Cognitive workload classification using cardiovascular measures and dynamic features

Eydis H. Magnusdottir, Kamilla R. Johannsdottir, Christian Bean, Brynjar Olafsson and Jon Gudnason
Center for Analysis and Design of Intelligent Agents, Reykjavik University, Iceland,
Email: eydis07@ru.is, kamilla@ru.is, christianbean@ru.is, brynjaro14@ru.is, jg@ru.is

***Abstract*—Monitoring cognitive workload has the potential to improve performance and fidelity in human decision making through a real-time monitoring model. Multiple studies have shown a successful binary classification of high and low workload using various methods and often focused on multiple physiological signals. A more detailed detection of cognitive workload is needed for a meaningful and reliable workload monitoring tool. This study focuses on trinary workload classification of parameters extracted from the cardiovascular system. The experiment was validated with the use of a database containing 96 participants performing tasks designed to induce slight variations in cognitive workload. Two distinct supervised learning classifying methods were used and their likelihood score used for the classification schemes of each heartbeat and each screens. The results show that the support vector classifier outperforms the random forest with the average misclassification rate of 20.44% using the whole screen classification scheme instead of individual heartbeat classification.**

I. Introduction

Monitoring cognitive workload in safety-critical job environments, e.g. aviation or surgery, is of paramount importance for enhanced human performance and decision making. The relationship between cognitive workload and performance has been well studied [1] and the connection between cognitive workload and physical health has also been highlighted [2]. The cardiovascular signal, measured through heart rate (HR) and heart rate variability (HRV) or blood pressure, provides a promising method to monitor cognitive workload. The measures are non-intrusive and the cardiovascular system is quite sensitive to cognitive stimuli [3], [4]. Moreover, multiple studies have linked increased cognitive workload to reliable and measurable changes in the cardiovascular signal [5], [6], [7], [8], [9].

However, despite great success, current methods used to analyze the cardiovascular signal, have not been able to detect small workload differences [10], [8], [9]. The majority of the work is based on comparing averages, calculated for the duration of a particular task of higher or lower workload. These studies often fail to detect adjacent levels of increasing or decreasing workload [10], [8], [9]. Few studies have attempted to use various machine learning algorithms to classify cognitive workload states (i.e. higher, lower) [11], [12], [13], [14]. The results, based on EEG signals along with other physiological measures, show a successful binary (high, low) classification of cognitive workload. A more detailed detection of multiple cognitive workload states, particularly based on the cardiovascular signal, remains a challenge.

The main objective of this work is to provide a trinary classification of cognitive workload with parameters extracted solely from the cardiovascular system. This was done by adding time dynamic measures to the static cardiovascular signals and using supervised learning methods for classification. A cognitive workload experiment was conducted with variations in cognitive workload introduced through different variations of the original Stroop task [15]. Speech- and cardiovascular information were recorded from 96 participants solving the set tasks by responding verbally to the stimuli. The feature set contained 10 cardiovascular signal parameters for each heartbeat comprising of a detailed hemodynamic profile and derived delta and delta-delta coefficients for each of these. A separate leave-one-out participant-dependent classifier of the three difficulty levels was trained for each participant. Two supervised learning classifiers were used; support vector machines (SVM) and random forest (RF) and the likelihood score was evaluated through two classification schemes, for each heartbeat and for each task screen. The conclusion is that a finer cognitive workload distinction can be reached with the combined feature set of cardiovascular signals and delta coefficients, classifying for each screen with the SVM classifier.

The context of the work is presented in Sec. II. The experimental methodology, cardiovascular measures, feature extraction, classification setup and design are described in Sec. III. The results are detailed in Sec. IV and the paper is concluded with discussion in Sec. V.

II. Cardiovascular signal analysis for detecting cognitive workload

Measurements of cognitive workload are generally gathered through three methods; (1) self assessment, (2) behavioral monitoring and (3) by measuring physiological signals. Most commonly the focus has been on physiological measures, in particular, cardiovascular reactivity of the individual to a performed task. Measuring the individuals cardiovascular reactivity is well suited for the automatic monitoring of cognitive workload as it is relatively non-intrusive, objective, and takes place in real time [16].

