) can be configured to provide ECG raw data and online RR intervals (based on R-peak detection), which are broadcast using a Bluetooth protocol. The Faros Streamer application[3] then allows streaming of these data through LSL. Additionally, a script was run to compute online RR_SD metrics (see Sec. 2.3) and to stream them to LSL. The SNAP markers, Faros raw data and RR intervals, and online computed RR_SD streams, were all recorded with LabRecorder (the LSL recording tool), the which makes use of the LSL clock for data synchronization.

2.3. Data processing

There were three main data analysis steps in this study:

1) Offline-metric validation: Whether the designed experiment had actually given statistically significant *Difficulty* effects over the offline metrics.

[1] https://github.com/sccn/SNAP
[2] https://github.com/sccn/labstreaminglayer
[3] https://github.com/bwrc/faros-streamer-2

2) Online-metric identification: From the results, and in light of the end goal, this was the identification of promising online metrics for cognitive load estimation;
3) Working memory load classification: based on selected online metrics.

For the first step, ECG metrics were initially computed for each period of interest using Kubios[4], state-of-the-art software for *offline* ECG data analysis using the ECG raw data acquired. After automated R-peak detection, Kubios software provided offline RR mean and RR standard deviation that was computed with Eq. 1, and is here called "offline RR_SD", for each period of interest.

In parallel, for the second step, the online RR intervals provided by the ECG device were exploited. The mean online RR and online RR standard deviation (RR_SD) metrics were also computed using our own Python-based code. In contrast to the conventional online RR_SD computed for a sliding window, in this work, we propose another type of online RR_SD computation based on a continuous averaged rolling standard deviation, which is achieved by an Exponentially Weighted Moving Average (EWMA) filter [25]. EWMA is a filtering tool particularly useful for detecting small shifts in a discrete time process. The EWMA is defined as:

$$y_i = \frac{x_i + N \cdot y_{i-1}}{1+N} \tag{2}$$

where, i is the sample index, N is a positive weighting constant used to balance previous and new values, x_i the new standard deviation value of the last 12 RR intervals (approximately 10 seconds), and y_{i-1} the last EWMA value. If $N = 1$, a simple average between the current and previous values is computed. Larger values of N give a greater weighting to the previous values, resulting in a smoother result. If $N = 0$, no filtering occurs. It is important to say that, each time a new RR interval was received, a new value of online RR_SD was computed using the EWMA filter. The online RR_SD processing script for this experiment provided a three-channel output stream (RR_SD-7, RR_SD-15, and RR_SD-20) with three empirically defined N-values (7, 15 and 20, respectively) for comparison of various levels of filtering.

In addition to the direct sample values of each previous metric, a number of different measurements were considered for every metric over each window of interest: mean, standard deviation, end value (in the window), maximum value, minimum value, range, and slope (mean gradient). These measurements are illustrated in Figure 3.

2.4. Statistical Analysis

Accuracy. A repeated two-way analysis-of-variance (ANOVA) measurement was carried out to evaluate the *type* (Block *vs.* Random) and *difficulty* (0 *vs.* 3 *vs.* 7 letters) effects for trials over the participants' span-response accuracy.

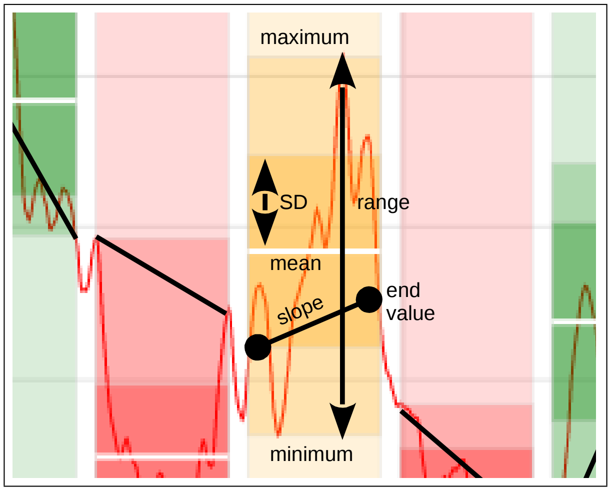

Figure 3. A portion of the RR interval signal illustrating the extra online measurements, annotated for one block. The orange line corresponds to the RR interval as a function of the time. Colored vertical blocks correspond to trials.

