Neurophysiological Estimation of Interruptibility: Demonstrating Feasibility in a Field Context

Santosh Mathan, Stephen Whitlow, Michael Dorneich, Patricia Ververs

Honeywell Laboratories
3660 Technology Drive, Minneapolis, MN 55418
santosh.mathan@honeywell.com

Gene Davis

Advanced Brain Monitoring
2237 Faraday Ave, Suite 100 Carlsbad, CA 92008
Gene@b-alert.com

Abstract

Inappropriately timed interruptions from task-relevant electronic devices have been shown to have a negative impact on accuracy and efficiency in difficult task contexts. Research has also shown that these risks can be minimized by timing interruptions appropriately based on estimates of a user's cognitive workload. The work reported here examines the potential for using body-worn electrophysiological sensors to assess cognitive workload in challenging field environments. Analysis of electroencephalogram (EEG) data gathered from a high fidelity military training exercise reveals that neurophysiological signals can provide the basis for accurate estimation of workload in harsh operational contexts.

1 INTRODUCTION

The impact of interruptions from electronic communication devices is an issue that has received substantial research scrutiny in recent years (Chen & Vertegaal, 2004; Fogarty, Hudson, & Lai, 2004; Iqbal, Adamczyk, Zheng, & Bailey, 2005; Lee & Tan, 2006). Inappropriately timed interruptions can increase errors and reduce efficiency (Iqubal et. al, 2005). The cost of inappropriate interruptions can be substantial in domains where cognitive workload is routinely high and where inefficiencies and errors could mean the difference between life and death—military personnel, police officers, paramedics, and fire fighters face these types of situations on a regular basis.

In recent years researchers have explored effective ways to minimize inappropriate interruptions. Studies have shown that environmental sensors and software instrumentation can provide an accurate basis for estimating a user's interruptibility (Fogarty, Hudson & Lai, 2004; Fogarty, Ko, Aung, Golden, Tang, & Hudson 2005). These estimates allow systems to interrupt when the potential for disruption is low.

While promising, interruption management approaches that rely on instrumentation of the task environments are ideally suited for physically static work environments such as offices. In physically static task contexts the objects and events contributing to cognitive workload can generally be enumerated and consistently monitored using environmental sensors such as cameras, motion detectors, microphones, and software instrumentation. However, in many domains, where the cost of inappropriate interruptions can be extreme, sources that contribute to workload are unpredictable and difficult to track. For example, a police officer could face uninterruptible moments while in vehicles, inside unfamiliar buildings, or on the street—sources of workload could range from planning under tight time constraints to pursuit of a suspect. For work that occurs in varied physical contexts, it may be more appropriate to rely on body worn sensors that provide estimates of interruptibility based on indices of cognitive workload.

1.1 Sensor-based Estimation of Cognitive Workload

While a broad range of sensors can provide insight into cognitive workload, effective interruptibility assessment in mobile contexts calls for sensor systems that: 1) are light-weight and wearable for extended periods of time 2) provide insight into underlying cognitive phenomenon on a near real-time basis 3) demonstrate a high degree of specificity with respect to cognitive activity . Each of these requirements is satisfied by EEG systems.

Mathan, S., Whitlow, S., Dorneich, M., Ververs, T., and Davis, G. (2007) Neurophysiological Estimation of Interruptibility: Demonstrating Feasibility in a Field Context. In Proceedings of the 4th International Conference of the Augmented Cognition Society. October 2007, Baltimore MD.

1.2 Electroencephalogram (EEG)

EEG sensors, worn on the scalp, record electrical activity associated with the firing of neurons. Signals recorded using these sensors are typically spectrally decomposed for analysis using the Fast Fourier Transform (FFT). The FFT operation decomposes each waveform into sinusoidal components that are described by three parameters: amplitude, frequency and phase. The amplitude of EEG over various frequency bands: delta (1 to 4 Hz), theta (4 to 8 Hz), alpha (8 to 13 Hz), beta (13 to 30 Hz), and gamma (30 to 40 Hz), have been shown to vary in conjunction with different brain states. For example, delta activity is dominant during deep sleep; alpha activity is typically observed wakeful but relaxed states; beta and gamma activity is prominent during problem solving and other complex cognitive tasks (Scerbo, Freeman, Mikulka, Parasuraman, Dinocera, and Prinzel, 2001).

