DETECTION AND ESTIMATION OF WORKING MEMORY STATES AND COGNITIVE FUNCTIONS BASED ON NEUROPHYSIOLOGICAL MEASURES

EDITED BY: Felix Putze, Christian Mühl, Fabien Lotte, Stephen Fairclough and Christian Herff

frontiers Research Topics

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

DETECTION AND ESTIMATION OF WORKING MEMORY STATES AND COGNITIVE FUNCTIONS BASED ON NEUROPHYSIOLOGICAL MEASURES

Topic Editors:
Felix Putze, University of Bremen, Germany
Christian Mühl, German Aerospace Center, Germany
Fabien Lotte, Inria/LaBRI (CNRS - Bordeaux INP - Univ. Bordeaux), France
Stephen Fairclough, Liverpool John Moores University, United Kingdom
Christian Herff, Maastricht University, Netherlands

Executive cognitive functions like working memory determine the success or failure of a wide variety of different cognitive tasks, such as problem solving, navigation, or planning. Estimation of constructs like working memory load or memory capacity from neurophysiological or psychophysiological signals would enable adaptive systems to respond to cognitive states experienced by an operator and trigger responses designed to support task performance (e.g. by simplifying the exercises of a tutor system when the subject is overloaded, or by shutting down distractions from the mobile phone). The determination of cognitive states like working memory load is also useful for automated testing/assessment or for usability evaluation. While there exists a large body of research work on neural and physiological correlates of cognitive functions like working memory activity, fewer publications deal witt the application of this research with respect to single-trial detection and real-time estimation of cognitive functions in complex, realistic scenarios. Single-trial classifiers based on brain activity measurements such as electroencephalography, functional near-infrared spectroscopy, physiological signals or eye tracking have the potential to classify affective or cognitive states based upon short segments of data. For this purpose, signal processing and machine learning techniques need to be developed and transferred to real-world user interfaces.

The goal of this Frontiers Research Topic was to advance the State-of-the-Art in signal-based modeling of cognitive processes. We were especially interested in research towards more complex and realistic study designs, for example collecting data in the wild or investigating the interaction between different cognitive processes or signal modalities. Bringing together many contributions in one format allowed us to look at the state of convergence or diversity regarding concepts, methods, and paradigms.

Table of Contents

frontiers
in Human Neuroscience

Editorial: Detection and Estimation of Working Memory States and Cognitive Functions Based on Neurophysiological Measures

Felix Putze[1]*, Christian Mühl[2], Fabien Lotte[3], Stephen Fairclough[4] and Christian Herff[1,5]

[1] *Cognitive Systems Lab, University of Bremen, Bremen, Germany,* [2] *Sleep and Human Factors Research, German Aerospace Center, Institute of Aerospace Medicine, Cologne, Germany,* [3] *Inria / LaBRI (CNRS - Bordeaux INP - Univ. Bordeaux), Talence, France,* [4] *School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool, United Kingdom,* [5] *School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands*

Editorial on the Research Topic

Detection and Estimation of Working Memory States and Cognitive Functions Based on Neurophysiological Measures

1. SCOPE

Executive cognitive functions like working memory determine the success or failure of a wide variety of different cognitive tasks, such as problem solving, navigation, or planning. Estimation of constructs like working memory load or memory capacity from neurophysiological or psychophysiological signals would enable adaptive systems to respond to cognitive states experienced by an operator and trigger responses designed to support task performance (e.g., by simplifying the exercises of a tutor system when the subject is overloaded Gerjets et al., 2014, or by shutting down distractions from the mobile phone). The determination of cognitive states like working memory load is also useful for automated testing/assessment or for usability evaluation. While there exists a large body of research work on neural and physiological correlates of cognitive functions like working memory activity, fewer publications deal with the application of this research with respect to single-trial detection and real-time estimation of cognitive functions in complex, realistic scenarios. Single-trial classifiers based on brain activity measurements such as electroencephalography (EEG, Kothe and Makeig, 2011; Lotte et al., 2018), functional near-infrared spectroscopy (fNIRS, Putze et al., 2014; Herff et al., 2015), physiological signals (Fairclough et al., 2005; Fairclough, 2008), or eye tracking (Putze et al., 2013) have the potential to classify affective (Koelstra et al., 2010; Heger et al., 2014; Mühl et al., 2014) or cognitive states based upon short segments of data. For this purpose, signal processing and machine learning techniques need to be developed and transferred to real-world user interfaces.