Several studies have shown a reliable cardiovascular reactivity to increased workload [5], [6], [17], [7], [18], [19]. However, the detection of cognitive workload through cardiovascular reactivity has to date mainly been binary. That is, current methods detect high and low workload, or task onset and task offset [20]. These method, often fail to distinguish between multiple adjacent levels of increasing or decreasing workload. For example, in a study by Wilson [9], HR successfully distinguished takeoff and landing from other segments of flight in a visual and instrumented flight rules simulated flights (VFR and IRF). However, neither HR nor HRV distinguished between other flight segments (22 in total), although they all varied in their load level. Furthermore, in a study by Vogt et al. [8], HR significantly detected increased number of aircrafts in two different simulations (en-route and tower) and increased number of conflicts in en-route simulation. HR however, did not distinguish

This work is sponsored ISAVIA, Icelandair and the Icelandic Center for Research (RANNIS) under the project *Monitoring cognitive workload in ATC using speech analysis*, Grant No 130749051.

higher load of vertical traffic or pilot error in en-route simulation nor predictable or unpredictable conflict in tower simulation.

The problem is largely methodological. For most parts, workload detection is based on calculating and comparing averages over extended task periods that vary in their level of workload. The problem with this approach is that it forgoes the information embedded within the time segments for different tasks such as the body's natural self-adjustment mechanism to long term fluctuations, the baroreflex [21]. It has therefore been suggested that if workload is being detected through comparing means, only the first 30 to 60 seconds should be considered [22], [23], [24], [25], [21]. Longer time segments may reflect a combined activity of two opposing systems, the sympathetic and the parasympathetic system [25], [21].

Another approach is to classify changes in cognitive workload state using various machine learning algorithms. Few studies have shown that a binary (high, low) classification of cognitive workload is possible by training neural networks using multiple physiological signals, including EEG [11], [12], [13], [14]. Other machine learning algorithms include discrimination analysis and support vectors (SVM) (see [26], [27], [28]), as well as logic regression and classification trees [11], and kernel partial least square classifier [29]. For most parts these multi signal classification methods gain high accuracy for a two level classification (i.e. high, low; present, absent). But a more detailed classification seems hard to reach. Brouwer et al. [26] used SVM to detect memory load using the n-back task. They trained the model on first three quarters of task performance and tested it on the last quarter of task performance. Binary classification between all three levels of memory load (0-back, 1-back, 2-back) were possible, but the highest accuracy was gained for classifying the greater load differences (high, low). Higher accuracy was also gained when both ERP and spectral power signals were used together. In general, most of the prior work has focused on EEG along with other physiological signals, showing high binary classification accuracy. Very little work has been done on classifying cognitive workload based solely on cardiovascular reactivity. But it has been pointed out that a classification based on the cardiovascular signal may be more suitable and practical in real task environments compared to using EEG [29].

Yin et al. [30], achieved 77.5% classification accuracy for three task-difficulty levels using only the speech signal. In their study, participants performed different variations of the original Stroop task in addition to a standard reading task. They concatenated standard mel-frequency cepstrum coefficients with prosodic features and used Gaussian mixture models to classify workload performance. Yap et al. [31] continued with this line of research by adding voice source features based on cepstral peak prominence and produced results for three Stroop test levels.

In the current study the focus is set on classification of cardiovascular signals with velocity and acceleration information from 5 heartbeat frames to include time dynamic information. The classification tasks are performed with two types of supervised learning methods with testing the classification schemes of screens or heartbeat.

III. Methodology

A. Experiments and data

A total of 96 participants visited the laboratory for a session of tasks lasting 45-50 min. During the session, speech and cardiovascular signals were recorded were the participant was engaged in Stroop tasks with 3 min. resting periods between cognitive workload levels. The Stroop tasks [15] were designed with a set (35+1) of words of five colors appearing on a screen and the participants task was to say the color of the words aloud. The cognitive workload levels were introduced with various levels of congruency, in-congruency and time limits.

Level 1 Congruent sets of screens with all the words appearing at the same time.

