Temporal Resolution Capabilities of the Mid-Frequency Heart Rate Variability-based Human-Computer Interaction Evaluation Method

Károly Hercegfi, Máté Köles

Department of Ergonomics and Psychology, Budapest University of Technology and Economics, Magyar tudósok körútja 2, 1117 Budapest, Hungary, hercegfi@erg.bme.hu, kolesm@erg.bme.hu

Abstract: The mid-frequency component of Heart Rate Variability (HRV) is utilized in many studies to measure the level of mental effort in Human-Computer Interaction (HCI). However, the temporal resolution that can be achieved using this method is underestimated. For refining the specification of the exact temporal resolution of this method, we employed a visual search task that required elevated levels of mental effort. Participants had to find one difference between pairs of pictures. Each of the twelve pairs was followed by a congratulation screen causing a short period of relief (5-6 seconds). Using our method based on power spectra analysis and windowing functions, we were able to differentiate between the HPV mid-frequency values of the visual search and the relief periods. These results, along with previous findings, seem to suggest that the temporal resolution of 5-6 seconds can be achieved with our method, widening the range of applications.

Keywords: human-computer interaction; empirical usability evaluation methods; ECG; heart rate variability; heart period variability; mental effort

1 Introduction

1.1 The Role of Mental Effort in Human-Computer Interaction

Mental workload required by Human-Computer Interaction (HCI) (or the self-imposed part of this mental workload, the mental effort invested by the user), as a measurement of “ease of use” is a key factor of usability or pragmatic aspect of user experience (UX) [1], [2]. The less mental effort one needs to operate a given software, the better it is in terms of usability. This approach is important in traditional HCI researches and in related research areas as ergonomics, UX, and Cognitive Infocommunications (CogInfoCom) as well [1]–[6]. We emphasize that mental effort can be an objective basis of the usability evaluation, and, there are practically applicable methods, such as the Heart Rate Variability based method described below. In this paper, new results are published on the temporal resolution capability of his method: The results raise hope that this method – over the previously developed HRV-based methods – is capable of exploring practical usability issues identifying quality attributes of software elements with a temporal resolution of only a few seconds: this time window can be narrow enough to analyse the mental effort caused by such short interactions as some clicks in a menu, or reading a message and pushing a button, etc.

In concordance with this, an everyday HCI is rarely characterized by sustained mental effort throughout the whole session. Instead, it most likely includes longer periods of relatively low mental effort with brief events of higher demand. Or vice versa, in more critical situations – such as air traffic management [7] or reading e-learning material for an exam, etc. –, it includes longer periods of relatively high mental effort with brief events of lower demand (e.g., reliefs). These events are very important from the usability standpoint. Sudden increases in mental effort demand can be caused by the users' previous experience (e.g., incompatible mental models) or a stage in the interaction where the mobilization of mental effort is unavoidable. However, they can also indicate a flaw in design that puts unnecessary burdens on the users. This extra load can lead to a set of errors [8], frustration, or higher levels of fatigue [9].

1.2 Measuring Mental Effort

There are many methods available for measuring mental effort (self-imposed mental work stress). Task performance on primary or secondary tasks are often used in ergonomics [10]. Subjective rating scales, such as the NASA Task Load Index (NASA-TLX) [11], are also widely used. They can even be used to differentiate between factors influencing mental effort, such as time pressure and frustration. Subjective ratings, however, have their limits. If a user is asked to recount a longer session, there will be events that will be forgotten or remembered differently than as it has actually happened. Simple observation techniques can prove invaluable in supporting other methods. Video recordings of facial expressions, body movements or postures can help disambiguate findings or highlight previously unnoticed periods of interest. The analysis of facial expressions using well defined coding schemes (e.g., [12]) can support decisions about the cause of mental effort change. However, their temporal resolution is not ideal for following the constantly changing states during HCI.

There are also a wide palette of psychophysiology-based methods that are capable of measuring mental effort. Change in electrodermal activity (EDA) [13], facial electromyography (EMG) [14], blink rate [15], pupillometry [15]–[19], visual Critical Flicker Frequency (CFF) [20], or even salivary cortisol levels [20], [21] are capable of identifying changes in mental effort. The advantage of these methods is that they do not require any recollection from the user, to uncover potential trouble spots in HCI.

However, most physiology-based methods are not selective enough in their output to be capable of measuring mental effort on their own. For example, pupil size is influenced by almost every external or internal event [19]. The EDA reacts much more profoundly to affective effects than to mental effort. To study emotions in HCI, our department hasexperiences with measuring EDA (Skin Conductance – SC) [22], [23].

CFF and measuring the cortisol have also been applied by colleagues [20], but they give an indicator for a relatively long period of several minutes to hours.

Applying pupillometry – among the mentioned problems – is a promising method [19]. Eye-tracking is also promising not only for detecting the focus of the user during effort required events, but its metrics can reflect to the mental state [24]. Even intraocular pressure changes can be used to identify cumulative or instantaneous changes in mental effort [25].

Electroencephalography (EEG) can also be used to measure mental effort in HCI research [26], [27]. Its temporal resolution is superb or on par with the previously mentioned methods. It also has the advantage of being a direct measure of central nervous system activity while others are indirect. To measure mental effort, spectral parameters of certain frequency bands are used. For example, the ratio of the beta (~12-30 Hz) and alpha frequency bands (~8-12 Hz) can be used as an index of mental effort [28]. Others use different frequency bands [29] or ERP based approaches [30], but all seem to be promising in measuring mental effort.

It is important to note, that to uncover the *cause* of those events, a single method will probably never be enough. To date no physiology-based method is able to completely eliminate other supporting techniques like observations, interviews, or retrospective think aloud protocol.

1.3 ECG in Measuring Mental Effort

Heart rate is the number of heart beats in each time interval. Heart rate usually increases during a mental effort demanding task, and the magnitude of its change can be informative to some degree [31]. However, there are more sensitive measures available when we are interested in the changes of mental effort.

