31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Mental Workload Classification using Heart Rate Metrics

Andreas Henelius, Kati Hirvonen, Anu Holm, Jussi Korpela and Kiti Müller

Abstract— **The ability of different short-term heart rate variability metrics to classify the level of mental workload (MWL) in 140 s segments was studied. Electrocardiographic data and event related potentials (ERPs), calculated from electroencephalographic data, were collected from 13 healthy subjects during the performance of a computerised cognitive multitask test with different task load levels. The amplitude of the P300 component of the ERPs was used as an objective measure of MWL. Receiver operating characteristics analysis (ROC) showed that the time domain metric of average interbeat interval length was the best-performing metric in terms of classification ability.**

I. Introduction

In several modern professions physical demands have decreased while mental workload (MWL) has increased [1]. It is therefore important to be able to evaluate the degree of MWL of different tasks, to promote safe and productive working environments. This is especially important in safety-critical professions. Measurements of cardiovascular responses have been used in psychophysiologic studies to investigate MWL in different task load conditions [2], [3].

Electrocardiography (ECG) is an easy and noninvasive method of recording cardiac activity. Several heart rate variability (HRV) metrics have been presented, that provide different information on cardiac chronotropy [4]. No optimal metric has been presented so far [5].

The aim of this study was to investigate the classification ability of short-term HRV metrics to identify the metrics that best separate two different levels of MWL during a computerised cognitive multitask test. The area under the receiving operating characteristics (ROC) curve, denoted AUC, was used to describe the classification performance of the metrics. The existence of different levels of MWL during a computerised multitask test was verified using event related potentials (ERPs).

II. Methods

A. Experimental Procedure

The Brain@Work computerised multitask test [6] developed at the Finnish Institute of Occupational Health (FIOH) was used to induce MWL. Three different tasks were used in the multitask test; an auditory task, a mental arithmetic task and a memory task.

All authors are with the Brain Work Research Centre, Finnish Institute of Occupational Health, Helsinki, Finland.
Corresponding author: Andreas Henelius
Email: andreas.henelius@ttl.fi
Address: Finnish Institute of Occupational Health
Topeliuksenkatu 41 a A, 00250 Helsinki, Finland

In the auditory task, two tones of frequencies 1000 Hz and 1200 Hz (target tone) were presented in an oddball paradigm, with the target tone occurring with a 20% probability. The interstimulus interval (ISI) was 1500 ms. The subject was instructed to press a button with the left hand whenever the target tone was presented. In the mental arithmetic task, the subject was to mentally sum three-digit numbers, entering the answer using the mouse by clicking a keypad shown on the screen. At the start of the computerised test, the subjects were shown a list of target letters. In the memory task, the subjects were then shown random letters one at a time and for each letter presented they had to recall whether the letter was among the target letters by clicking a "yes" or "no" button shown on the screen.

It was hypothesised, that a greater number of parallel tasks exerts a greater degree of MWL. Using the three tasks, three different task blocks with a different number of simultaneous tasks were formed. The single task (S) consisted only of the auditory task, the dual task (D) consisted of the auditory and mental arithmetic task, and the multi task (M) consisted of all three tasks. Each task block lasted for 7 minutes and each task block was presented twice for the subject, in the order S1–D1–M1–S2–M2–D2, where the letter denotes the task type (single, dual or multi) and the number following the task type denotes the first or second occurrence of the task block. In the analysis only the low MWL (blocks S1 and S2) and high MWL (blocks M1 and M2) task conditions were used to guarantee a maximum difference in MWL.

B. Subjects

The study group consisted of 13 voluntary subjects (7 men) with a mean age of 32 years (range 23–51, standard deviation 9.62) All subjects were employees of the FIOH.

C. Data Acquisition

The ECG was measured using a single-lead, with one electrode placed below the right clavicle and the other electrode on the lower left ribcage. This configuration was chosen to maximise the amplitude of the ECG R-wave. EEG (channels Fz, Cz, Pz and Oz, according to the International 10-20 electrode system [7]) were also recorded. All data was digitised at a sampling rate of 500 Hz using a SynAmps amplifier (Compumedics Neuroscan, Charlotte, New Carolina, USA). The EEG data was used for the calculation of ERPs. The ERP P300 component amplitude has been found to decrease with an increase in mental workload [8]. Respiration was not controlled. Measurements were made in the sitting position.