Level 2 In-congruent sets of screens with the alternating two levels of 0.3 and 0.7 of in-congruency with all the words appearing at the same time.

Level 3 Sets of screens with one word appearing at a time at randomly timed intervals of 0.75 sec. and 0.65 sec. Here the same in-congruency setup was used as in Level 2.

A Latin square technique was used to alternate the order of the cognitive workload levels into six between participants.

B. Cardiovascular measures and delta features

The Finometer Pro from Finapres was used to record cardiovascular signals of the participants during the experiments [32], [33]. The signals were obtained using a finger cuff and an upper arm cuff was used for calibration of the reconstructed blood pressure. Ten measures for each heartbeat were obtained from the output of the Finometer Pro system. These measures are 1) Heart rate, 2) Systolic pressure, 3) Diastolic pressure, 4) Mean pressure, 5) Pulse interval, 6) Stroke volume, 7) Left ventricular pressure energy, 8) Cardiac output, 9) Total peripheral resistance, 10) Maximum steepness and are designated as $c_{t,i}$ where $i \in \{1, 2, \ldots, 10\}$ is the index for the measure and t is the integer time index of the heartbeat. The ten dimensional feature vector for a heartbeat at time t is therefore $\mathbf{c}_t = [c_{t,1}, c_{t,2}, \ldots, c_{t,10}]^T$.

The vector $\mathbf{c}_t$ only includes information about the current heartbeat. Time derivatives are appended to the vector to include information about the rate of change of the static cardiovascular measures. These dynamic measures are called delta $\delta_t^{(1)}$ and delta-delta $\delta_t^{(2)}$ coefficients and are calculated from the measure $c_{t,i}$ along the time index by using,

$$\delta_{t,i}^{(1)} = \frac{\sum_{n=1}^{N} n(c_{t+n,i} - c_{t-n,i})}{2\sum_{n=1}^{N} n^2}, \tag{1}$$

where N denotes the number adjacent heartbeat measures before and after t used to derive the delta feature. This is set to $N = 2$ for this work (resulting in a window size of 5). A ten dimensional delta vector is therefore obtained by $\boldsymbol{\delta}_t^{(1)} = [\delta_{t,1}^{(1)}, \delta_{t,2}^{(1)}, \ldots, \delta_{t,10}^{(1)}]^T$ and the delta-delta feature vector $\boldsymbol{\delta}_t^{(2)}$ are calculated using the same formula but using $\delta_{t,i}^{(1)}$ instead of $c_{t,i}$. Three feature sets were evaluated in this work, one without delta and delta-delta features, i.e. just the ten dimensional $\mathbf{c}_t$, the second only with the delta features, i.e. the twenty dimensional $\mathbf{d}_t = [\mathbf{c}_t^T, \boldsymbol{\delta}_t^{(1)T}]^T$ and the third with all the vectors concatenated, i.e. the thirty dimensional $\mathbf{e}_t = [\mathbf{c}_t^T, \boldsymbol{\delta}_t^{(1)T}, \boldsymbol{\delta}_t^{(2)T}]^T$.

C. Classifier design

The classifiers designed and implemented in this work are based on the cardiovascular measures either from a single heartbeat feature vector $\mathbf{x}_t$ (without delta features, $\mathbf{x}_t = \mathbf{c}_t$ or with delta features, $\mathbf{x}_t = \mathbf{d}_t$ or $\mathbf{x}_t = \mathbf{e}_t$) or a sequence of heartbeats from a single screen represented with the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]^T$. The classifiers return soft scores and are denoted either as $y_k(\mathbf{x}_t)$ for the single heartbeat classifier or $y_k(\mathbf{X})$ for the sequence classifier. The index $k \in \{1, 2, 3\}$ denotes Stroop level one, two or three respectively. The vector $\mathbf{y}(\mathbf{x}_t) = [y_1(\mathbf{x}_t), y_2(\mathbf{x}_t), y_3(\mathbf{x}_t)]^T$ contains the soft scores for the heartbeat at time index t and the heartbeat classification simply chooses the class with the maximum value in that vector. For the sequence classification, the soft scores are collected in an output matrix $\mathbf{Y} = [\mathbf{y}(\mathbf{x}_1), \mathbf{y}(\mathbf{x}_2), \ldots, \mathbf{y}(\mathbf{x}_T)]^T$. The classification is then obtained by summing the soft scores together over the screen to obtain $\mathbf{y}(\mathbf{X}) = [y_1(\mathbf{X}), y_2(\mathbf{X}), y_3(\mathbf{X})]^T$ and then the class according to the maximum value of that vector is chosen.