The variance or standard deviation of heart rate can also be used; however, these measures also contain influences from various physiological sources independent of mental effort. The spectral analysis of Heart Rate Variability (HRV) (or its reciprocal expression, Heart Period Variability (HPV), where the power spectral density estimation is based on Interbeat Intervals (IBI)) can be used to minimize effects from other sources. The most frequently used IBI is based on the component of the electrocardiogram (ECG) recording with the biggest amplitude, the so-called R peaks. These time periods can be referred to as RR intervals. The power spectral density estimation is either based on RR intervals, or HR values.

The most important frequency band, in case of mental effort, is the so-called Mid-Frequency (MF) peak between 0.07-0.15 Hz. A number of studies [32]–[36] reported lower power in this frequency band during mental effort. Both sympathetic and parasympathetic activity is believed influence this component [37]. High peak in the Mid-Frequency Power (MFP) band may also be caused by movements (as the baroreflex controls the blood pressure). To separate the effect of the mental effort from the effect of baroreflex, a ratio of the MF component and the below mentioned higher frequency respiratory component can be applied [38]. However, in case of HCI, users typically sit continuously, and their larger muscle movements (e.g., stretching) eventually can be filtered from the records via video analysis. Furthermore, practically, some significant movements seem not to affect the indication of mental effort [39]. Therefore, the mental effort can be characterized sensitively enough by the MFP band itself, as it is shown by the current results presented in this paper.

The high frequency band (0.15-0.45 Hz) represents respiratory function through the so-called respiratory sinus arrhythmia. It is influenced by parasympathetic activity. The power of low-frequency band between 0.04-0.07 Hz is related to thermoregulatory fluctuations of the blood vessels [40], [41]. An ultra-low frequency band with a range of $1.15*10^{-5}$-0.00335 Hz can also be defined and is believed to reflect circadian variation [42], [43].

There are many ways to calculate these spectral frequency measures. There are non-parametric methods based on Fast Fourier Transformation (FFT). Their advantages are their ease of computing and low processing requirements. The FFT-based power spectral density estimation is derived from all the data present in the recording [44]. This means that FFT is computed using the whole variance of frequency components regardless of them being at certain frequency peaks or not. It also requires a longer recording to achieve its best spectral resolution. An alternative would be parametric methods based on Autoregressive (AR) modeling. These methods produce smoother spectral components and give more precise power spectral density measures in case of shorter recordings. This property makes them a better candidate for use in HCI research, where often shorter time periods are more informative that the whole session. The AR models only use specified band powers for their estimation; other components are discarded as noise. The method described in this paper is based on a special application of AR spectral power computation. We discuss the analysis in Section 2.4.

2 Methods

2.1 The INTERFACE Methodology

The present study applied the INTERFACE (INTegrated Evaluation and Research Facilities for Assessing Computer-users' Efficiency) software evaluation methodology, developed by Izsó and his colleagues at the Budapest University of Technology and Economics (BME) [36], [40]. The strength of the methodology lies in recording (and later replaying and analysing) multiple channels simultaneously. The default setup of the INTERFACE workstation records various aspects of HCI. First of all, key presses, mouse clicks and, sometime, other events of the HCI are recorded as well as the content of the screen. Optimally two cameras film the participant. One focusing on the face and the other on the whole body. The earlier is necessary for the observation of facial expressions. The latter is mainly useful to spot any major movement or changes in posture that could have influenced the physiology channels. However, postures and gestures can also show the users' metal state similarly as the facial expressions do. These recordings help clarify ambiguous periods observable in the physiology channels recordable with the current setup – in this paper, namely the MFP of the HPV.

The recording of the physiological data was accomplished with the ISAX module (Integrated System for Ambulatory Cardio-respiratory data acquisition and Spectral analysis). It is a specific hardware and software solution developed by the Psychophysiology Research Group of Hungarian Academy of Sciences and the BME [40] for easy and portable physiological measurement.

2.2 Experimental Setup and Procedure

2.2.1 Participants

All 11 participants (8 female) were graduate or undergraduate students at BME, with a mean age of 21.9 and the minimum of 19 and maximum of 26. They all had normal or corrected to normal vision, and reported no cases of previous cardiovascular surgeries or diseases. They were instructed to refrain from the consumption of any stimulants (coffee, cigarettes, energy drinks, etc.) for at least 2 hours before the experiment.

2.2.2 Electrode Placement

A bipolar lead was used for ECG measurement. The setup of the two main electrodes close to the electrical axis of the heart is found most suited to maximize the amplitude of the R peaks. The exploring, or positive electrode was placed on the sixth or seventh rib, below the left nipple. The indifferent, or negative electrode was placed high up on the right side of the sternum, i.e. on the right side of the manubrium of the sternum, close to the right clavicle, or in the left side of the right infraclavicular fossa. The ground electrode was located on the seventh or eighth rib on the left median auxiliary line (see Figure 1). Depending on the real electrical axis of the heart, the texture of the tissues, and the build of the person, other ECG electrode placements can also be chosen to maximize the magnitude of the R wave and minimize the artefacts caused by movements.

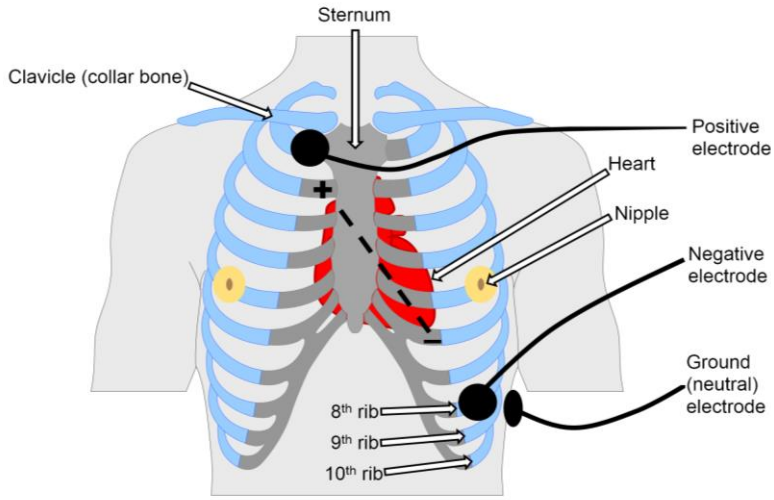

Figure 1
The electrode placement used for our experiment

We used Type 2228, Ag/AgCl electrodes manufactured by 3M. The skin was cleaned using alcohol before electrode placement. The ECG data was recorded by the ISAX module. It was connected to a laptop running from battery for safety measures.