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

TABLE I
TIME AND FREQUENCY DOMAIN HRV METRICS.

Metric	Unit	Description
Time domain metrics		
Average IBI	ms	Average IBI length
Std IBI	ms	Standard deviation of IBI lengths
RMSSD	ms	Root mean square of successive differences of IBIs
Frequency domain metrics		
P_{tot}	ms^2	Total power (sum of absolute power in VLF, LF and HF bands)
LF_p, HF_p	Hz	Peak frequency in LF and HF band
LF_{abs}, HF_{abs}		Absolute power in LF and HF band divided by square of mean IBI.
$LF_{\%}$, $HF_{\%}$	%	Relative power in LF or HF band to total power
LF_{nu}, HF_{nu}	nu	Normalised power in LF and HF band (absolute VLF band power subtracted from total power)
$^{LF}/_{HF}$		Ratio of LF_{abs} to HF_{abs}

D. Data Analysis

The data was analysed using the Matlab-based Biosignals analysis package [9]. All HRV analyses were performed using Matlab R2008a (The MathWorks Inc., Natick, Massachusetts). The data was visually screened for artifacts and correct identification of the R-peak was verified. Artifacts were manually corrected.

The analysis was performed both in the time domain and frequency domain, using 140 s analysis segments, with 130 s overlap, yielding 29 data points per subject in each task block, giving a total of 116 data points per subject. These individual data points were used in the ROC analysis, where they were classified according to the known level of MWL (low or high).

The use of 140 s analysis segments allows both the LF and HF frequency components to be determined from the data [5]. The two frequency bands used in HRV analysis were defined as follows: low frequency (LF): 0.04-0.15 Hz and high frequency (HF): 0.51-0.40 Hz [10]. Measures from the very low frequency band (VLF: $\leq$0.04 Hz) were not included since short-term HRV analysis was used [10]. The HRV metrics are presented in Tab. I. The raw IBI series was detrended using the method of smoothness priors [11], using a value of 500 for α. The average IBI was calculated from the non-detrended time series. The frequency domain metrics were obtained as the discrete Fourier transform (DFT) of the 4 Hz cubic-spline interpolated IBI series, using a Hanning window. The spectral values LF_{abs} and HF_{abs} were divided by the square of the mean IBI length in the analysis segment, to make them statistically independent of the mean IBI [12]. The power spectral values become unitless in the process and can be labelled squared modulation index per Hz [13].

1) Baseline Correction: The calculated HRV metrics were baseline corrected to compensate for the interindividual variation of the physiologic data. The baseline correction was performed individually for every HRV metric for each subject. The average value of the combined data in the D1 and D2 conditions (medium MWL) was subtracted from all data points.

2) Pooling of Data: After baseline correction, the HRV metrics show the relative difference for each subject from the medium MWL condition, allowing the data points from all subjects from task blocks S1, S2, M1 and M2 to be pooled for group-level comparisons.

3) ROC Analysis: The ROC analysis was performed in R [14] using the ROCR-package [15]. Classifier performance was evaluated using the AUC. Of two classifiers, the classifier with greater AUC is considered to have better average performance [16]. A classifier with an $AUC_{orig} < 0.5$ can be negated to produce a classifier with an AUC of $1 - AUC_{orig}$.

III. RESULTS

A. Existence of MWL

The P300 ERP component amplitude calculated from the Pz EEG channel was used to verify the existence of MWL in the different task blocks. The ERPs were calculated from correctly acknowledged target tones in the auditory task.

The mean and 95% confidence intervals for the P300 amplitudes are shown in Fig. 1. The P300 amplitude decreased with an increasing number of simultaneous tasks. This confirmed, that task blocks M1 and M2 indeed induced a greater MWL than task blocks S1 and S2.

B. AUC-distributions

Distributions of AUC-values calculated from the individual data are shown in Fig. 2. The median of AUC values of average IBI was very high (close to 1) and the weight of the distribution was located above 0.5. The metrics $LF_{\%}$, LF_{nu}, $^{LF}/_{HF}$ and HF_{nu} had similar medians and distributions, with LF_{nu} having a more narrow distribution. The other HRV metrics did not perform well on the individual level.