RR stream. In order to validate the online RR intervals provided by the Faros device, a Z-transform Lin's concordance correlation coefficient [26] was calculated between the RR data series from the Faros and those obtained offline with the Kubios software, across all subjects.

Offline ECG metrics. To evaluate the propensity of our letter-span task to induce effects over the cardiac metrics, two-way analysis-of-variance tests (ANOVAs) were conducted for the offline metrics (i.e. Mean RR interval and Mean RR_SD) with the same two factors used for the accuracy analysis (*Type* and *Difficulty*). For both offline ECG metrics and accuracy ANOVAs, p values were adjusted, in case of violation of sphericity, using the Greenhouse Geisser correction. Post-hoc pairwise t-tests with false discovery rate (FDR) correction were used to identify the nature of any identified effects.

Online ECG metrics. As we were interested in finding online metrics that could distinguish between working memory difficulty levels, we then conducted Welch 2-sample, 2-tailed t-tests between the three difficulty levels, both for the block sections and the randomized sections, over all online metrics (i.e., minimum, maximum, mean, slope and end value of RR, RR_SD-7, RR_SD-15 and RR_SD-20).

Working-memory load classification. Finally, to evaluate the possibility of classifying working-memory load, the most promising metrics from the online metric group-level results were selected to build two-dimensional classifiers. Among all existing classifiers we chose to evaluate those most commonly used in the literature – that is a Support Vector Machine (SVM), Linear and Quadratic Discriminant Analyses (LDA, QDA), k-Nearest Neighbors (kNN), and Naive Bayes (NB) using the scikit-learn library version 0.20.3. For SVM, KNN and NB classifiers, a grid search algorithm was applied for parameter-optimisation. Each chosen classifier was cross-validated (cv=5). Each training-test episode used 80% of the data for training

[4] http://www.kubios.com/

and kept the rest for testing purposes. Then the average precision, recall and f1-scores were computed.

3. Results

For the sake of clarity, only significant results are reported in the following section.

3.1. Letter-span accuracy

The ANOVA over the task accuracy revealed an effect of the *Trial Type* ($p < 0.05$; $F(1,10) = 5.67$) with 92.3% ($SD = 7.92$) accuracy within the blocks and 95.2% ($SD = 4.31$) for random difficulty trials as well as a *Difficulty* effect ($p < 0.01$; $F = 13.67$) with better accuracy for the 0-letter difficulty compared to the 7-letter (83.58%; $SD = 18.86$) and better for the 3-letter compared to the 7-letter difficulty (98.51%; $SD = 5.98$). No significant difference was found between 0-letter and 3-letters difficulty. Finally, a significant *Type* × *Difficulty* interaction was found ($p < 0.05$; $F = 4.75$). However, no significant result was found from the post-hoc pairwise t-tests, meaning that this interaction was only due to a trend in different *Difficulty* effects between the blocks and the random trials.

3.2. Offline *versus* Online RR streams

The Lin's test revealed a strong correlation coefficient between the online and offline RR detection ($\rho = 0.99$; bias correction factor= 0.99; average deviation of 0.02 ms) validating the almost perfect similitude between the two RR series.

3.3. Offline metrics

RR interval. A significant main effect of the *Difficulty* ($p < 0.001$, $F(2,20) = 13.01$) was found in the offline mean RR intervals, corresponding to shorter RR intervals for 7 letters compared to 0 letters ($p < 0.05$) or 3 letters ($p < 0.05$). No significant effect was found between 0 and 3 letters. No effect of *Trial Type* (i.e. block *versus* random) was observed. However, a significant *Type*×*Difficulty* interaction ($p < 0.001; F(2.20) = 12.67$) was found. The post-hoc test did not reveal any significant common difference across the conditions, meaning that this interaction effect was due to differing trends – specifically, higher RR values for the block type for 0 and 3 letters, and higher values for the random type for 7 letters (see Fig. 4).