2.2.3 Additional Software Used

The video capture of the screen content with the eye-gaze paths was realized by Tobii Studio v.2.1.14 software – analysing the eye tracking data gained by the Tobii T120 equipment can be subject of further analysis in another paper. Virtual Dub software recorded the view of an external camera connected to a second computer. Additional software for recording event logs, experimenter’s notation, and synchronizing the records via serial wires and special button as parts of the INTERFACE frame system developed by our team. For statistical analyses, IBM SPSS Statistics 22 was used.

2.3 Procedure

All participants were briefed before electrode placement. They were asked to abstain from any major movements during the whole session. Speech is also known to influence HRV, so we asked them to speak during recordings only if they had some trouble with the tasks.

At the beginning of the session, all participants were asked to sit as comfortable as they could and relax for four minutes. The instructions of these periods were to seat themselves in a comfortable posture, without any movement while keeping their eyes open. Also, trying not to think about anything in particular, in spite of the known fact that it is not trivial for people untrained at this (at least trying to avoid thinking specific things), and ensuring them that there are no good or bad personal results: we have no expectations, we would only like to investigate some differences between this period and the next one.

The following period consisted of a mental arithmetic task. Participants were told that a number will appear in the middle of the screen and they will have to count backwards by seven silently from this number until a question mark appears on the screen. We instructed them before the task that they should avoid movements and also avoid speech, counting aloud, or voiceless movement of mouth. In case they lost track, they should continue from any number they seem to remember (to ensure that the level of mental effort is kept up during the whole period).

After 20 seconds of waiting, the number of 11558 appeared for five seconds. Two minutes later a question mark appeared on the screen. Then the participants had to speak out loudly the number they reached. They were given positive feedback on their performance.

After a short break, participants were presented a visual search task. Twelve pair of pictures were displayed. The participants had to find the difference between one pair of pictures at a time and click on it with the left mouse button to proceed. Clicking anywhere else caused no effect. The pictures were created applying twelve holiday photos on various topics with various atmosphere, taken by Károly Hercegfi. Each stimulus contained duplicated pictures with only one difference between them. To create the differences, the pictures were edited using Adobe Photoshop (see Figure 2). It was either a missing or extra object or the change of colour of an object. If a participant had not found the difference within three minutes, they were given clues verbally by the instructor, to avoid the building up of frustration. The order of the pictures was fixed. They were either aligned left and right or top and bottom according to the original format of the picture.

Once the difference was found and clicked upon, a congratulation screen appeared for 5-6 seconds (meantime: 5.4 s), then the next pair was loaded. The reason for the variability in duration comes from the JavaScript animated HTML design we used for this experiment. The pictures themselves were not stored on the hard drive of the computer running the experiment, but on a server. Because of this, there was always a delay in loading the next set of pictures, however, the duration of the congratulation screen never exceeded six seconds.

We expect the MFP of HPV values to be significantly lower during relaxation compared to mental arithmetic. This is done to illustrate that our method of calculation is able to separate high and low levels of mental effort using only the MF band.

The main goal of this study is to examine the MFP of HPV value differences between a mental effort demanding task and a short relief period immediately after it. We expect to find significantly higher MFP of HPV values during the Congratulation screens opposed to the Visual search task. If we would find such a relationship, it would mean that the temporal resolution below 6 seconds is possible to achieve.

Figure 2
An example of the visual search task stimuli

2.4 HPV Analysis

For our goal to analyze HCI events, we need the MFP of HPV values as a quasi-continuous function of time. Such a curve would make spotting changes in invested mental effort more convenient. For the estimation of power spectral density, we are using all-pole autoregressive (AR) modeling. To create the MFP profile curve, a few transformations have to be made. The main steps of our analysis are shown by Figures 3a and 3b.

