Examining the Reliability of Using fNIRS in Realistic HCI Settings for Spatial and Verbal Tasks

Horia A. Maior[*†], Matthew Pike[*], Sarah Sharples[‡] and Max L. Wilson[*]

[*]Mixed Reality Lab
University of Nottingham
Nottingham, UK

[†]Horizon Centre for Doctoral Training
University of Nottingham
Nottingham, UK

[‡]Human Factors Research Group
University of Nottingham
Nottingham, UK

{horia.maior, pike, sarah.sharples, max.wilson}@nottingham.ac.uk

ABSTRACT

Recent efforts have shown that functional near-infrared spectroscopy (fNIRS) has potential value for brain sensing in HCI user studies. Research has shown that, although large head movement significantly affects fNIRS data, typical keyboard use, mouse movement, and non-task-related verbalisations do not affect measurements during *Verbal* tasks. This work aims to examine the *Reliability* of fNIRS, by 1) confirming these prior findings, and 2) significantly extending our understanding of how artefacts affect recordings during *Spatial* tasks, since much of user interfaces and interaction is inherently spatial. Our results show that artefacts have a significantly different impact during Verbal and Spatial tasks. We contribute clearer insights into using fNIRS as a tool within HCI user studies.

ACM Classification Keywords

H.5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces

Author Keywords

functional near-infrared spectroscopy; fNIRS; brain-computer interface; human cognition; BCI

INTRODUCTION

Recent research has shown functional near-infrared spectrsocopy (fNIRS) to be a highly suitable brain sensing technology for typical HCI user studies, providing an objective, non-intrusive measure correlating to what is known as human Mental Workload. Solovey et al. [10] showed that some typical interactions, like typing on a keyboard and using a mouse, did not create significant artefacts in fNIRS measurements during Verbal memory tasks, as long as forehead and major head movements were avoided. Further, Pike et al. [7] showed that only non-task-related verbalisation created additional workload measured with fNIRS during Verbal memory tasks. This work aims to explore the reliability of fNIRS data as a measure of mental workload, by 1) replicating the prior findings of Solovey et al. [10] and Pike et al. [7], and 2) extending our understanding of how these artefacts affect *Spatial* tasks, since much about user interfaces and interaction involves Spatial working memory [8]. We hypothesised that each artefact would have different measurable affects on spatial tasks than with verbal tasks.

Mental Workload

Mental Workload (MWL) is a concept used to describe how much mental effort is being experienced by an individual when completing a task. Baddeley and Hitch first proposed that Working Memory (WM) was composed of multiple components [2]. The model distinguishes between two types of encodings: Spatial (Visuospatial Sketchpad) and Verbal (Phonological Loop). Using this model as a foundation, we can develop tasks to target each of these encoding types, allowing us to investigate whether measurement techniques can detect them.

The Multiple Resource Model (MRM) proposed by Wickens [12] illustrates how resource limitations and coordination affects the interrelation of MWL in tasks. Wickens describes that necessary resources are limited, and aims to illustrate how elements of the human information processing system such as attention, perception, memory, decision making and response selection interconnect.

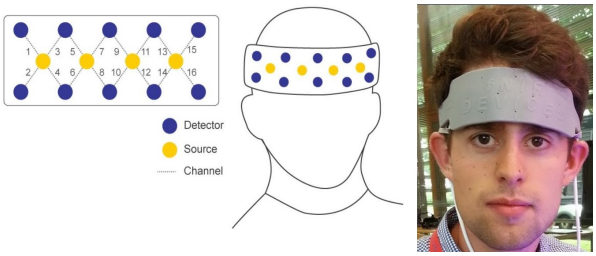

Figure 1. Source detector diagram and placement.

Measuring Mental Workload

A wide variety of measures are used to capture MWL. Subjective measures (e.g. questionnaires, think aloud protocols, interviews, and NASA-TLX scale [3]) are useful for capturing a user's perception of a system and are minimally intrusive if applied infrequently. Objective measures (e.g. task performance, secondary measures, physiological changes) are

		Control	Artefacts				
			Head Movement	Typing	Mouse Movement	Facial Movement	Verbalising
Task	Verbal	✓HbO	℧	✓HbO			✓HbO
	Spatial	✓HbO	✓Hb	℧			✓HbO
Non-Replicated Conditions, Replicated Conditions, Novel Conditions							

Table 1. Results and Contributions of the current and Solovey et al. study. ✓HbO means fNIRS is fine to use in the presence of the investigated artefact, best measure to use HbO. ℧ means that the artefact needs to be avoided or filtered.

complex methods, but do have the potential to provide a real-time continuous measure of the user's state. To be valuable for HCI research however, the equipment should be non-intrusive, ideally allowing "normal" interaction with the system.

fNIRS offers the potential to provide continuous, detailed insight into human mental workload, enabling an objective means of detecting overload conditions during complex tasks in a minimally intrusive manner. Our fNIRS device (Figure 1) allows for the easy and direct application to an individual's forehead, targeting the prefrontal cortex; an area of the brain typically associated with Working Memory [2, 4]. fNIRS measures the change in Oxygenated (HbO) and Deoxygenated (Hb) haemoglobin as the body responds to the individuals' cognition (when measured on the head).