RR_SD. A significant main effect of the *Difficulty* ($p < 0.01$, $F(2.20) = 5.90$) was found over the offline RR_SD, corresponding to smaller standard deviation for higher difficulty. That is, smaller SD for 7 letters compared to 3 letters ($p < 0.05$) or 0 letters ($p < 0.001$), and for 3 letters compared to 0 letters ($p < 0.05$). In addition, a main *Trial Type* effect was found ($p < 0.001, F(1.10) = 22.01$) with smaller SD values for the random *Type*. No interaction between these two factors was found for offline RR_SD (see Fig. 5).

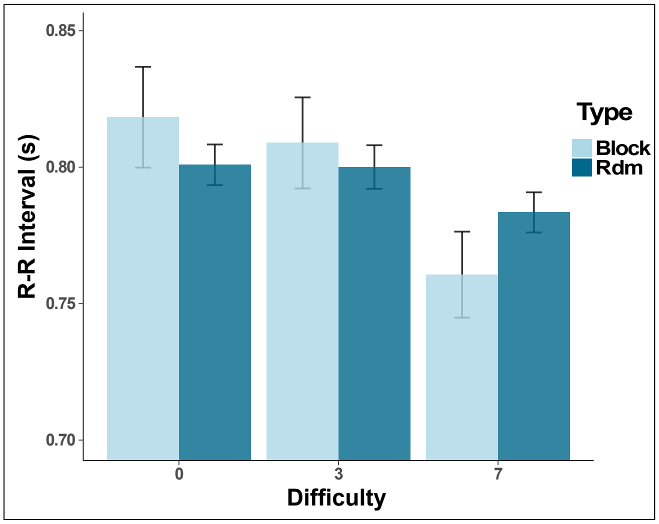

Figure 4. Mean offline RR intervals as a function of trial *Type* and *Difficulty*. Vertical bars indicate the standard deviations.

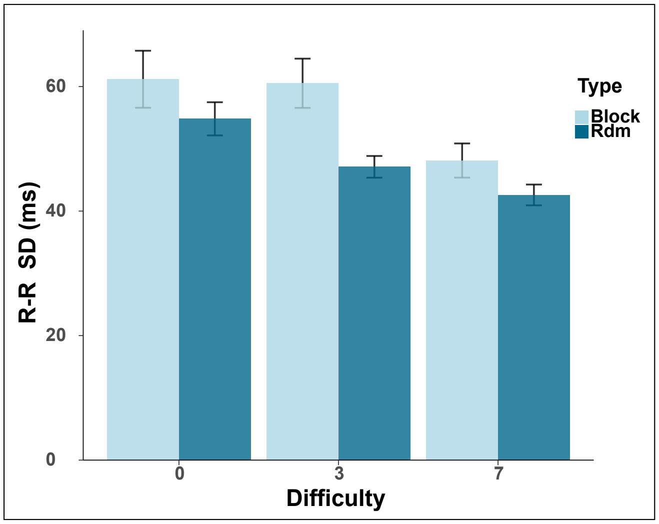

Figure 5. Mean offline standard deviations of the RR intervals as a function of trial *Type* and *Difficulty*. Vertical bars indicate the standard deviations.

3.4. Online metrics

RR interval. As the offline and online RR data series were almost identical (Lin's $\rho = 0.99$), we found the same results as those obtained with the offline method (see section 3.3).

Block type RR_SDs. Within the same-difficulty block condition (long time-window), several online metrics showed significant effects between difficulties. The RR maximum value revealed differences between 7 and 0 letters ($p < 0.005$) or 7 and 3 letters ($p < 0.01$), giving a smaller maximum for 7 letters in both cases. There was no significant effect between 0 and 3 letter blocks. The RR_SD-7 mean also revealed significant differences between 7 letters and 3 or 0 letters ($p < 0.05$ for both comparisons), giving a smaller RR_SD-7 mean for 7 letters (see Fig. 6). Finally, the RR_SD-20 slope, showed a significant effect between all difficulties, but not all in the same direction: an increase in