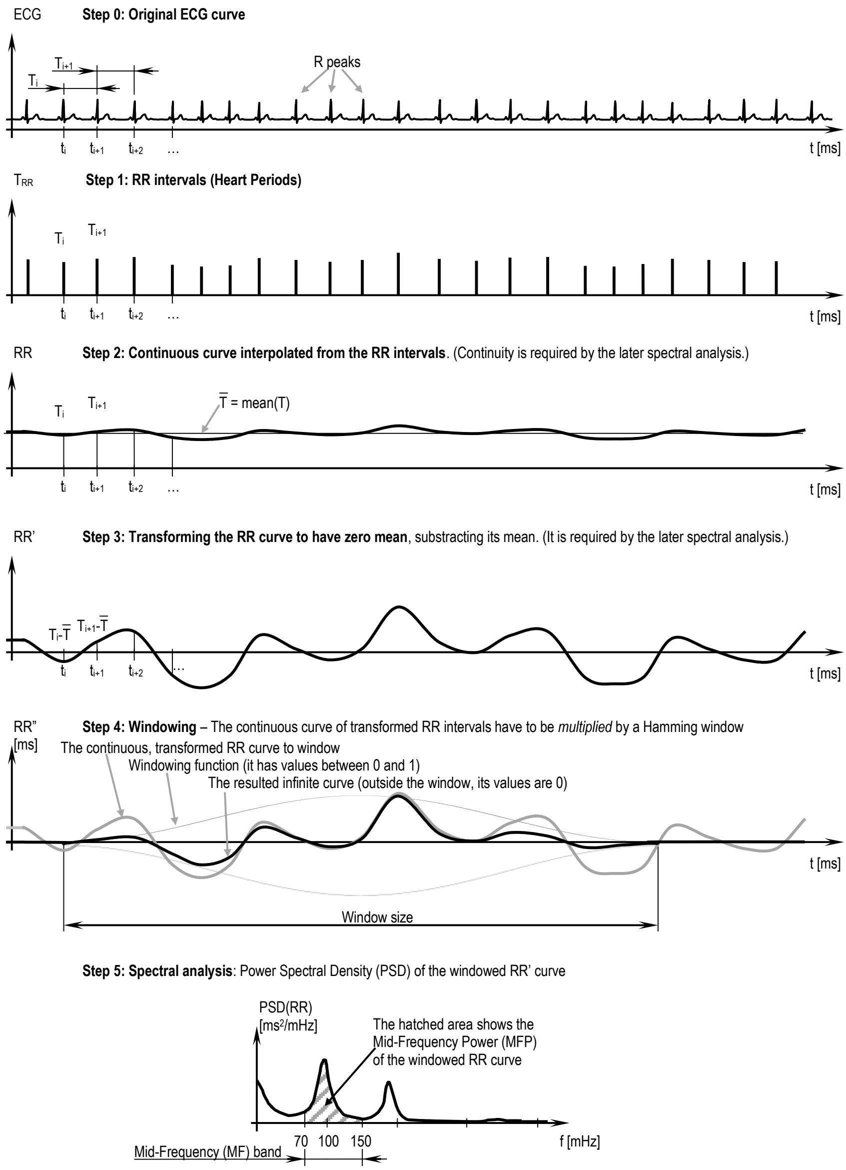

Figure 3a

Calculation of the Mid-Frequency Power (MFP) of the Heart Period Variability (HPV) for a particular time-window

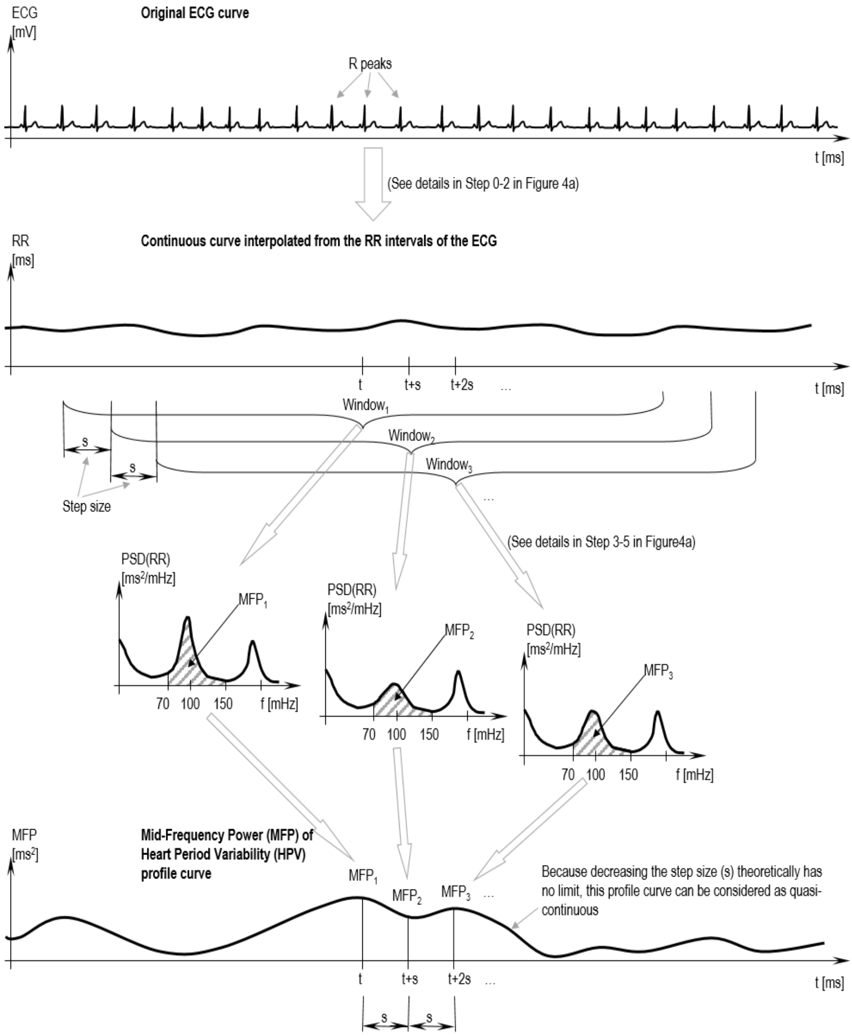

Figure 3b
Creation of the Mid-Frequency Power (MFP) of the Heart Period Variability (HPV) profile curve applying windowing technique

The first step is to identify the R peaks, and calculate the RR interval times. A continuous time signal is created by using a linear interpolation. Creating a time series applying equidistant sampling of the RR interval function is required by the later spectral analysis, and is performed at 1 Hz. Prior to the AR model fitting, another requirement has to be fulfilled; the signal has to have a mean of zero. In order to do this, we subtract the mean of the whole series from each value. The signal is now prepared for AR model fitting.

In order to create a MFP profile curve, we estimate the power spectral density for highly overlapping 32 second wide frames (shifted repeatedly by 1 second). To avoid side-effects caused by truncation of the time series, we use the Hamming windowing procedure. The power spectral density estimation is carried out using a modified Burg's algorithm and Akaike's Final Prediction Error criteria [45] is used for model selection.

The summary of the parameters used:

- MF band: 0.07-0.15 Hz
- Maximum model order number = 12
- Window length = 32 sec
- Step size = 1 sec

3 Results

The normal range of HRV shows great variety in the normal population. In order to make our data comparable between subjects we normalized the available HPV data. To do this, we calculated an average HPV value for every subject using the data recorded during the visual search task and the congratulation screens. We divided the original values with this average, thus we received a new set, where 1 could be viewed as 100%, 1.6 as 160% compared to the participants personal average, and so on. All of the following analyses will be conducted with these normalized values. Due to the small sample size, we used the nonparametric Wilcoxon Signed Ranks Tests for most of our analysis. Pearson's r values are also provided for measures of effect size.