Reliability of fNIRS

Sharples and Megaw [9] state the appropriate criteria for MWL measures as being: Validity, Reliability, Generalisability, Sensitivity, Interference, Diagnosticity, Selectivity, Granularity/Bandwidth, Feasibility of use, Acceptability/Ethics and Resources. Pike et al., Maior et al. and Peck et al. [7, 5, 6] provide evidence of fNIRS correlating with NASA-TLX, a widely used measure of MWL (**Validity**). fNIRS is inherently generalisable as it simply measures oxygenation and is not specific to a particular domain (**Generalisability**).

Afergan et al. [1] demonstrated the ability to distinguish between different workload states (**Sensitivity**) during a UAV simulation task. Additionally, the study identified workload changes over time (**Bandwidth**). Pike et al [7] identified non-related verbalisations as being a contributing factor to increased mental workload (**Diagnosticity**). Solovey et al [10] demonstrated that fNIRS was able to distinguish between common human behaviours (typing, mouse movement, head and facial movement) and a Verbal Memory task (**Selectivity**).

fNIRS has been deployed in a number of studies and has caused minimal **Interference**, with many reporting ecological validity whilst using fNIRS (also demonstrating **Feasibility of use, Acceptability/Ethics** and **Resources**)[7, 10, 1]. In this context however we are particularly interested in exploring the **reliability** of fNIRS within HCI.

As fNIRS is an emerging technology in this field, replicating the findings of existing work is one step towards establishing the reliability of the technology.

EXPERIMENT DESIGN

The aim of this study is to identify the reliability of fNIRS as a measure of MWL in the context of HCI. We have chosen to replicate the work of Solovey et al. [10] and Pike et al. [7] on *verbal* memory, but significantly extend our understanding of the impact of artefacts on fNIRS measurements, by also examining *Spatial* tasks. Reliability of the measure is one of the criteria identified by Sharples and Megaw [9] as being appropriate for measuring MWL. In this study we followed much of the original procedure as described by Solovey et al., we did however remove some of the behaviours under study deciding instead to focus on the behaviours that had the greatest impact on the fNIRS signal (Typing and Head Movement). We included another human behaviour in our study - Verbalisation, a common part of HCI studies that was absent in the original study, but studied by Pike et al. [7].

In this study we also addressed the issue of memory encoding types and how the artefacts under study can affect these different encodings. In the original study, the task of memorising a 7 digit number was Verbal since the encoding of the digits would reside within the Phonological Loop of Baddeley and Hitch's model of WM [2]. To understand the impact of a different encoding, we introduced a Spatial task of memorising a 6x6 grid. This task will be encoded in the Visuospatial Sketchpad (according to the same model [2]), allowing us to investigate whether there are differences in results according to the encoding type of the task.

The study had 4 conditions, which were tested under both task types: 1)Task Only (No Artefact), 2)Task + Head Movement, 3)Task + Typing, 4)Task + Verbalising. We followed the same repeated measures, within-participants approach as the original study to compare conditions.

Participants

Fifteen participants (11 male, 4 female) with an average age of 22.06 (SD = 2.31) were recruited to take part in the study. All participants had normal or corrected vision and reported no history of head trauma or brain damage. The study was approved by the school's ethics committee. Participants provided informed consent and were compensated with gift vouchers.

Tasks and Study Procedure

The study procedure closely followed that of the original study by Solovey et al. Participants were asked to memorise a 7 digit number for the Verbal Task and then submit the number using an on-screen form. For the Spatial Task, participants were required to memorise a 6x6 black and white grid (Figure 2), and recreate the grid using an on-screen form. Participants completed 8 experiments (3 Artefact conditions and 1 Control (No Artefact) condition, under 2 tasks), with each experiment composed by 8 trials. Each trial started with 15s rest, followed

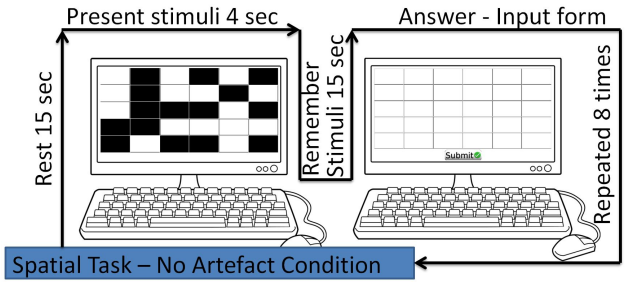

Figure 2. Experiment Procedure with Spatial Task

by 4s presented stimuli (Verbal number or Spatial grid), 15s remembering the stimuli, and ended with an input form for answering the remembered stimuli. For the artefact conditions the 15s remembering period also included performing the specific artefact, and an additional 15s period of performing the artefact alone was performed after the task.