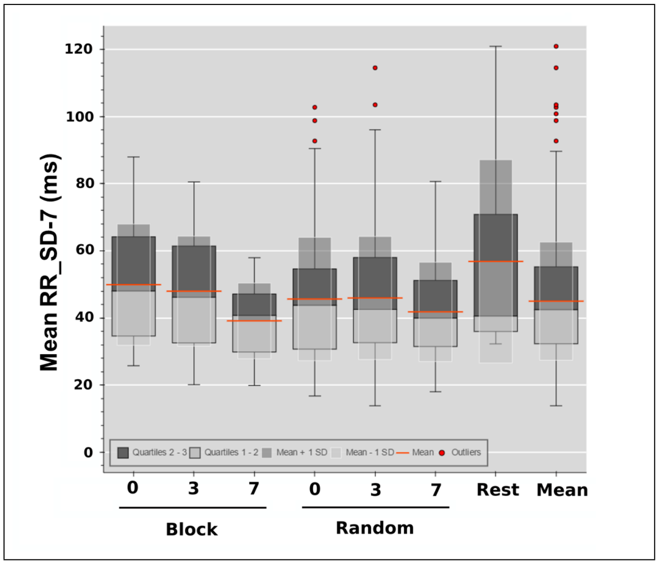

Figure 6. Augmented box plot for the online RR_SD-7 mean stream (ms), grouped by condition, block or random *types* and difficulty (0, 3 and 7-letter span task).

slope from 0 to 3 ($p < 0.05$), a decrease in slope from 0 to 7 ($p = 0.010$), and a decrease in slope from 3 to 7 ($p < 5e^{-6}$). Decreasing the N-values reduced those significant results, although in all cases there was a significant decrease in slope from 3 to 7 letters ($p < 5e^{-4}$).

Random type RR_SDs. Similarly, within the randomized event-related type, several metrics have been found to be sensitive to the trial difficulty. First, the RR maximum showed significant smaller values for 7 letters compared to the two other difficulties ($p < 0.05$). The RR_SD-7 end value revealed lower values for the 7-letter trials compared to the 0-letter ones ($p < 0.01$) as shown in Fig. 7. However, none of the effects between 3 letters and 7 or 0 letters were significant. Stronger values of N reduced the significance to $p < 0.05$. For RR_SD slope, depending on the level of EWMA filtering, statistically significant slope decreases were found either from 0 to 7 letters ($p < 0.05, N = 20$ or 15) or from 0 to 3 letters ($p < 0.05, N = 7$). At all EWMA levels, 7 and 3 letters were similar.

3.5. Classification

Given that the mean offline RR interval and RR_SD metrics led to significant *Difficulty* effects (see Sec. 3.3) at the group level, we used them as control features to evaluate two-dimensional working memory load classifiers using low and high difficulty (0 *versus* 7 letter span task) as labels. These input features are here called Offline metrics. As for the online classifier evaluation, we used the online mean RR interval coupled with the online RR_SD-7 end value as input features. Note that these metrics led to significant *Difficulty* effects, in particular, RR_SD-7 end value for the event-related trial type (see Sec. 3.4). These input features are then here called Online metrics. Inter-subject classification results are summarized in Table 1. Among all used classification techniques, the LDA and the SVM reached the most interesting averaged classification accuracy, ranging from 62.87 to 86.14% accuracy ($1e^{-4} < p - value < 0.01$), depending on the input features, the trial and the classifier types.

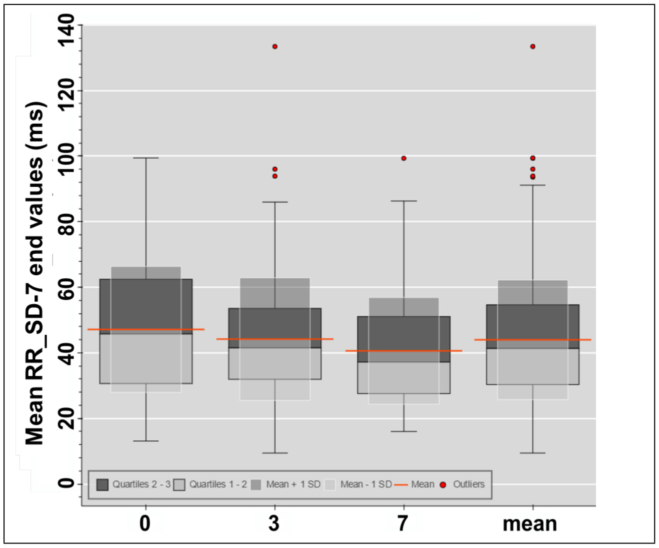

Figure 7. Augmented box plot for the online RR_SD-7 end value stream (ms) as a function of the difficulty for the random event-related trials.