1.3 EEG in the Field

EEG sensors have been used in laboratory settings and clinical contexts for decades. In recent years, researchers have begun exploring the potential for using EEG systems as a human computer interaction (HCI) modality (e.g. Chen & Vertegaal, 2004; Fogarty et. al. 2004). Much of the existing EEG research—including the HCI work—has focused on seated subjects in laboratory settings. Evaluations of traditional EEG systems in mobile task contexts have shown that these signals are severely compromised by noise artifacts (Kerick, Oie & McDowell, in print). EEG signals are prone to contamination from sources such as eye blinks, facial muscle activity, and electrical interference. Additionally, most EEG systems used in laboratory contexts require considerable wiring to link electrodes to data acquisition hardware—even modest movement of sensor cables and wires has the potential for inducing artifacts that can overwhelm signals associated with cognitive activity. Typical lab hardware is also too heavy and bulky to be practical in many work contexts.

Recently, lightweight and low profile wireless electrophysiological sensor systems have begun appearing on the market. For example, the B-Alert™ EEG sensor set (Advanced Brain Monitoring Inc., Carlsbad, CA) used in this effort takes the form factor of a snug baseball cap—the system is designed to relay neurophysiological signals wirelessly to processing computers over a Bluetooth connection. The system is compact enough to fit under a helmet. While the compact profile of newer EEG system reduces the potential for many motion related artifacts, these benefits come at the cost of a lower density of EEG sensors. For example, unlike laboratory systems that employ anywhere from 32 to 256 electrodes, the B-Alert system has only 6 electrodes—providing limited information about underlying cognitive activity. The question of whether low density EEG systems, in conjunction with relatively simple artifact reduction algorithms, provide enough information for accurate cognitive workload estimation in artifact rich environments remains an open one.

2 METHOD

Is EEG based workload estimation feasible outside the pristine environment of the laboratory? We sought an answer to this question by evaluating workload classification in the environment of a combat training exercise. Tasks called for problem solving and decision making in an environment characterized by substantial physical activity, including gross head movements, running, walking, and talking.

2.1 Participants and Task

EEG data was gathered during an hour-long simulated military mission at a United States Army training and testing facility. A platoon, consisting of 32 members of a National Guard unit, participated in an urban operations combat training mission. The scenario called for entering and clearing buildings in an urban environment. Simulated enemy forces and simunitions were employed for operational fidelity.

Three team leaders—a platoon leader (PL), platoon sergeant (PSG), and a squad leader (SL) —were instrumented with the B-Alert wireless EEG system under their helmets. Team leaders were chosen as subjects because their task required both tactical combat activity and substantial interaction with electronic communications technology. The cognitive workload imposed on these leaders varied as a function of combat activities and communications necessary to direct and coordinate teams.

2.2 Hardware Design

Several adaptations had to be made to the B-Alert system in order to minimize noise artifacts in the harsh environment of a high fidelity training exercise. First, the system relied on flexible flat cables – a design that integrated all wiring within a thin, flat, film-like insulating layer. This not only reduced the overall vertical profile of the system, but also eliminated independent, loose wires that can become the source of noise artifacts. Second, the system relied on differential amplifiers – that amplify differences between electrodes while rejecting common signals—usually large artifacts—through a process called Common Mode Rejection (CMR). Third, the system amplified and digitized signals close to the scalp in order to reduce the potential for transmission related signal attenuation. Fourth, customized foam padding was developed to keep electrodes in place on the scalp and to reduce the possibility of contact with the user's helmet.

Figure1. The B-Alert EEG System adapted for use under a helmet in ambulatory conditions

2.3 Signal Processing

EEG was collected using the B-Alert EEG system. Signals were sampled from 6 bipolar channels (CzPOz, FzPOz, F3Cz, F3F4, FzC3, C3C4) at a rate of 256Hz. EEG signals were processed in real time to minimize the impact of noise artifacts on signals associated with cognitive activity.

As described in (Berka, Levendowski, Lumicao, Yau, Davis, Zivkovic, Olmstead, Tremoulet, and Craven, 2007), prior to all other signal processing, a 60 Hz notch filter was applied to remove environmental electrical noise. Artifacts such as spikes, amplifier saturation and excursions -- related to movement -- were analyzed in the time domain. Spikes and excursions were identified when the EEG amplitude changed significantly (e.g., > 40 μV) over short durations (e.g., ~12 to 27 ms.). Amplifier saturation was recognized when the change in amplitude between two data points exceeded predefined thresholds (e.g., 440 μV) or the EEG amplitude approached the maximum/minimum of the amplifier dynamic range. These thresholds were established based on the unique characteristics of the amplifier circuit which was utilized. The data points associated with these artifacts are stored in memory and used in a later step to decontaminate the EEG.