C. Group-Level Results

The AUC-values calculated from the pooled data are shown in Tab. II. Average IBI showed the highest AUC for the pooled data with good inter-subject consistency. Also $LF_{\%}$ showed a pooled AUC above 0.7 with good consistency.

D. Individual Results

The number of subjects with AUC<0.3, 0.3$\leq$AUC$\leq$ 0.7 and AUC>0.7 for the different HRV metrics is shown in Tab. II as a measure of the consistency of the HRV metrics. For both average IBI and $LF_{\%}$ some subjects showed opposite results. This is, for average IBI, seen as some subjects having individual AUC values below 0.3 and some subjects having AUC values above 0.7. The classification performance was good in both cases, but the underlying physiologic reactions were different.

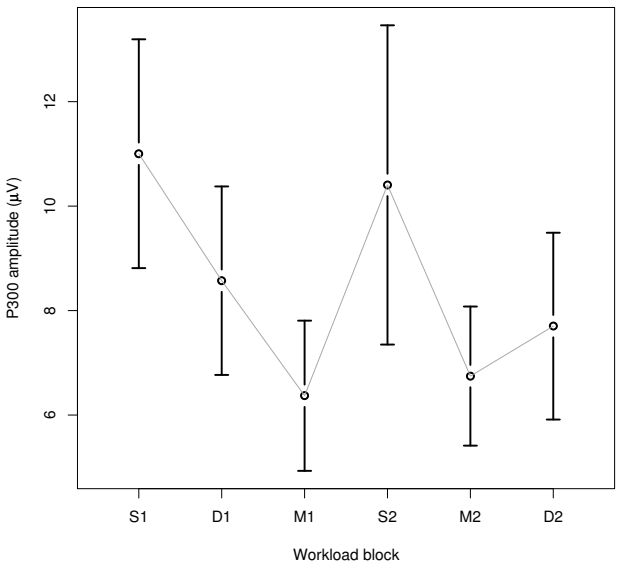


Fig. 1. The mean and 95% confidence intervals for the amplitude of P300 ERP component of the Pz EEG channel in the six blocks of the computerised multitask test.

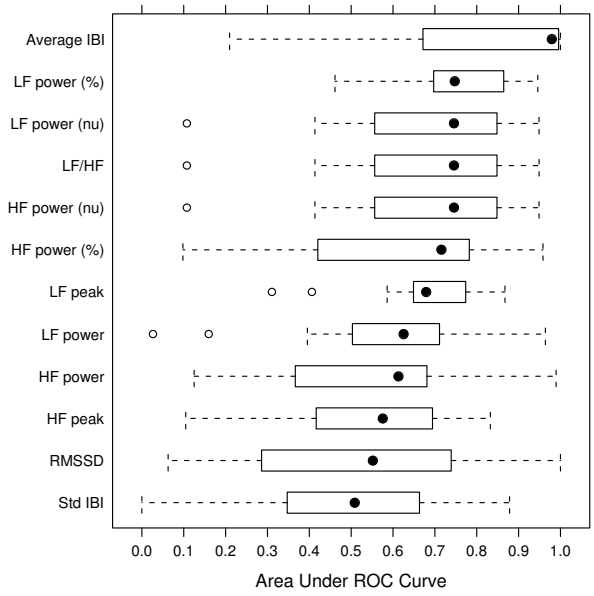


Fig. 2. Distributions of AUC-values calculated from individual data. The black dot marks the median. The box spans the first to the third quartile. Outliers (over 1.5 times the interquartile range) are marked with circles. Whiskers extend to the minimum and maximum non-outlier values.

E. ROC Curves

The ROC-curve for the four HRV metrics with the greatest pooled AUC values are shown in Fig. 3. Average IBI and $LF_{\%}$ reached their best classification performance in the upper right region of the ROC space where both classifiers had a high true positive rate (TPR) but also a high false positive rate (FPR). Average IBI also dominated all of the other HRV metrics in the ROC space. The distributions of the other classifiers presented were more overlapping which lead to degraded classification performance.

TABLE II

AUC VALUES CALCULATED FROM POOLED DATA, AND THE NUMBER OF SUBJECTS WITH AN AUC <0.3, $0.3\leq$ AUC ≥ 0.7 AND AUC > 0.7, CALCULATED FROM THE INDIVIDUAL DATA.