Two types of supervised learning classifiers were implemented using the statistics toolbox in Matlab: Support Vector Machine (SVM), $\mathbf{y}_{SVM}(\mathbf{x}_t)$ and Random Forests (RF), $\mathbf{y}_{RF}(\mathbf{x}_t)$. A support vector machine is fundamentally a binary classifier so to solve the trinary classification problem, three two-class one-vs-all binary SVM classifiers were implemented, one for each Stroop level. The soft score output $y_k(\mathbf{x}_t)$ is the signed distance from the decision boundary for each of the three classifiers where $k \in \{1, 2, 3\}$. The class is then determined by the one-vs-all classifier which obtains the maximum signed distance from the decision boundary. If all scores are negative then the class that is the closest to the decision boundary with the least negative score is chosen.

A random forest classifier is trained for each heartbeat using one hundred decision trees. The minimum number of observations in a leaf was set to one. The soft score $y_k(\mathbf{x}_t)$ for the random forest classifier is the proportion of trees in the ensemble predicting class k and is interpreted as the probability of this observation $\mathbf{x}_t$ originating from this class.

D. Classifier training and evaluation

As this work does not deal with differences between individuals, only participant-dependent classifiers were trained in this work. This means that 96 separate classifiers were trained and tested, one for each participant. A leave-one-out strategy was used where the test sample that is left out corresponds to a single screen. The other twenty screens for that participant were used to train the classifiers. In the case of single heartbeat classifiers the number of test results corresponded to the number of heartbeats contained within the test screen (an average of 776) but in the case of the sequence classification for the entire screen only one result was obtained. The experiment was then repeated where another screen from the set of twenty-one screens was reserved for testing. The results were then cumulated over the twenty-one trials. This procedure was then repeated for each participant in the study.

IV. Results

The results are first presented for the support vector machine using using delta-delta features. We show how we construct a confusion table from the leave-one-out experiments and compare the single heartbeat classifier with the sequence screen classifier. The results are then summarized as single average misclassification rates and compared with the other feature sets and the random forest classifier. Finally, a further analysis of the performance of individual participants is presented.

A. Support vector machines using delta-delta features

Table I shows the confusion table of how all heartbeats in the data set are classified using the support vector machine classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$ and the delta-delta feature set $\mathbf{e}_t$. The numbers are obtained by summing individual confusion tables of each participant. The table also shows the misclassification rate and the mistrust rate of each Stroop level. For example, if there is a 32.29% chance of misclassification if the heartbeat is known to be from Stroop Level 2 and if an unknown heartbeat is classified as Stroop Level 2, there is a 30.99% chance of an error. There errors seem to be evenly distributed over the remaining classes which indicates that no level is dominating the classification. However, Stroop Level 3 seem to be performing slightly better than the other levels with a 23.13% misclassification rate and 25.69% mistrust rate. The participant-average test set misclassification rate is given as 29.26%.

TABLE I. *Confusion matrix for the number of heartbeats classified with the $\mathbf{e}_t$ feature set classifying with the SVM.*

	Stroop L1	**Stroop L2**	**Stroop L3**	**MCR [%]**
Stroop L1	**13790**	3749	3554	34.62
Stroop L2	3745	**16536**	4142	32.29
Stroop L3	3022	3676	**22254**	23.13
Mistrust rate [%]	32.92	30.99	25.69	**29.26**

Table II shows the confusion table of how all screens in the data set are classified also using the support vector machine classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$ and the delta-delta feature set $\mathbf{e}_t$. As with the heartbeat confusion Table I, the numbers are obtained by summing individual confusion tables of each participant. The misclassification and mistrust rates are also shown for each Stroop level and the participant-average test set misclassification rate is reported. The table shows that the screen based classifier is also balanced with respect to the Stroop levels.