The EEG was then deconstructed using a wavelets transformation into the 0-2, 2-4, 4-8, 8-16, 16-32, 32-64 and 64-128 Hz wavelets bands. The wavelets power 64-128 Hz band was used to identify epochs with excessive muscle activity (EMG) that were rejected from further analysis. To detect eye blinks, a linear discriminant function analysis was employed which uses the absolute value of the 0-2, 2-4, 4-8, 8-16, and 16-32 Hz wavelet coefficients from the 50th, 40th, 30th, 20th and 10th data points before and after the target data point from FzPOz and CzPOz as variables to classify each data point as an eye blink, theta wave or non-eye blink. Decontamination of eye blinks was

accomplished by computing mean wavelet coefficients for the 0-2, 2-4 and 4-8 Hz bins from nearby non-contaminated regions and replacing the contaminated data points. The EEG signal was then reconstructed using all wavelets bands except 64-128 Hz.

The data points previously associated with spikes, excursions or saturation were recalled from memory and replaced with zero values at zero crossing before and after spikes, excursions and saturations. Finally, EEG absolute and relative power spectral density (PSD) variables for each 500 millisecond epoch were computed using a Fast-Fourier transform applied using a 50% overlapping Kaiser window ($\alpha = 6.0$). The PSD values were scaled to accommodate the insertion of zero values as replacements for the artifact.

PSD values were integrated over various ranges provide estimates of power in frequency bands associated with cognitive activity (delta, theta, alpha, beta, and gamma). These estimates were output every 500 milliseconds in the form of a 30 element vector (6 channels x 5 frequency bands). Classification of cognitive workload was based on the analysis of these vectors.

2.4 Characterization of Ground Truth

An important step towards creating a system that can classify EEG data is an accurate characterization of ground truth—i.e. associating each sample of EEG with accurate estimates of the user's cognitive workload. EEG samples from the training exercise, coupled with ground truth labels, serve an important role in training and testing the classifier.

The training scenario that is the focus of this paper was designed to have distinct and prolonged periods of low and high workload. In high cognitive work-load conditions participants had to carry out multiple tasks concurrently under considerable time pressure. Any distraction from these critical tasks had the potential for compromising performance. In low cognitive workload conditions these tasks could be carried out serially, with little time pressure. Ground truth characterization was carried out based on inputs from the three participants—PL, PSG, and SL. All participants viewed video recordings of their performance and provided retrospective estimates of their interruptibility during exercise events. Additionally, two independent raters, one with military experience, coded 15-second segments of experimental video protocol for each subject. The categorized each segment as either high or low workload. Raters considered both their own judgments and participant characterizations of cognitive workload. All EEG samples falling within a given segment received a common workload label. Distractions could be accommodated without compromising performance. Agreement between the two rates was high (88%); discrepancies were reconciled through mutual consensus.

2.5 Classification

The PSD features that form the basis for classification contains information pertinent to the classification of cognitive states, as well as irrelevant components and noise. Accurate classification of workload based on EEG calls for a system that can estimate workload by identifying dimensions or features of EEG that are informative with respect to distinctions among workload levels. Our efforts relied on the logistic classifier. A logistic classifier assumes that the relationship between a set of independent variables (EEG features in this context) and the estimated probability of membership in a class (high or low workload) can be modelled in terms of a sigmoid function: $P(c|y)=1/(1+ey)$. Model parameters are identified using maximum likelihood estimation. The decision boundary created by this classifier is linear. Linear classifiers are widely used by EEG researchers as their inherently low complexity limits the possibility of overfitting – an issue of concern in artifact rich mobile task contexts.

The metric used to evaluate classification performance in our effort is the Area under the Receiver Operating Characteristic (ROC) curve (Duda, Hart, & Stork, 2001). ROC curves plot true positives (on the y-axis) against false positives (on the x-axis) as a threshold for discriminating between targets and distractors is varied. It is widely used to evaluate human and machine signal detection capabilities. The ROC curve provides a way to assess the degree of overlap between the output of a classifier for two classes of data. Perfect classification produces an area under the curve value (Az) of 1.0, while chance performance produces an Az value of 0.5.