Measurements and Equipment

In this study we have collected two types of measures, namely brain activity using fNIRS and task performance.

fNIRS data was recorded using an fNIRS300 device and the associated COBI Studio recording software provided by Biopac Systems Inc. Using the Matlab Toolbox NIRS-SPM [13] we applied filtering algorithms to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. Finally we separated each trial according to the condition under test (rest/ task/ artefact) considering the slow hemodynamic response [11], and averaged the data accordingly.

Task performance for both task types was calculated using 2 measures: Absolute performance - where an answer is simply correct or not, and Relative Performance - where answers were scored according to distance from the target answer (calculated with Levenshtein distance).

RESULTS

Performance data

No significant difference between conditions was reported by Solovey et al. in task performance, where the number of correct (in-place) digits was used as the dependent variable. We hypothesised that non-related verbalisation will negatively impact performance during the Verbal task, as demonstrated by Pike et al. [7]. Based on Wickens MRM [12], we expect no performance differences under Spatial conditions as the resources are complementary. However, a within participants, one-way repeated measure ANOVA with LSD correction, revealed that participants performed significantly worse under the typing artefact compared to all other conditions during the Verbal task ($N = 15, p < 0.05$). Participants also performed significantly worse in the Verbalisation artefact condition compared to the no artefact one during the Spatial task ($N = 15, p < 0.025$). The findings disprove our hypothesis, but do lead to an interesting discussion; For the Verbal task, the greatest interference was typing, which could be interpreted as being a Spatial input modality since the keys have a physical mapping. Whereas for the Spatial task, the verbalising artefact had the greatest interference providing a **crossing** of resource modalities, which is the opposite of our original hypothesis.

Experiments: No artefacts

The main observation we wanted to reproduce in this experiment was distinguishing between states of rest and cognition, a distinction described as "fundamental" in the original study. We hypothesised that in addition to reproducing the distinction in the Verbal task (as identified by Solovey et al.), we would also identify the difference in the 2 states for our Spatial task. For both Verbal and Spatial tasks, a paired-sample t-test, within participants, revealed significant differences over multiple channels between rest periods and task periods. In both task conditions HbO was significantly higher in 13 out of the 16 channels of data, with $N = 15, p < 0.05$. Our results are in line with those identified by Solovey et al. and with our hypothesis regarding the Spatial task. We note that our results favoured HbO over Hb in the detection of these states. To provide an interesting visual representation of fNIRS ability to distinguish between rest and cognitive states, Figure 3 visualises a participant's fNIRS data (ch.1) for the no artefact experiment (consisting of 8 trials hence the 8 peaks in HbO data).

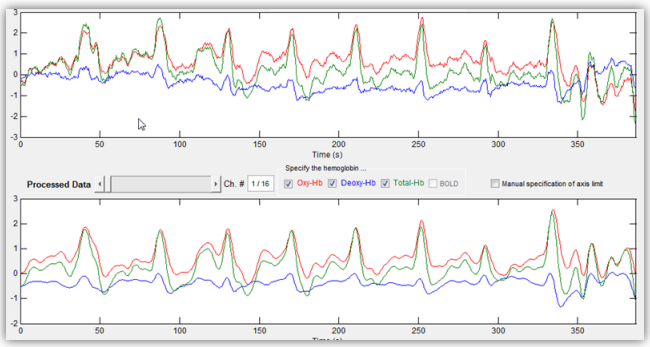

Figure 3. Oxygenation level peaks for 8 Verbal trials.

Experiments: With artefacts

Our interest here lies in distinguishing cognition in the presence of artefacts (Table 1 provides the summary of our findings). To achieve this, we combine the following three stages: rest periods, artefact (alone) periods, and cognitive task under artefact into paired comparisons. We have applied a series of one-way repeated measure ANOVAs within participant design with LSD correction for each of the artefact conditions.