TABLE II. *Confusion matrix for the number of screens classified with the $\mathbf{e}_t$ feature set classifying with the SVM.*

	Stroop L1	**Stroop L2**	**Stroop L3**	**MCR [%]**
Stroop L1	**499**	95	78	25.74
Stroop L2	63	**440**	73	23.61
Stroop L3	38	65	**665**	13.41
Mistrust rate [%]	16.83	26.66	18.50	**20.44**

The advantage of using a sequence of heartbeats is clearly seen by comparing Table I and Table II. The participant-average test set misclassification rate improves from 29.26% to 20.44%.

B. Overall performance results

Table III shows the participant-average test set misclassification rate for the support vector machine classifier. Each line shows the results for the three feature sets: without delta features, $\mathbf{c}_t$, with delta features, $\mathbf{d}_t$ and with delta and delta-delta features $\mathbf{e}_t$. The two columns show the results for the screen sequence classifier $\mathbf{y}_{SVM}(\mathbf{X})$ and the individual heartbeat classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$. Each result shows the average and the standard deviation of the individual participant test set misclassification rates. For example the SVM screen classifier using the cardiovascular feature set without delta features, $\mathbf{c}_t$, obtains 26.34% misclassification rate on average over all participants and the standard deviation over all participants is 16.95%. The table shows that adding delta and delta-delta features improves performance considerably for both the screen based and the heartbeat based SVM classifier. The best results of $20.44 \pm 15.48\%$ are obtained using delta and delta-delta features, $\mathbf{e}_t$ and the screen based classifier $\mathbf{y}(\mathbf{X})$. These results are also the best results obtained in this work and correspond to the confusion table presented in Table II. Further analysis of these results are presented in Section IV-C.

TABLE III. *Average MCR [%] over all participants comparing different combinations of classification schemes and feature sets for the SVM classifier.*

Feature set $\mathbf{x}_t$	**Screen** $\mathbf{y}(\mathbf{X})$	**Heartbeat** $\mathbf{y}(\mathbf{x}_t)$
$\mathbf{c}_t$	26.34±16.95	36.11±13.88
$\mathbf{d}_t$	23.42±17.08	33.73±14.11
$\mathbf{e}_t$	**20.44±15.48**	29.26±13.27

Table IV shows the set of results for the random forest classifier. The overall results are inferior to the support vector machine results. For the heartbeat based random forest classifier the performance improves by adding delta and delta-delta features which is consistent with the support vector machines results. However the screen based classifier is not improved by adding the delta or delta-delta features.

TABLE IV. *Average MCR [%] over all participants comparing different combinations of classification schemes and feature sets for the RF classifier.*

Feature set $\mathbf{x}_t$	**Screen** $\mathbf{y}(\mathbf{X})$	**Heartbeat** $\mathbf{y}(\mathbf{x}_t)$
$\mathbf{c}_t$	22.72±15.99	37.21±12.79
$\mathbf{d}_t$	23.91±15.69	35.75±12.73
$\mathbf{e}_t$	24.21±15.89	34.50±12.76

C. Participant distributions

Figure 1 shows the histogram of the test set misclassification rates for all participants for the support vector machine classifier and delta and delta-delta features. The first panel shows the results for the single heartbeat classifier and the second panel shows the screen based classifier. The figure corresponds to the last line in Table III and demonstrates how the results improve when the classifier can accumulate the results over an entire screen before making decision. The figure also shows how widely distributed the results are depending on the participants, as is evident in the standard deviation. The range for the screen based classifier is from zero misclassification rate to 76.19%. More interesting results evident from the histogram is that 34 participants (out of 96) achieve misclassification rate under 10%.

Figure 2 shows the histogram of the misclassification rate when using the Random forest classifier without dynamic features. The figure shows how well the performance improves when using a sequence of heartbeats to do the classification. The average misclassification rate for the screen-based classification is 22.72% and the range if from zero to 85.71% (not shown in the figure). The number of participants achieving misclassification rate under 10% is 22 participants (out of 96).