HRV Metric	**AUC**	**N<0.3**	**0.3≤N≤ 0.7**	**N>0.7**
Average IBI	0.80	9	1	3
LF power (%)	0.73	0	4	9
LF peak	0.67	0	8	5
LF power (nu)	0.66	1	4	8
HF power (nu)	0.66	8	4	1
LF/HF	0.63	1	4	8
HF power (%)	0.62	8	4	1
HF peak	0.58	3	8	2
LF power	0.57	2	7	4
HF power	0.55	3	7	3
Std IBI	0.52	2	8	3
RMSSD	0.51	4	4	5

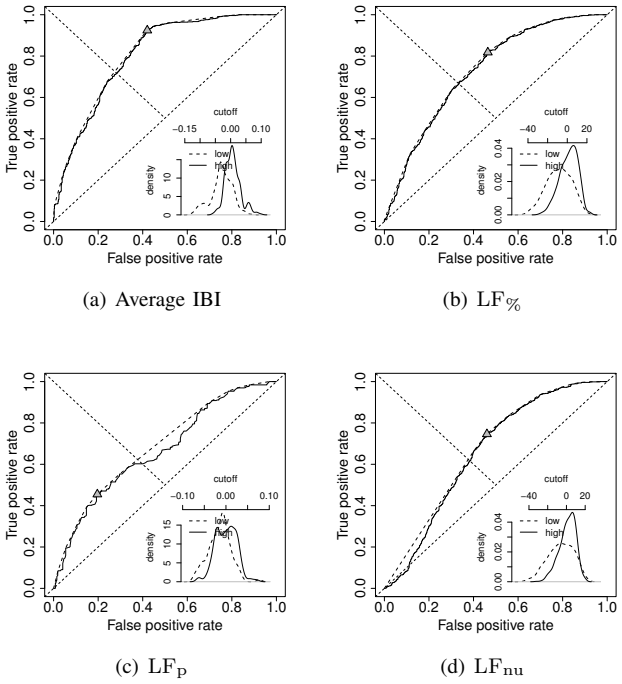
IV. DISCUSSION

The ERPs showed, that a difference of MWL existed between the different task blocks in the multitask test. This MWL was also reflected in the cardiac activity of the subjects, allowing the physiologic responses in the different task blocks to be compared.

Since HRV metrics with an AUC$<$0.7 cannot be considered usable for the purpose of classifying the degree of MWL, the data suggests that only the HRV metrics of average IBI and $LF_{\%}$ are suitable for MWL classification.

The average IBI showed the best classification performance, but it is possible that this metric only distinguishes situations with pronounced differences in task difficulty. Other metrics could be more useful in the classification of MWL in situations with similar MWL levels, but with tasks involving different types of mental processing. The fact that the average IBI reached a higher classification performance than the other HRV metrics does not mean that the other metrics were not reactive to changes in MWL; the effects were only more visible and consistent for the average IBI than for the other metrics.

The average IBI HRV metric outperformed all of the other HRV metrics considered in this study in terms of MWL classification ability. Since the average IBI is a liberal classifier, the trade-off to the high classification performance is an increased FPR. Despite the drawback of a high FPR, a liberal classifier can be useful in situations where the HRV metric is used as a part of a pre-alarm system. Such a system could warn for instance operators in safety-critical occupations of increased MWL leading to degraded performance.



(a) Average IBI (b) $\mathrm{LF}_{\%}$

(c) LF_{p} (d) $\mathrm{LF}_{\mathrm{nu}}$

Fig. 3. ROC curves for the four best-performing HRV metrics calculated from pooled data. True positive rate indicates the fraction of correctly classified high MWL at given threshold. The ROC convex hull has been drawn with a dotted line and a triangle marks the point of maximum accuracy. Density plots of the pooled data are shown in the insets, for low MWL (dotted line) and high MWL (solid line).

The use of short-term HRV measurements allows the degree of MWL to be estimated with good temporal accuracy.