There was a significant difference between the relaxation (mean = 1.84, SD = 1.32) and the mental arithmetic (mean = 0.87, SD=0.76) periods ($z = -2.09$; $p = 0.0185$ (one-tails), $r = -0,63$, Figure 4). This difference was even visually evident in most cases, based on the MFP of HPV profile curve (Figure 5). A perfect relaxation would provide a curve that is relatively high the whole time, but given the circumstances the participants were in, it was not expected; some can easily relax in an experimental setup, some was disturbed by the wiring and observation, and the subjects were not trained to use advanced relaxation techniques.

The difference between the MFP of HPV values of the visual search task (mean = 1.06, SD = 0.23) and the congratulation screen (mean = 1.47, SD = 0.44) was significant ($z = -2.223$, $p = 0.013$ (one-tailes), $r = -0.67$, Figure 6). This means that, using this method, we were able to differentiate between periods thought to invoke higher and lower levels of mental effort. Hereinafter, we will present additional data, to support this claim.

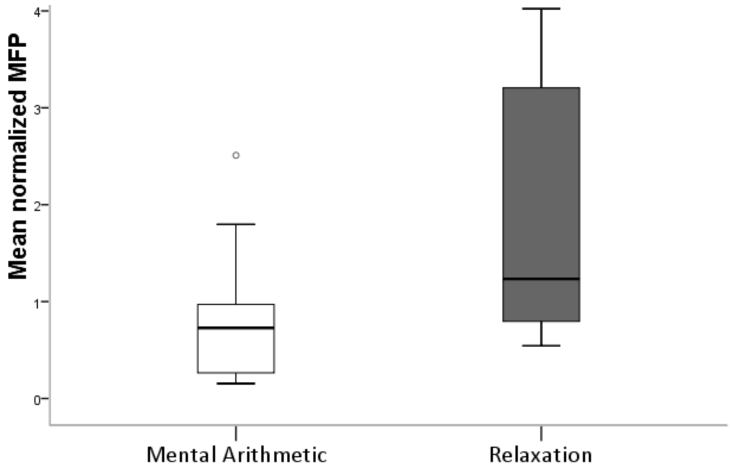

Figure 4

Boxplots of normalized MFP of HPV values for the relaxation and mental arithmetic periods. According to the defaults of the SPSS software, the T-bars (also called inner fences or whiskers) extend to 1.5 times the height of the box, or, if no case/row has a value in that range, to the minimum or maximum values. The circle represent an outlier (value that does not fall in the inner fences).

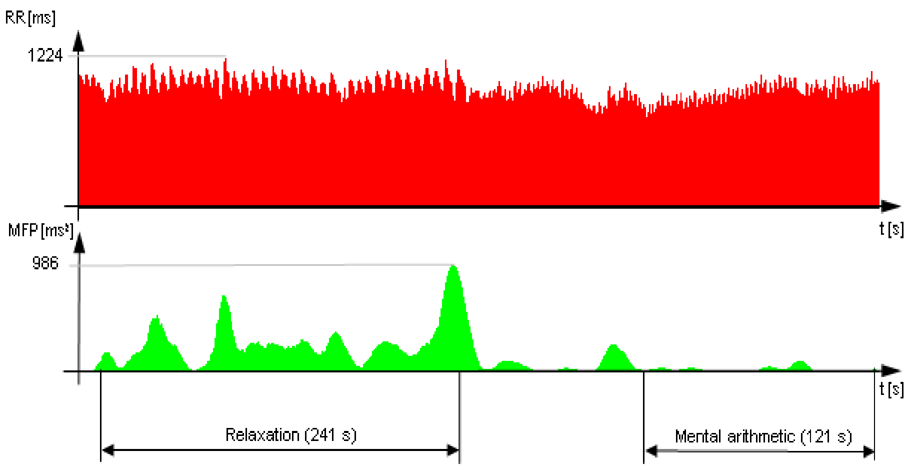

Figure 5

The difference between relaxation and mental arithmetic is clearly observable on the MFP of HPV profile curve of subject #5 in visualization style of the INTERFACE Viewer software. The upper (red) curve displays the RR intervals; the bottom (green) one represents the Mid-Frequency Power (MFP) profile curve of the Heart Period Variability (HPV). While the participant is relaxing, the MFP profile curve has much higher values and the RR curve has big zigzags as opposed to the mental arithmetic phase, where the MFP is low and the RR curve smoothens out

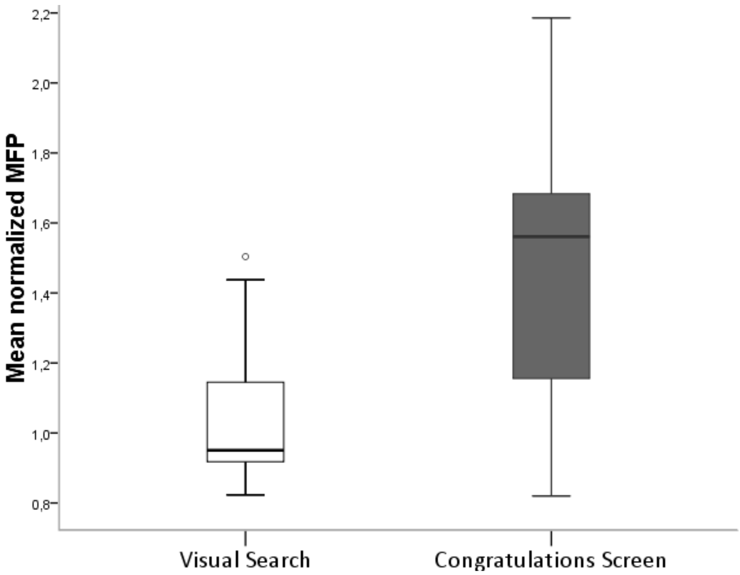

Figure 6

Boxplots of normalized MFP of HPV values for the visual search task and the congratulation screen. (T-bars extend to 1.5 times the height of the box, or, if no case/row has a value in that range, to the minimum or maximum values. The circle represents an outlier.)