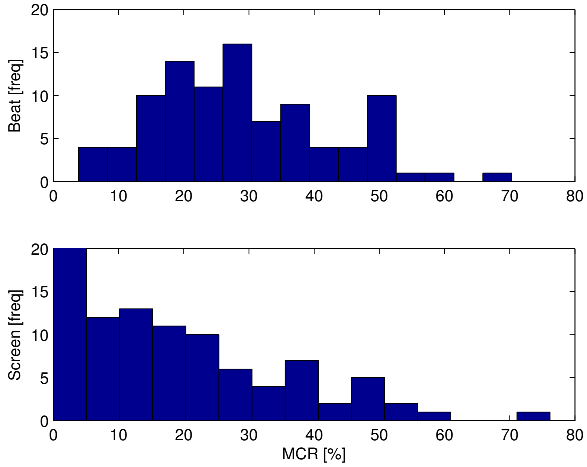

Fig. 1. *Histogram of test set misclassification rate (MCR) [%] for all participants in the set using the SVM classifier with the delta and delta-delta features. The upper panel shows the individual heartbeat classifier results and the lower panel shows the screen-based classifier results.*

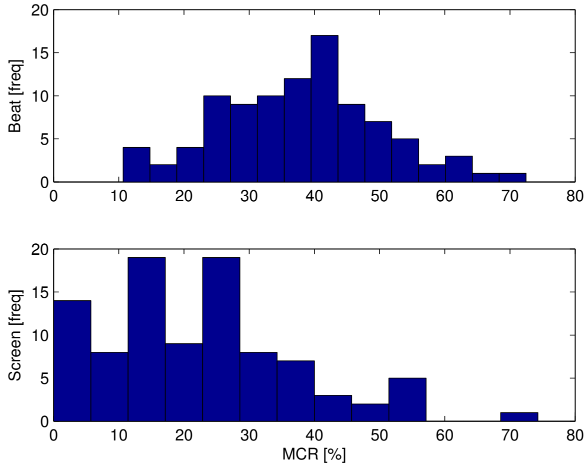

Fig. 2. *Histogram of test set misclassification rate (MCR) [%] for all participants in the set using the RF classifier without the dynamic features. The upper panel shows the individual heartbeat classifier results and the lower panel shows the screen-based classifier results.*

V. Discussion and Conclusions

The experiments presented in this work show how the cardiovascular system responds to cognitive stimuli when using the Stroop tests. Support vector machines and Random forests were used as classifiers for static cardiovascular measures and their dynamic features. The results show obvious cardiovascular reactivity to the different cognitive stimuli. The Stroop levels that were tested in this work all demand the same or very similar physical workload from the participant. The study does not include cardiovascular measures from the baseline period when the participant is completely at rest. The classification results can therefore be interpreted solely on the cognitive workload level that is elicited using the different Stroop levels. The best classifier is able to distinguish between low, medium and high Stroop level with an average of 20.44% test set misclassification rate. Further analysis shows that more than 34 participants out of 96 achieve 10% misclassification rate or less. This shows that cognitive workload strongly affects the cardiovascular system. The methodology does not however give an insight into what aspect of the cardiovascular measures are affected.

The results also showed that the support vector machine classifier outperformed the random forest classifier. The best SVM results is 20.44% while the best RF result is 22.72%. Both methods benefit greatly from classifying on a whole sequence of heartbeats (screens) instead of individual heartbeats.

The work also introduced dynamic features for the cardiovascular methods that have been called delta and delta-delta features in the speech processing community. Both the SVM and RF classifiers that are based on single heartbeats benefited from the addition of dynamic features. However the screen-based RF classifier did not benefit from the addition.

Signal classification offers a new approach to cognitive workload monitoring. The results presented here compare well with other classification work in the field. For most parts, reported results are based on a binary classification, sometimes using no vs. some or low vs. high cognitive workload [11], [12], [13], [14], [26]. It is clear from this work, that high accuracy in binary classification is possible, in particular, if combining multiple physiological signals. The obstacle however, has been to move beyond the binary classification.