The experimental procedure could be improved by individualising the multitask workload levels for every subject, e.g. by varying the ISI in the tasks, as described by [6]. A limitation of this study is the small sample group. A larger sample would permit an estimation of the variance in the ROC analysis. The number of blocks in the multitask test could be increased to allow for more transitions between the different task levels. Blocks of longer duration would allow more data from each MWL level to be collected, reducing the possibility of bad data corrupting an entire block.

In the future, it would be interesting to study the repeatability of the measurements. Repeated measurements performed with the same subjects on different days would provide information on the individual variation in the reactivity to the computerised multitask test. This would also show how consistent the responses are. The underlying reasons for the inconsistent reactions exhibited by some of the HRV metrics in this study, including average IBI, is an issue that should be further investigated. The subjects' performance in the multitask test could also be used to gain insight into these paradoxical reactions.

Instead of only using the two generic levels of "low" and "high" MWL in the ROC analysis, the individual subjects' P300 amplitudes could be used in the classification of MWL. This would allow a more precise classification of MWL.

V. Conclusion

The use of ERPs in the external verification of MWL followed by ROC analysis allowed the classification performance of the HRV metrics to be compared.

Based on the results of this study, it can be concluded, that the short-term HRV metrics of average IBI was the best HRV metric in terms of classification ability using short-term measurements. This HRV metric can hence be used to distinguish between a low and high MWL condition during a computerised multitask test.

References

[1] L. Ding-Yu and H. Sheue-Ling, "The Development of Mental Workload Measurement in Flexible Manufacturing Systems," *Human Factors and Ergonomics in Manufacturing*, vol. 8, no. 1, pp. 41–62, 1998.

[2] G. Mulder, L. J. M. Mulder, T. F. Meijman, B. P. Veldman, Johannes, and A. M. van Roon, "A psychophysiological approach to working conditions," in *Engineering Psychophysiology: Issues and Applications*, R. W. Backs and W. Boucsein, Eds. Lawrence Erlbaum Associates, Inc., 2000, ch. 6, pp. 139–159.

[3] J. Aasman, G. Mulder, and L. J. Mulder, "Operator effort and the measurement of heart-rate variability." *Hum Factors*, vol. 29, no. 2, pp. 161–170, 1987.

[4] J. J. B. Allen, A. S. Chambers, and D. N. Towers, "The many metrics of cardiac chronotropy: a pragmatic primer and a brief comparison of metrics." *Biol Psychol*, vol. 74, no. 2, pp. 243–262, Feb 2007.

[5] G. G. Berntson, J. T. J. Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, and M. W. van der Molen, "Heart rate variability: origins, methods, and interpretive caveats." *Psychophysiology*, vol. 34, no. 6, pp. 623–648, 1997.

[6] M. Sallinen, A. Holm, J. Hiltunen, K. Hirvonen, M. Härmä, J. Koskelo, M. Letonsaari, R. Luukkonen, J. Virkkala, and K. Müller, "Recovery of cognitive performance from sleep debt: do a short rest pause and a single recovery night help?" *Chronobiol Int*, vol. 25, no. 2, pp. 279–296, Apr 2008.

[7] E. L. Reilly, *Electroencephalography: basic principles, clinical applications, and related fields*, 5th ed. Philadelphia: Lippincott Williams & Wilkins, 2005, ch. EEG Recording and Operation of the Apparatus.

[8] A. Kok, "On the utility of P3 amplitude as a measure of processing capacity," *Psychophysiology*, vol. 38, no. 3, pp. 557–577, 2001.

[9] J. Niskanen, M. Tarvainen, P. Ranta-aho, and P. Karjalainen, "Software for advanced HRV analysis," *Computer methods and programs in biomedicine*, vol. 76, no. 1, pp. 73–81, 2004.

[10] M. Malik, J. Bigger, A. Camm, R. Kleiger, A. Malliani, A. Moss, and P. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354–381, 1996.

[11] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to HRV analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[12] J. A. Veltman and A. W. Gaillard, "Physiological indices of workload in a simulated flight task." *Biol Psychol*, vol. 42, no. 3, pp. 323–342, 1996.

[13] L. J. M. Mulder, "Assessment of Cardiovascular Reactivity by means of Spectral Analysis," Ph.D. dissertation, Rijksuniversitet Groningen, Groningen, The Netherlands, July 1988.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[15] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r." *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[16] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.