Figure 7 shows the average values by pictures and the related congratulation screens. As it can be seen, the MFP of HPV values are always higher except for the first picture. This can be attributed to the novelty of the task, as there was no practice set before it. The fact that the aggregated values of the participants show this kind of consistency, supports our claims. We do not know of any other methods to date that are able to identify such short periods of change in mental effort levels.

This consistency is also observable if we take a look at the values of our participants separately. As it can be seen in Figure 8, the MFP of HPV values were much higher during the congratulations screens in most cases. Only three participants showed a different pattern. However, these differences are minuscule compared to others with the expected pattern.

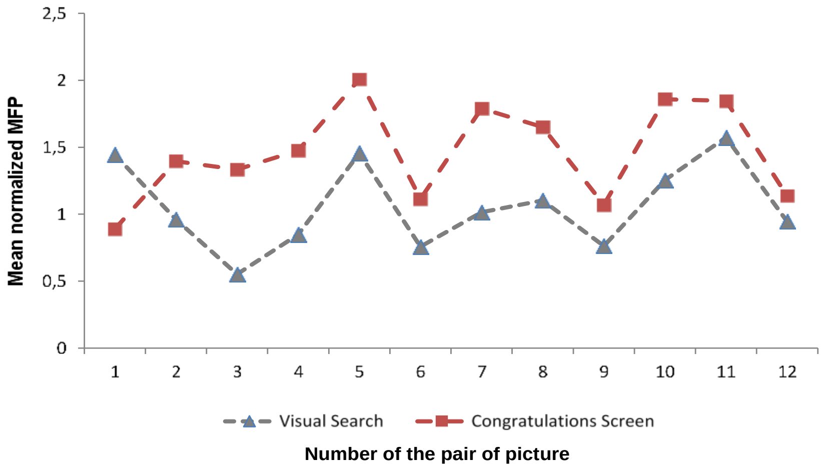

Figure 7
The average MFP of HPV values for each pair of pictures and the related congratulating screen. Only the first picture showed a pattern, where the average values were higher during the visual search task

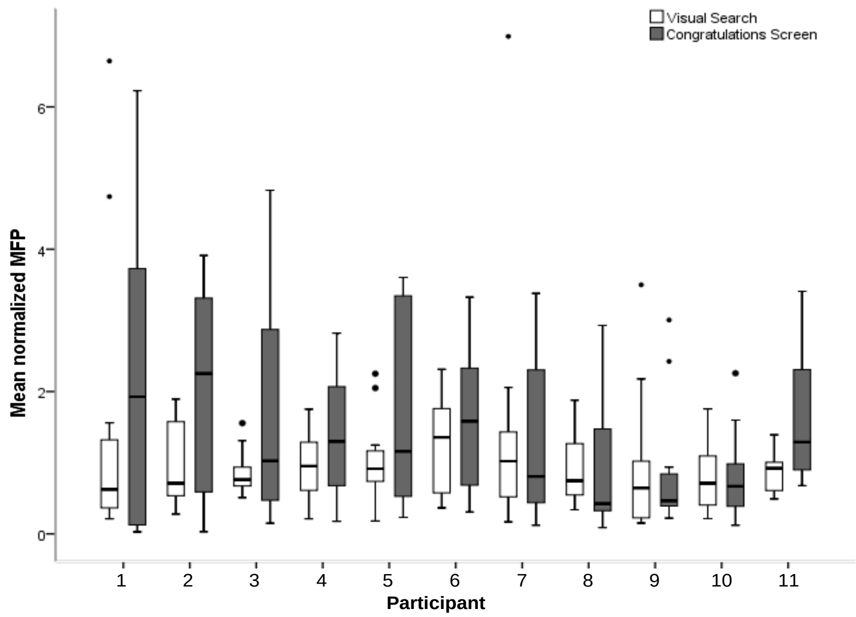

Figure 8
The mean normalized values of all visual search and relief periods by participants. Blank boxplots represent visual search MFP of HPV values; patterned boxplots represent the same for congratulation screens. (T-bars extend to 1.5 times the height of the box, or, if no case/row has a value in that range, to the minimum or maximum values. Circles represent outliers.)

4 Discussion

Our aim was to further explore the temporal resolution capabilities of our method based on ECG data. A visual search task was applied that required longer periods of sustained mental effort, followed by short intervals of relief. We found significant difference in the MF power of the HPV between these two periods. Post hoc effect size measures were also classifiable as large, based on Cohen's (1992) criteria [46]. Taken together with previous results [20], [36], [40], our method can be considered capable of identifying problematic events in HCI that are no longer than few seconds. The profile curve created is also a useful tool in itself because it allows for the identification of periods of interest by simple visual inspection. In our opinion, this makes our AR model based approach easy-to-apply, quick and informative usability evaluation tool of Human-Computer Interaction.

However, as we stressed in Chapter 1.2, a single method is not sufficient to get a complete picture of an interaction. Some increases in mental effort are unavoidable, normal, or even beneficial. If the goal is to improve the rate at which users retain knowledge regarding the user interface, effortful recall is favourable [47], [48]. To identify the cause of each change in mental effort levels, other supporting methods such as video based observations, interviews, or retrospective think aloud protocol must also be applied.

As mentioned earlier, EEG based methods promise high temporal resolution with great differentiating ability between different levels of mental effort [49]. In the near future, it might prove to be the best method of measurement. However, it also has some issues that have to be overcome first. To get a good estimate of power spectral parameters, noise levels should be kept at minimum. There are great filters available to identify the effects of eye-blinks on a recording [50]. It is noteworthy that in a more natural HCI setup (e.g., no head rest) other muscle activity can influence quality. Filtering out these effects requires the use of EMG which makes the experiment more complex and less natural for the participant. Even cardiovascular activity introduces noise into the EEG recording [51]. Independent Component Analysis is often used to remove these artefacts, but the component selection is often based on subjective judgement of the person conducting the analysis and not on specific rules. The abundance of different approaches that all seem to measure mental effort very well is also peculiar in light of significant individual differences in reactions to the increases of mental load [52]. In case of our ECG-based method, movement artefacts are also an issue, but are more easily avoided, and electrode placement leaves more room for error. It is also faster to set up than even a cheaper EEG cap (reliable dry electrode EEG systems might change that). In conclusion, we feel that at present, the ECG-based approach is more reliable than the EEG.