3 RESULTS

3.1 Qualitative analysis

We performed a qualitative analysis of the EEG power spectrum for both subjects (Figure 2). Our analysis revealed that power in the gamma (30 to 40 Hz) band, and to a lesser extent in the beta (13 to 30 Hz) band, rose as a function of workload over several electrode sites. In contrast, alpha power (8 to 12 Hz) decreased under high workload conditions. While the prominence of alpha and beta activity varied considerably across sites and subjects, gamma activity was consistently higher in high workload conditions across sensors and subjects. These finding are consistent with prior (e.g. Laine, Bauer, Lanning, Russell, and Wilson, 2002).

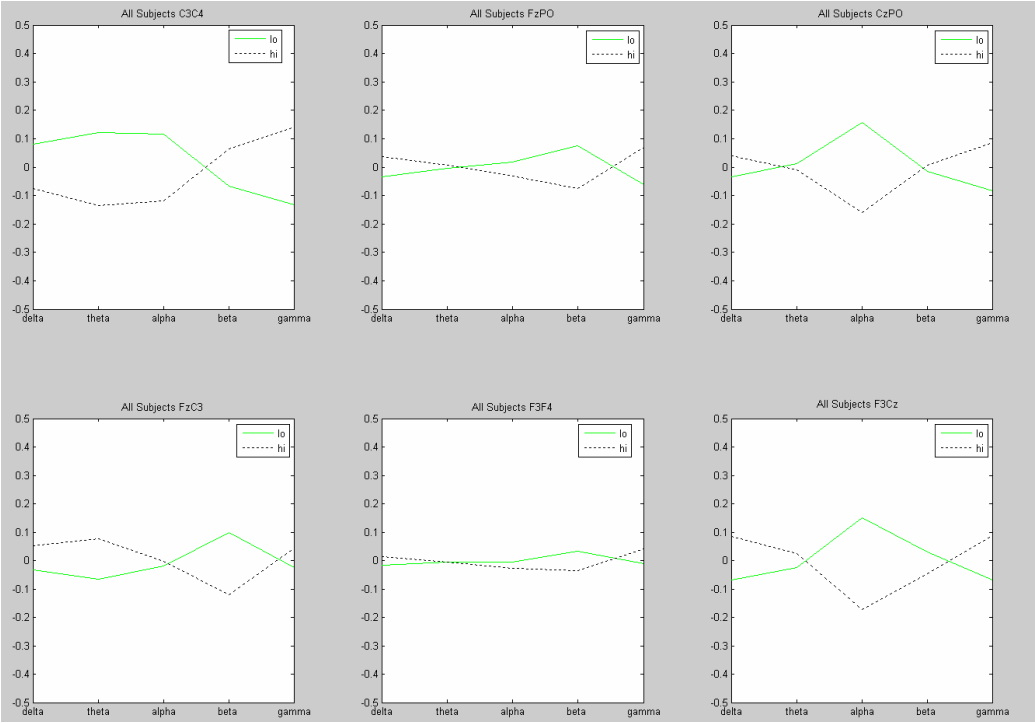

Figure 2. Average activity across subjects in high and low workload conditions. Gamma activity increases with workload at all channels. Beta increases, while alpha decreases, with workload at several channels.

3.2 Data Set

Only a subset of data acquired was used for workload classification. On average, 26% of samples were rejected by the real-time signal processing components due to artifact contamination. Rejection rates were similar in high and low workload conditions (14.7% vs 10% respectively). The data set used in this analysis was fairy balanced between high and low workload data.

	PL	PSG	SL	Average
Overall sample rejection	40	20	18	26.0
% samples rejected in LO	9	17	4	10.0
% samples rejected in HI	31	2	11	14.7
% Samples HI	53	39	54	48.7
% Samples LO	46	61	41	49.3
% Unknown	1	0	6	2.3

Table 1 Sample Rejection Statistics

3.3 Cross Validation

We assessed classification performance using cross validation (Figure 3). We employed both ten-fold cross validation and a more conservative two-fold cross validation procedure. With ten-fold cross validation, the data set is split into 10 subsets. Over the course of ten iterations, a new subset is picked to serve as testing data, while the remaining 9 folds serve as the training data. In contrast, with two fold cross validation, the data set is split into two. Over two iterations, each half of the data set is used in turn to train the classifier while the other half is used for testing. Two-fold cross validation provides a more pessimistic estimate of classification accuracy because less of the data set is used for training.