In Wilson and Russel [13] a high accuracy was gained using an ANN for classifying high and low cognitive workload. Their classification was based on multiple physiological signals, including EEG. The network however, was not capable of classifying accurately between more than two cognitive workload states (high and low). When the training scenario included four and seven different states based on complexity and number of aircraft, the classifier confused adjacent states, not distinguishing between low and medium or medium and high. Similarly, Brouwer et al. [26] using a SVM to classify mental load, found a high accuracy in classifying high and low load, whereas, low to medium or medium to high classification was not as accurate. Their classification was also based on EEG and higher accuracy was gained when two EEG signals (ERP and spectral power) were combined.

The present study goes beyond the state of the art by successfully demonstrating a trinary classification of low, medium and high cognitive workload states. Furthermore, the classification here is based entirely on the cardiovascular signal whereas prior work has for most parts focused on EEG often along with other physiological signals. Few studies have attempted to classify workload based on the speech signal. In Magnúsdóttir et al. [34], three load states were classified using vocal tract features and SVM. However, the classification accuracy was 67,5%. The present results show a better classification accuracy using the cardiovascular signal. In both cases, variation of the original Stroop task were used to induce cognitive workload.

In conclusion, cognitive workload classification based on the cardiovascular signal is possible and might provide a reliable, non-intrusive monitoring tool in the field. The results here provide an important input to the literature on workload monitoring as prior work has mainly focused on binary (high, low) classification using multiple physiological signals, including EEG. But as pointed out by Fong [11], basing the classification on the cardiovascular signal may be more suitable in the field compared to using the EEG signal. The trinary classification accuracy reached here is comparable to binary (high, low) classification accuracy reached in prior work using multiple physiological signals.

References

[1] E. Galy, M. Cariou, and C. Mélan, "What is the relationship between mental workload factors and cognitive load types?" *International Journal of Psychophysiology*, vol. 83, no. 3, pp. 269–275, Mar. 2012.

[2] M. A. Kompier, B. Aust, A. M. van den Berg, and J. Siegrist, "Stress prevention in bus drivers: evaluation of 13 natural experiments." *Journal of occupational health psychology*, vol. 5, pp. 11–31, 2000.

[3] B. Mehler, B. Reimer, and M. Zec, "Defining workload in the context of driver state detection and HMI evaluation," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2012, pp. 187–191.

[4] M. W. Scerbo, "Stress, workload, and boredom in vigilance: a problem and an answer." *Stress, workload, and fatigue*, 2001.

[5] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, Jul. 2014.

[6] N. Dahlstrom and S. Nahlinder, "Mental workload in aircraft and simulator during basic civil aviation training," *The International journal of aviation psychology*, vol. 19, no. 4, pp. 309–325, 2009.

[7] R. N. Nolte, R. A. Wright, C. Turner, and R. J. Contrada, "Reported fatigue, difficulty, and cardiovascular response to a memory challenge," *International Journal of Psychophysiology*, vol. 69, no. 1, pp. 1–8, 2008.

[8] J. Vogt, T. Hagemann, and M. Kastner, "The impact of workload on heart rate and blood pressure in en-route and tower air traffic control," *Journal of psychophysiology*, vol. 20, no. 4, pp. 297–314, 2006.

[9] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, Jan. 2002.

[10] D. B. Kaber, C. M. Perry, N. Segall, and M. A. Sheik-Nainar, "Workload state classification with automation during simulated air traffic control," *The International Journal of Aviation Psychology*, vol. 17, no. 4, pp. 371–390, 2007.

[11] A. Fong, C. Sibley, A. Cole, C. Baldwin, and J. Coyne, "A comparison of artificial neural networks, logistic regressions, and classification trees for modeling mental workload in real-time," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 2010, pp. 1709–1712.

[12] G. F. Wilson, "Real-time adaptive aiding using psychophysiological operator state assessment," *Publication of: Ashgate Publishing Company*, 2001.

[13] G. F. Wilson and C. A. Russell, "Operator Functional State Classification Using Multiple Psychophysiological Features in an Air Traffic Control Task," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 45, no. 3, pp. 381–389, Sep. 2003.