3.6. Summary

With both the online and the offline RR and RR_SD streams we found significant effects of the difficulty. However, only the offline RR_SD demonstrated a main effect of the trial type (lower RR_SD for random event-related trials). The RR streams demonstrated some *Difficulty*×*Type* effect that differed in sense from 0 and 3 to 7 letters. Online streams provided different candidates with different observed effects depending on the length of the time-per-difficulty (i.e. the *trail type*). For blocks of 5 trials of the same difficulty (72s each), the RR mean, RR maximum, RR_SD mean, and RR_SD slope demonstrated significant trial difficulty effects. For random shorter trials (14s each), the RR maximum, RR_SD-7 end value, and RR_SD-20 slope demonstrated similar effects. The classification results showed that the proposed online metrics (mean RR and RR_SD-7 end value) are promising, because they present equivalent or better results for random and block trials type than those obtained with offline metrics.

4. Discussion

The intention of this discussion is to explore the meaning and implications of our findings, in a specific and then a more broad sense.

In this study, the first two mandatory validations have been successfully achieved. First of all, the behavioral results confirm that the chosen difficulty levels induced at least a very low WM load (0 letters) and a highly demanding one (7 letters) with nearly 83% correct responses in the latter.

TABLE 1. MAIN WORKING MEMORY LOAD TWO-DIMENSIONAL CLASSIFICATION RESULTS FOR ECG-BASED METRICS.

		Block					Random				
Metrics vs Classifier	**Score**	**LDA**	**QDA**	**SVM**	**kNN**	**NB**	**LDA**	**QDA**	**SVM**	**kNN**	**NB**
Offline metrics	precision	77.5%	73.5%	**78.72%**	67.12%	71.66%	60.36%	59.65%	**62.87%**	57.94%	59.40%
	recall	0.75	0.725	**0.75**	0.67	0.705	0.582	0.582	**0.614**	0.573	0.594
	f1-score	0.744	0.722	**0.74**	0.67	0.701	0.566	0.566	**0.589**	0.565	0.553
	p-value	p$< 1e^{-3}$	p$< 1e^{-3}$	**p**$< 1e^{-4}$	p$< 1e^{-4}$	p$< 1e^{-3}$	p<0.01	p<0.01	**p**$< 1e^{-3}$	p<0.05	p<0.01
Online metrics	precision	**86.14%**	81.30%	84.64%	76.99%	78.64%	58.48%	60.36%	**63.07%**	57.99%	59.37%
	recall	**0.845**	0.8	0.8	0.755	0.75	0.591	0.591	**0.614**	0.577	0.582
	f1-score	**0.843**	0.798	0.791	0.752	0.743	0.582	0.595	**0.592**	0.5745	0.561
	p-value	**p**$< 1e^{-4}$	p$< 1e^{-4}$	p$< 1e^{-4}$	p$< 1e^{-3}$	p$< 1e^{-4}$	p<0.01	p<0.01	**p**$< 1e^{-4}$	p<0.05	p<0.01

Offline metrics: mean RR and RR_SD offline metrics (windowed). **Online metrics**: Mean RR and RR_SD-7 end value. **LDA**: Linear Discriminant Analyses; **QDA**: Quadratic Discriminant Analyses; **SVM**: Support Vector Machine (with kernel=rbf, $\gamma = 0.1$ and $C = 1$ for blocks, and kernel=rbf, $\gamma = 0.15$ and $C = 2.2$ for random trials); **kNN**: k-Nearest Neighbors (with n_neighbors= 4 for block, and n_neighbors= 6 for random trials); **NB**: Gaussian Naive Bayes.

This also confirms that participants did not disengaged from the task because of its difficulty which would have prevented classifying the 7-letter as a valid high-WM-demanding condition. Secondly, we found that the two RR streams (online and offline) were almost identical, providing proof that the Faros eMotion 360°broadcasts reliable online RR-interval values.