Apart from usability testing, our method could be applied to other fields as well. For example, dynamic difficulty adjustment in games based on physiology is an upcoming trend [53], [54]. If our aim is to create an experience that is engaging to the player, maybe even eliciting Flow [55], an index of mental effort can prove to be useful.

We are aware that in our present study the periods of relief were short, not the mental effort. However, by being able to differentiate between the two, we have shown that the AR based method is capable of a relatively high temporal resolution. Our next project should aim at a more natural HCI setting, where longer periods of low mental effort are interrupted by short, but more demanding "trouble spots". This would provide a more direct support to our claims.

Acknowledgement

The authors thank Prof. Lajos Izsó and Prof. Eszter Láng for their earlier developments, and the participants for their valuable contribution.

References

[1] K. Lukanov, H. A. Maior, and M. L. Wilson, "Using fNIRS in Usability Testing: Understanding the Effect of Web Form Layout on Mental Workload," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4011-4016

[2] L. Longo, "Subjective Usability, Mental Workload Assessments and Their Impact on Objective Human Performance BT - Human-Computer Interaction - INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part II," R. Bernhaupt, G. Dalvi, A. Joshi, D. K. Balkrishan, J. O'Neill, and M. Winckler, Eds. Cham: Springer International Publishing, 2017, pp. 202-223

[3] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, Vol. 58, No. 1, pp. 1-17, 2015

[4] J. Sauro, *Quantifying the user experience : practical statistics for user research.* Cambridge, MA: Morgan Kaufmann, 2016

[5] L. Gazdi, K. Pomázi, B. Radostyán, M. Szabó, L. Szegletes, and B. Forstner, "Experimenting with classifiers in biofeedback-based mental effort measurement," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* 2016, pp. 331-336

[6] P. Baranyi, A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)* Cham: Springer International Publishing, 2015

[7] S. M. B. M. B. A. Rahman, C. Borst, M. M. M. va. Paassen, and M. Mulder, "Cross-sector transferability of metrics for air traffic controller workload," *IFAC-PapersOnLine*, Vol. 49, No. 19, pp. 313-318, Jan. 2016

[8] J. A. Cheyne, J. S. A. Carriere, G. J. F. Solman, and D. Smilek, “Challenge and error: critical events and attention-related errors.,” *Cognition*, Vol. 121, No. 3, pp. 437-46, Dec. 2011

[9] T. M. Moore, A. P. Key, A. Thelen, and B. W. Y. Hornsby, “Neural mechanisms of mental fatigue elicited by sustained auditory processing,” *Neuropsychologia*, Vol. 106, pp. 371-382, Nov. 2017

[10] X. Wu and Z. Li, “Secondary Task Method for Workload Measurement in Alarm Monitoring and Identification Tasks,” in *Cross-Cultural Design. Methods, Practice, and Case Studies SE - 39*, Vol. 8023, P. L. P. Rau, Ed. Springer Berlin Heidelberg, 2013, pp. 346-354

[11] S. G. Hart, “Nasa-Task Load Index (NASA-TLX); 20 Years Later,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, Vol. 50, No. 9, pp. 904-908, Oct. 2006

[12] A. M. Kring and D. M. Sloan, “The Facial Expression Coding System (FACES): development, validation, and utility.,” *Psychol. Assess.*, Vol. 19, No. 2, pp. 210-24, Jun. 2007

[13] G. H. Gendolla, a E. Abele, and J. Krüsken, “The informational impact of mood on effort mobilization: a study of cardiovascular and electrodermal responses.,” *Emotion*, Vol. 1, No. 1, pp. 12-24, 2001

[14] W. Waterink and A. van Boxtel, “Facial and jaw-elevator EMG activity in relation to changes in performance level during a sustained information processing task,” *Biol. Psychol.*, Vol. 37, No. 3, pp. 183-198, Jul. 1994

[15] S. Chen, J. Epps, N. Ruiz, and F. Chen, “Eye Activity As a Measure of Human Mental Effort in HCI,” in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 2011, pp. 315-318

[16] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011

[17] T. Čegovnik, K. Stojmenova, G. Jakus, and J. Sodnik, “An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers,” *Appl. Ergon.*, Vol. 68, pp. 1-11, Apr. 2018

[18] H. K. Wong and J. Epps, “Pupillary transient responses to within-task cognitive load variation,” *Comput. Methods Programs Biomed.*, Vol. 137, pp. 47-63, Dec. 2016

[19] M. Köles, “A Review of Pupillometry for Human-computer Interaction Studies,” *Period. Polytech. Electr. Eng. Comput. Sci.*, Vol. 61, No. 4, pp. 1-7, Sep. 2017

[20] L. Izsó, *Developing evaluation methodologies for human-computer interaction*. Delft: Delft University Press, 2001

[21] M. A. Wetherell, O. Craw, K. Smith, and M. A. Smith, "Psychobiological responses to critically evaluated multitasking," *Neurobiol. Stress*, Vol. 7, pp. 68-73, Dec. 2017

[22] A. Komlodi and K. Hercegfi, "Exploring cultural differences in information behavior applying psychophysiological methods," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, 2010, p. 4153

[23] L. Laufer and B. Németh, "Predicting user action from skin conductance," in *Proceedings of the 13th international conference on Intelligent user interfaces - IUI '08*, 2008, p. 357

[24] M. Marchitto, S. Benedetto, T. Baccino, and J. J. Cañas, "Air traffic control: Ocular metrics reflect cognitive complexity," *Int. J. Ind. Ergon.*, Vol. 54, pp. 120-130, Jul. 2016

[25] J. Vera, R. Jiménez, J. A. García, and D. Cárdenas, "Intraocular pressure is sensitive to cumulative and instantaneous mental workload," *Appl. Ergon.*, Vol. 60, pp. 313-319, Apr. 2017