[14] G. F. Wilson, J. Estepp, and I. Davis, "A comparison of performance and psychophysiological classification of complex task performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53. SAGE Publications Sage CA: Los Angeles, CA, 2009, pp. 141–145.

[15] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of experimental psychology*, vol. 18, no. 6, p. 643, 1935.

[16] F. Psychophysiologie, "Recording methods in applied environments," *Engineering psychophysiology: Issues and applications*, p. 111, 2000.

[17] B. Mehler, B. Reimer, and Y. Wang, "A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload," in *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2011, pp. 590–597.

[18] R. C. Smith, "Stress, anxiety, and the air traffic control specialist: Some conclusions from a decade of research," Federal Aviation Administration Washington DC Office of Aviation Medicine, Tech. Rep., 1980.

[19] J. Taelman, S. Vandeput, E. Vlemincx, A. Spaepen, and S. Van Huffel, “Instantaneous changes in heart rate regulation due to mental load in simulated office work,” *European journal of applied physiology*, vol. 111, no. 7, pp. 1497–1505, 2011.

[20] E. A. Byrne and R. Parasuraman, “Psychophysiology and adaptive automation,” *Biological psychology*, vol. 42, no. 3, pp. 249–268, 1996.

[21] A. Stuiver, K. A. Brookhuis, D. de Waard, and B. Mulder, “Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload,” *International Journal of Psychophysiology*, vol. 92, no. 1, pp. 35–41, Apr. 2014.

[22] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. M. Mulder, “Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight,” *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[23] D. de Waard, A. Kruizinga, and K. A. Brookhuis, “The consequences of an increase in heavy goods vehicles for passenger car drivers’ mental workload and behaviour: a simulator study,” *Accident Analysis & Prevention*, vol. 40, no. 2, pp. 818–828, 2008.

[24] A. F. Kramer, “Physiological metrics of mental workload: A review of recent progress,” *Multiple-task performance*, pp. 279–328, 1991.

[25] A. Stuiver, D. De Waard, K. A. Brookhuis, C. Dijksterhuis, B. Lewis-Evans, and L. J. M. Mulder, “Short-term cardiovascular responses to changing task demands,” *International Journal of Psychophysiology*, vol. 85, no. 2, pp. 153–160, 2012.

[26] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, “Estimating workload using EEG spectral power and ERPs in the n-back task,” *Journal of neural engineering*, vol. 9, no. 4, p. 045008, 2012.

[27] M. A. Hogervorst, A.-M. Brouwer, and J. B. van Erp, “Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload,” *Frontiers in neuroscience*, vol. 8, 2014.

[28] C. A. Russell and G. F. Wilson, “Feature saliency analysis for operator state estimation,” in *Proceedings of the 11th International Conference on Human-Computer Interaction*, vol. 11. Foundations of Augmented Cognition, 2005.

[29] L. J. Trejo, N. J. McDonald, R. Matthews, and B. Z. Allison, “Experimental design and testing of a multimodal cognitive overload classifier,” *Foundations of Augmented Cognition*, pp. 13–22, 2007.

[30] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, “Speech-based cognitive load monitoring system,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2041–2044.

[31] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. C. E. Member), “Formant Frequencies under Cognitive Load: Effects and Classification,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 219253, Dec. 2011.

[32] “Finapres Medical Systems | Products - Finometer® PRO.” [Online]. Available: http://www.finapres.com/Products/Finometer-PRO

[33] I. Guelen, B. E. Westerhof, G. L. van der Sar, G. A. van Montfrans, F. Kiemeneij, K. H. Wesseling, and W. J. Bos, “Finometer, finger pressure measurements with the possibility to reconstruct brachial pressure,” *Blood pressure monitoring*, vol. 8, no. 1, pp. 27–30, 2003.

[34] E. H. Magnusdottir, M. Borsky, M. Meier, K. Johannsdottir, and J. Gudnason, “Monitoring Cognitive Workload Using Vocal Tract and Voice Source Features,” *Periodica Polytechnica Electrical Engineering and Computer Science*, May 2017.