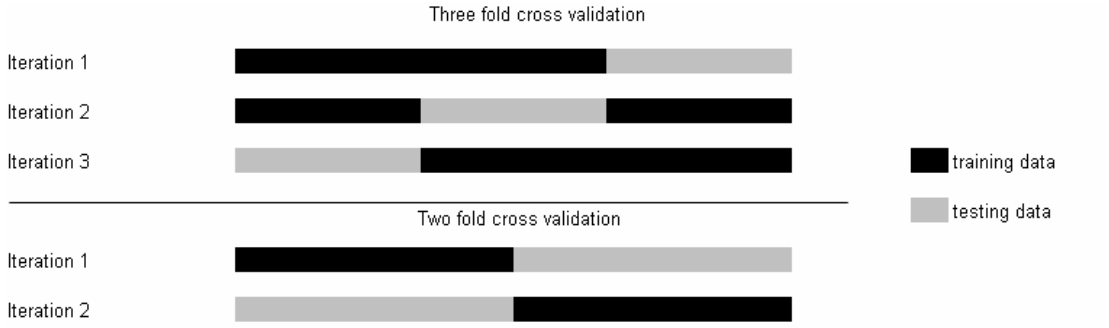

Figure 3. N-fold cross validation illustrated with n ==3 and n=2. Data is split into n equally sized segments. Over n iterations each data subset is used in turn for testing while the remaining data is used for training.

Results from these analyses point to the feasibility of accurate cognitive state estimation in challenging field settings. Average base classification ranged from 0.72 area under the ROC curve (AUC) with 10-fold cross validation to 0.65 AUC with 2-fold cross validation.

We considered temporal smoothing as a strategy for dealing with intermittent classification errors stemming from the noise inherent in the field environment. This strategy assumes that task demands remain stable over the span of the smoothing window. Smoothing was accomplished using a median filter on the output of the classifier over specific time windows. One consequence of temporal smoothing of classifier output is to introduce a lag in the decision process. Our analysis considered the trade off in accuracy as the temporal window of output smoothing was varied.

Classification accuracy for both subjects rose monotonically up to a one minute long temporal smoothing window. However, the rate at which temporal smoothing benefits accuracy appeared to diminish as window sizes increase. Temporal smoothing of 10 seconds contributed to a rise in classification accuracy— average accuracy rose from 0.72 to 0.86 with ten-fold cross validation and from 0.65 to 0.78 with two-fold cross validation.

	Temporal Smoothing Window (seconds)																
	base	1	1.5	2	2.5	5	10	15	20	25	30	35	40	45	50	55	60
PL 2-fold	0.65	0.70	0.71	0.72	0.73	0.77	0.81	0.83	0.84	0.86	0.86	0.87	0.88	0.88	0.88	0.89	0.89
PSG 2-fold	0.73	0.78	0.79	0.81	0.82	0.85	0.87	0.88	0.89	0.9	0.91	0.92	0.92	0.93	0.93	0.93	0.93
SL 2-fold	0.57	0.59	0.59	0.6	0.6	0.62	0.64	0.66	0.67	0.67	0.68	0.68	0.69	0.7	0.71	0.71	0.71
mean 2-fold	0.65	0.69	0.7	0.71	0.72	0.75	0.78	0.79	0.8	0.81	0.82	0.82	0.83	0.84	0.84	0.84	0.84
PL 10-fold	0.71	0.76	0.78	0.8	0.81	0.85	0.88	0.91	0.92	0.92	0.92	0.93	0.93	0.94	0.94	0.94	0.94
PSG 10-fold	0.81	0.86	0.88	0.89	0.9	0.93	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
SL 10-fold	0.63	0.68	0.7	0.71	0.71	0.74	0.77	0.78	0.79	0.79	0.79	0.8	0.8	0.8	0.81	0.81	0.81
mean 10-fold	0.72	0.77	0.79	0.8	0.81	0.84	0.86	0.88	0.88	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90

Table 2 Classification performance as a function of temporal smoothing

4 DISCUSSION

These findings open up the possibility of managing interruptions from electronic communication devices based on estimates of cognitive workload derived from EEG sensors. Knowledge of a user's workload could be leveraged in a variety of ways. For example, originators of messages could assess a recipient's interruptibility before sending radio messages. Alternatively, devices could defer messages that are not critical to the task at hand.