[26] R. Chavarriaga, A. Biasiucci, K. Forster, D. Roggen, G. Troster, and J. D. R. Millan, "Adaptation of hybrid human-computer interaction systems using EEG error-related potentials.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, Vol. 2010, pp. 4226-9, Jan. 2010

[27] N. Kumar and J. Kumar, "Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study," in *Procedia Computer Science*, 2016, Vol. 84, pp. 70-78

[28] M. K. Choi, S. M. Lee, J. S. Ha, and P. H. Seong, "Development of an EEG-based workload measurement method in nuclear power plants," *Ann. Nucl. Energy*, Vol. 111, pp. 595-607, Jan. 2018

[29] J. C. Lee and D. S. Tan, "Using a low-cost electroencephalograph for task classification in HCI research," in *Proceedings of the 19th annual ACM symposium on User interface software and technology - UIST '06*, 2006, p. 81

[30] S. Mun, M. Whang, S. Park, and M. C. Park, "Effects of mental workload on involuntary attention: A somatosensory ERP study," *Neuropsychologia*, Vol. 106, pp. 7-20, Nov. 2017

[31] L. (Ben)JM Mulder, D. De Waard, and K. Brookhuis, "Estimating Mental Effort Using Heart Rate and Heart Rate Variability," in *Handbook of Human Factors and Ergonomics Methods*, 2005, pp. 20-28

[32] T. Lin and A. Imamiya, "Evaluating usability based on multimodal information," in *Proceedings of the 8th international conference on Multimodal interfaces - ICMI '06*, 2006, p. 364

[33] T. Heine, G. Lenis, P. Reichensperger, T. Beran, O. Doessel, and B. Deml,

"Electrocardiographic features for the measurement of drivers' mental workload," *Appl. Ergon.*, Vol. 61, pp. 31-43, May 2017

[34] D. W. Rowe, J. Sibert, and D. Irwin, "Heart rate variability: indicator of user state as an aid to human-computer interaction," in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '98*, 1998, pp. 480-487

[35] G. Mulder and van der Mulder-Hajonides, "Mental load and the measurement of heart rate variability.," *Ergonomics*, Vol. 16, No. 1, pp. 69-83, Jan. 1973

[36] K. Hercegfi, "Heart Rate Variability Monitoring during Human-Computer Interaction," *Acta Polytech. Hungarica*, Vol. 8, No. 5, pp. 205-224, 2011

[37] S. H. Fairclough and L. J. M. Mulder, "Psychophysiological processes of mental effort investment.," in *How motivation affects cardiovascular response: Mechanisms and applications*, R. A. Wright and G. H. E. Gendolla, Eds. American Psychological Association, 2011

[38] M. Pagani *et al.*, "Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog," *Circ. Res.*, Vol. 59, No. 2, pp. 178-193, Aug. 1986

[39] M. Koles and K. Hercegfi, "Is the Heart Period Variability-based usability measurement of mental effort robust enough against arm movements?," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)* 2012, pp. 601-604

[40] L. Izsó and E. Láng, "Heart period variability as mental effort monitor in Human Computer Interaction," *Behav. Inf. Technol.*, Vol. 19, No. 4, pp. 297-306, Jan. 2000

[41] E. Láng and N. Szilágyi, "Significance and assessment of autonomic indices in cardiovascular reactions.," *Acta Physiol. Hung.*, Vol. 78, No. 3, pp. 241-60, Jan. 1991

[42] Y. Suda *et al.*, "Changes in ultra-low and very low frequency heart rate variability after coronary artery bypass grafting.," *Biomed. Pharmacother.*, Vol. 55, Suppl 1, pp. 110s-114s, Jan. 2001

[43] A. R. Bilge *et al.*, "Assessment of ultra low frequency band power of heart rate variability: validation of alternative methods.," *Int. J. Cardiol.*, Vol. 71, No. 1, pp. 1-6, Sep. 1999

[44] G. Parati, J. P. Saul, M. Di Rienzo, and G. Mancia, "Spectral analysis of blood pressure and heart rate variability in evaluating cardiovascular regulation. A critical appraisal.," *Hypertension*, Vol. 25, No. 6, pp. 1276-86, Jun. 1995

[45] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Stat. Math.*, Vol. 21, No. 1, pp. 243-247, Dec. 1969

[46] J. Cohen, "A power primer.," *Psychol. Bull.*, Vol. 112, No. 1, pp. 155-9, Jul. 1992

[47] B. Pastötter and K.-H. T. Bäuml, "Retrieval practice enhances new learning: the forward effect of testing.," *Front. Psychol.*, Vol. 5, p. 286, Jan. 2014

[48] M. A. Pyc and K. A. Rawson, "Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?," *J. Mem. Lang.*, Vol. 60, No. 4, pp. 437-447, May 2009

[49] P. Zarjam, J. Epps, F. Chen, and N. H. Lovell, "Estimating cognitive workload using wavelet entropy-based features during an arithmetic task.," *Comput. Biol. Med.*, Vol. 43, No. 12, pp. 2186-95, Dec. 2013

[50] C. Zhang, X. Wu, L. Zhang, X. He, and Z. Lv, "Simultaneous detection of blink and heart rate using multi-channel ICA from smart phone videos," *Biomed. Signal Process. Control*, Vol. 33, pp. 189-200, Mar. 2017

[51] W. Zhou and J. Gotman, "Removal of EMG and ECG artifacts from EEG based on wavelet transform and ICA.," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, Vol. 1, pp. 392-5, Jan. 2004

[52] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. N. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, 2008, p. 835

[53] L. Szegletes, M. Köles, and B. Forstner, "The design of a biofeedback framework for dynamic difficulty adjustment in games," *Cognitive Infocommunications (CogInfoCom) 2014, 5th IEEE Conference on*. pp. 295-299, 2014

[54] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, "Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback," *Int. J. Hum. Comput. Interact.*, Vol. 25, No. 6, pp. 506-529, Aug. 2009

[55] M. Csíkszentmihályi, *Beyond boredom and anxiety*. Jossey-Bass Publishers, 1975