Regarding the offline physiological assessment of the WMC, we found that both the RR interval and the RR standard deviation revealed significant *Difficulty* effects, demonstrating the ability of the physiological measurements to resolve differing trial difficulties. The fact that the main *Trial type* effect was only observable with the RR_SD and that the *Type* × *Difficulty* interaction effect was only observable with the RR interval is an additional argument in favor of using these two measures in complement [10]. In addition, the *Type* × *Difficulty* interaction could be explained by the fact that *changing* high cognitive load results in stronger *transient* effects on the RR intervals than lower loads. The higher transient effects would dominate in the short windows of the random trials, whereas the effects of lower difficulties would be better observed over longer windows.

Finally, using the online mean RR and RR_SD-7 end-value to classify low *versus* high working memory loads, we achieved significant classification results. Given that any objective information about human operator's mental state is a valuable information, one could consequently consider the proposed online metrics as supplementary features for mixed-initiative system development. For instance de Souza and colleagues [6] have proposed to use such physiological data with a Partially Observable Markov Decision Process (POMDP) to improve human-machine interactions.

The experiment here allows us to argue that the online RR mean measurement is effective in distinguishing high-difficulty working memory tasks from less demanding ones. The RR_SD-7 end-value was also effective at distinguishing high vs. low difficulty for the short windows. This measurement encodes in the EWMA (see Eq. 2) the raw online RR_SD stream value taken 14s after the start of a random trial. This means that the stream itself is effective in distinguishing changes from low to high difficulty with a delay of at most 14s. Moreover, these metrics coupled with the online mean-RR conventional metric presented the best classification results for block and random trial types.

The literature review for this work indicates more candidate metrics than those tested here. In the present study, RR and RR_SD, were demonstrated to be successful as online indicators of working memory load, which is an encouragement for testing the remaining metrics, in particular frequency analyses such as: low-frequency (LF), high-frequency (HF), and LF/HF frequency-power [12], [13], [16], [27], [28]. In addition, the results of the candidate identification indicated other metrics that may be of interest that have not appeared in the literature yet, namely the derivative of the RR_SD stream and the first and second derivatives of the RR stream. The ability to detect high cognitive load or changes in engagement in a mere 14 seconds and to resolve subtle load differences in 72 seconds suggests that it should be feasible to integrate such metrics to estimate human operator/driver/pilot cognitive states in adaptive systems as a mixed-initiative framework.

5. Future work

From these results, two branches of future work could be indicated: (i) broadening this research to develop a more comprehensive pool of candidate metrics and algorithms for online cardiac monitoring, and (ii) use of these results to experimentally validate the approach only simulated in [6]. Related to this last point, the experiment described in [9] is now taking place in our lab. It is expected to demonstrate the application of such online cardiac metrics in mixed-initiative human-robot interactions. Note that POMDPs are promising decision frameworks to handle with the *imprecise* output of classifiers. It is known that this sequential decision making model can integrate such uncertainties by means of an observation function. In POMDPs, a belief state (probability distribution over states) is maintained and updated at each decision taken and observation perceived using the Bayes' rule. In this sense, and as already discussed, any information, even uncertain, about the human operator mental state, is valuable information to be integrated.

Acknowledgements

This work has been financially supported by the Dassault Aviation Chair (CASAC 2016-2021).

References

[1] C. W. Nielsen, D. A. Few, and D. S. Athey, "Using mixed-initiative human-robot interaction to bound performance in a search task," in *Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on*. IEEE, 2008, pp. 195–200.

[2] J. A. Adams, P. Rani, and N. Sarkar, "Mixed initiative interaction and robotic systems," in *AAAI Workshop on Supervisory Control of Learning and Adaptive Systems*, 2004, p. 613.

[3] G. Durantin, J.-F. Gagnon, S. Tremblay, and F. Dehais, "Using near infrared spectroscopy and heart rate variability to detect mental overload," *Behavioural brain research*, vol. 259, pp. 16–23, 2014.

[4] N. Régis, F. Dehais, E. Rachelson, C. Thooris, S. Pizziol, M. Causse, and C. Tessier, "Formal Detection of Attentional Tunneling in Human Operator–Automation Interactions," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 3, pp. 326–336, 2014.

[5] S. Scannella, V. Peysakhovich, F. Ehrig, E. Lepron, and F. Dehais, "Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft," *Human factors*, 2018.