EEG based assessments of interruptibility are particularly appropriate in work contexts where instrumentation of the task environment with sensors may not be practical. Unfortunately, traditional laboratory EEG systems have fared poorly in field contexts due to their susceptibility to artifacts. The work reported here suggests that wireless EEG sensors, used in conjunction with a relatively simple signal processing approach, can provide the basis for accurate estimation of cognitive workload. Validation in the challenging context of a combat exercise—an environment in which participants ran, walked, and moved their heads around freely—has positive implications for the likely efficacy of EEG based workload estimation across a wide range of task settings.

While the results presented here are promising, there are several issues that have to be addressed in future work.

Robustness: Cognitive state classification in the effort reported here was based entirely on likelihood estimates provided by a logistic classifier. However, Bayesian decision theory suggests that other factors such as prior probability of workload states and the cost of classification errors must considered in order to make optimal decisions based on noisy data. We will examine the benefits of considering priors and costs in future work. We will also examine avenues for improving likelihood estimates provided by the classifier by fusing information from other physiological sensor sources.

Other Data Sources: While the analysis here has been limited to EEG based workload classification. Other data sources such as cardiac sensors and accelerometers can complement EEG data to raise overall classification accuracy.

Generalization: The analysis described here was limited to data from an hour long session. However, generalization of classifiers over long periods of time—particularly with changes in tonic physiological states such as fatigue and stress is an issue that will be examined and addressed in our future efforts. Cognitive responses to tasks and their electrophysiological manifestation are likely to change as a function of underlying states. While the results presented here provide a proof of principle that accurate cognitive state classification is possible in challenging mobile contexts, validation with larger groups of subjects is necessary to determine if these results are broadly representative.

Individualization: Classification performance varied considerably across individuals. It is possible that the placement of sensors may have been suboptimal for some individuals. Customization of EEG sensor placement to each individual may provide an avenue for improving system performance.

5 ACKNOWLEDGMENTS

This research was supported by DARPA and the U.S. Army Natick Soldier Center (DAAD16-03-C-0054). The opinions expressed here are those of the authors and do not necessarily reflect the views of DARPA or the U.S. Army.

6 REFERENCES

1. Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT, Olmstead RE, Tremoulet PD, Craven PL. (2007) *EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks.* Aviation Space and Environmental Medicine; 78(5, Suppl.):B231-B244.
2. Chen, D., & Vertegaal, R. (2004) Using mental load for managing interruptions in physiologically attentive user interfaces. Proceedings of CHI 2004, 1513-1516
3. Duda, R.O, Hart, P.E., and Stork, D. G (2001). Pattern Classification. John Wiley and Sons, 2nd edition.

4. Fogarty, J., Hudson, S. and Lai, J. (2004) Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. Proceedings of CHI 2004, 207-214.

5. Fogarty, J., Ko, A.J., Aung, H.H., Golden, E., Tang, K.P. and Hudson, S.E. (2005) Examining Task Engagement in Sensor-Based Statistical Models of Human Interruptibility. Proceedings of CHI 2005, 331-340.

6. Iqbal, S.T., Adamczyk, P.D., Zheng, X. S., Bailey, B.P. (2005) Towards an index of opportunity: understanding changes in mental workload during task execution. Proceedings of CHI 2005, 311-320.

7. Kerick, S.E., Oie, K.S., & McDowell, K. (Manuscript in preparation). Assessment of EEG Signal Quality in Motion Environments. Manuscript in preparation.

8. Laine, T.I., Bauer, K.W., Lanning, J.W., Russell, C.A., and Wilson, G. F., (2002) Selection of input features across subjects for classifying crewmember workload using artificial neural networks. IEEE Transactions on Systems, Man and Cybernetics, Part A,32(6):691–704

9. Lee, J., Tan, D. Using a Low-Cost Electroencephalograph for Task Classification in HCI Research. (2006) Proceedings of UIST 2006, 81-90.

10. Scerbo, M.S., Freeman, F.G., Milkulka, P.J., Parasuraman, R., Dinocera, F., & Prinzel, L.J. (2001) The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology (Technical Paper NASA/TP-2001-211018). Hampton, VA: NASA Langley Research Center.

.