Head Movement Artefact - For the Verbal task we were not able to significantly distinguish between participants at rest and participants performing the cognitive task in the presence of major head movement, as reported by Solovey et al. We were however able to distinguish between participant at rest and participants performing just the artefact ($Hb, p < 0.025$), indicating that head movement is detrimental to the fNIRS signal. We were also able to distinguish between cognition in the presence of head movement and performing the artefact alone ($Hb, p < 0.01$), indicating the potential for filtering of

this artefact in the future. Accordingly, we advise the sampling of major head movements as a part of studies involving a Verbal task. For the same artefact (head movement) under the Spatial task our results suggest that we could relax the restrictions. For the Spatial Task we report significance in Hb for all comparisons ($p < 0.05$) indicating that Spatial based tasks are less prone to head movement artefacts.

Keyboard Input Artefact - For the typing condition we were able to distinguish between rest and task periods ($HbO, p < 0.05$) during the Verbal Task. During the Spatial Task, the difference was not significant. Potential for filtering exists again due to the significance difference between the remaining two comparisons (rest vs artefact and artefact vs cognitive task = $HbO, p < 0.05$). The findings suggest that keyboard input does not affect the fNIRS signal during verbal tasks, however, it should only be controlled for the spatial ones.

Verbalisation Artefact - There were significant differences between rest and cognition periods for both Verbal and Spatial tasks ($HbO, p < 0.01$ Verbal Task and $HbO, p < 0.025$ Spatial Task) in the presence of verbalisation artefact. This finding implies that fNIRS could be reliably used in the presence of Verbalisation artefacts, confirming the findings of Pike et al. [7]. The results also show that Verbalization artefact is the most compatible with fNIRS for a typical HCI Settings.

CONCLUSIONS

In this study we sought to replicate and extend the work performed by Solovey et al., investigating the effect of common human behaviours on fNIRS ability to distinguish states of cognition from other states. Our aim in doing so was to prove the reliability of fNIRS as a measure and extend our understanding of artefacts effects on different task types (Verbal *and* Spatial). The fundamental finding confirmed by this study is that we are able to distinguish between cognitive and rest states in both Verbal (as confirmed by Solovey et al.) and Spatial tasks. The two types of tasks, however, were differently affected, according to the two key fNIRS measures, for each artefact. Our addition of a Spatial task, therefore, provided a greater understanding of fNIRS' ability to distinguish cognition under tasks using such encodings. Further, our inclusion of the verbalisation artefact also provided this greater understanding for an additional, but very common user study behaviour. These findings contribute towards a body of evidence to suggest that, in a HCI context, fNIRS is a indeed valuable measure. To provide further practical advice to other researchers about fNIRS reliability and portability, future work might examine other untested artefacts, such as: age of participants, interface familiarity, task expertise etc.

ACKNOWLEDGMENTS

This work was partially supported by the EPSRC ORCHID project (EP/I011587/1) and the Horizon Centre for Doctoral Training at the University of Nottingham (EP/G037574/1).

REFERENCES

1. Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., Chang, R., and Jacob, R. J. Dynamic difficulty using brain metrics of workload. In *Proc. SIGCHI*, ACM (2014), 3797–3806.
2. Baddeley, A. D., and Hitch, G. Working memory. *The psychology of learning and motivation 8* (1974), 47–89.
3. Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload* (1988).
4. Kane, M. J., and Engle, R. W. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review 9*, 4 (2002), 637–671.
5. Maior, H. A., Pike, M., Wilson, M. L., and Sharples, S. Continuous detection of workload overload: An fnirs approach. In *Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014, Southampton, UK, 7-10 April 2014*, CRC Press (2014), 450.
6. Peck, E. M., Yuksel, B. F., Ottley, A., Jacob, R. J., and Chang, R. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *Proc. SIGCHI*, ACM (2013).
7. Pike, M. F., Maior, H. A., Porcheron, M., Sharples, S. C., and Wilson, M. L. Measuring the effect of think aloud protocols on workload using fnirs. In *Proc. SIGCHI*, ACM (2014), 3807–3816.
8. Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., and Van Dantzich, M. Data mountain: using spatial memory for document management. In *Proc. UIST*, ACM (1998), 153–162.
9. Sharples, S., and Megaw, T. *The definition and measurement of human workload*. In Wilson, J.R. & Sharples, S. (Eds) Evaluation of Human Work. Boca, 2015.
10. Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., and Jacob, R. J. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST*, ACM (2009), 157–166.
11. Villringer, A., and Chance, B. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences 20*, 10 (1997), 435–442.
12. Wickens, C. D. Multiple resources and mental workload. *The Journal of the Human Factors and Ergonomics Society 50*, 3 (2008), 449–455.
13. Ye, J. C., Tak, S., Jang, K. E., Jung, J., and Jang, J. NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy. *Neuroimage 44*, 2 (2009), 428–447.