[6] P. E. U. de Souza, C. P. C. Chanel, and F. Dehais, "MOMDP-Based Target Search Mission Taking into Account the Human Operator's Cognitive State," *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 729–736, 2015.

[7] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with computers*, vol. 21, no. 1-2, pp. 133–145, 2008.

[8] T. McMahan, I. Parberry, and T. D. Parsons, "Evaluating player task engagement and arousal using electroencephalography," *Procedia Manufacturing*, vol. 3, pp. 2303–2310, 2015.

[9] N. Drougard, C. Carvalho Chanel, R. Roy, and F. Dehais, "An online scenario for mixed-initiative planning considering human operator state estimation based on physiological sensors," in *IROS Workshop in Synergies Between Learning and Interaction (SBLI)*, 2017.

[10] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, 2018.

[11] M. Pagani, F. Lombardi, S. Guzzetti, O. Rimoldi, R. Furlan, P. Pizzinelli, G. Sandrone, G. Malfatto, S. Dell'Orto, and E. Piccaluga, "Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog." *Circulation research*, vol. 59, no. 2, pp. 178–193, 1986.

[12] K. Kudrynski and P. Strumillo, "Real-time estimation of the spectral parameters of Heart Rate Variability," *Biocybernetics and Biomedical Engineering*, vol. 35, no. 4, pp. 304–316, 2015.

[13] K. kudrynski and P. Strumillo, "Real-time estimation of heart rate variability parameters from passband filtered interbeat interval series," *2011 Computing in Cardiology (CinC)*, pp. 297–300, 2011.

[14] A. Andreoli, R. Gravina, R. Giannantonio, P. Pierleoni, and G. Fortino, "SPINE-HRV: A BSN-based toolkit for heart rate variability analysis in the time-domain," *Wearable and Autonomous Biomedical Devices and Systems for Smart Environment*, pp. 369–377, 2010.

[15] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A diagnostic human workload assessment algorithm for human-robot teams," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 123–124.

[16] R. N. Roy, S. Charbonnier, and A. Campagne, "Probing ECG-based mental state monitoring on short time segments," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2013, pp. 6611–6614.

[17] F. Sauvet, J. C. Jouanin, C. Langrume, P. Van Beers, Y. Papelier, and C. Dussault, "Heart rate variability in novice pilots during and after a multi-leg cross-country flight," *Aviation, space, and environmental medicine*, vol. 80, no. 10, pp. 862–869, 2009.

[18] J. Veltman and A. Gaillard, "Indices of mental workload in a complex task environment," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 72–75, 1993.

[19] A. R. A. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and user's guide," *Psychonomic Bulletin & Review*, vol. 12, no. 5, pp. 769–786, oct 2005.

[20] P. J. Gianaros, F. M. Van der Veen, and J. R. Jennings, "Regional cerebral blood flow correlates with heart period and high-frequency heart period variability during working-memory tasks: Implications for the cortical and subcortical regulation of cardiac autonomic activity," *Psychophysiology*, vol. 41, no. 4, pp. 521–530, 2004.

[21] Y.-H. Lee and B.-S. Liu, "Inflight workload assessment: Comparison of subjective and physiological measurements," *Aviation, space, and environmental medicine*, vol. 74, no. 10, pp. 1078–1084, 2003.

[22] American Psychological Association, "Ethical Principles of Psychologists and Code of Conduct," 2017. [Online]. Available: http://www.apa.org/ethics/code/

[23] World Medical Association, "World Medical Association Declaration of Helsinki," *Bulletin of the world health organization.*, 2001.

[24] UCSD Swartz Center for Computational Neuroscience, "Lab Streaming Layer (LSL) GitHub Repository," 2017. [Online]. Available: https://github.com/sccn/labstreaminglayer

[25] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–12, 1990.

[26] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[27] J. F. Thayer, F. Åhs, M. Fredrikson, J. J. Sollers Iii, T. D. Wager, J. J. Sollers, and T. D. Wager, "A meta-analysis of heart rate variability and neuroimaging studies : Implications for heart rate variability as a marker of stress and health," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 2, pp. 747–756, 2012.

[28] G. E. Billman, "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance," *Frontiers in Physiology*, vol. 4, no. February, pp. 1–5, 2013.