The goal of this Frontiers Research Topic was to advance the State-of-the-Art in signal-based modeling of cognitive processes. We were especially interested in research toward more complex and realistic study designs, for example collecting data in the wild or investigating the interaction

between different cognitive processes or signal modalities. Bringing together many contributions in one format allowed us to look at the state of convergence or diversity regarding concepts, methods, and paradigms.

The accepted manuscripts in this research topic cover a large range of aspects of cognition, reflecting the broadness of the field and its many application domains. A dominant challenge in the research topic is the analysis of cognitive workload (or memory load) from neurological signals. This does not come as a surprise because workload is a thoroughly studied construct and workload models can be immediately exploited, e.g., for adaptive human-machine interaction. While all these manuscripts share a joint research interest, they tackle the challenge of workload modeling in different application domains, with different signals, different classification approaches, and different features. Working memory and attentional control represent two recurring themes through the collection of papers in this research theme. The most prominent modality in this research topic is EEG, drawing from both spectral as well as time-domain features. In multiple articles, EEG is complemented by other modalities: Two of them use fNIRS as a different mode to capture neural activity (two others use fNIRS as single modality); two others use eye tracking as a way to capture visual attention and one also uses physiological signals (such as heart rate and breath rate). This shows that researchers today routinely select which (combinations of) signals are most promising for a given task.

2. HIGHLIGHTS

In the following paragraphs, we give an overview of the contributed manuscripts and their main contributions to the field. We structure them according to the cognitive processes which are in the respective focus of interest, according to the core themes identified in the previous section.

2.1. Cognitive Workload

Gateau et al. perform a study in which they investigated cognitive workload (induced by a serial memorization task in two difficulty levels) in pilots from prefrontal fNIRS. They show that binary classification of workload level was feasible both in flight simulator as well as in a real aircraft. Additionally, the authors investigated the differences between these two conditions and revealed that workload tends to be higher in the real aircraft. Workload classification performances were similar in both conditions though. Grissmann et al. investigated in a study how the simultaneous stimulation of multiple cognitive and affective states, namely workload and valence, influenced the respective neural markers in the EEG signal. They found that while correlates of workload (in frontal theta and parietal alpha power) could still be detected, these correlates were affected by negative valence. Moreover they found no evidence of typical neural correlates of affect. This has implications for many real-world applications for which different concurrent mental states can be expected. In a study with 17 healthy volunteers, Aghajani et al. investigated mental workload classification using EEG and fNIRS. Workload was induced using a letter n-back task and

could be classified using Support Vector machines in EEG, fNIRS, and the combination of both, which outperformed the individual modalities.

The impact of a multimodal approach for the classification of working memory load was also investigated by Liu et al.. These authors presented a study in which they discriminated three levels of mental workload, induced by an n-back task, using EEG, fNIRS, and physiological measures, with data combined over modalities performing best. They could show that all three modalities yielded better than chance level classification. To reduce calibration time for each subject, they demonstrated how results could be improved by learning from other subjects.

Extending offline analysis to a closed-loop system, Walter et al. demonstrate in their study that cognitive workload can be estimated from EEG to automatically adapt the difficulty level in a learning environment. The prediction model was trained on 10 subjects beforehand and no user-specific calibration data was needed, highlighting the feasibility in realistic scenarios. Labonte-Lemoyne et al. investigate a closed-loop system which adapted a game of Tetris to the cognitive load of the player, as measured from EEG and facial expression, with dynamic thresholding. In their user study, they show that adaptation to measured cognitive user states is not a trivial task and neither perceived experience, nor objective performance could be improved compared to the control condition.

2.2. Memory Processes

Noh et al. used ERP features from the EEG signal to discriminate between successful and unsuccessful retrieval of memorized items. In their study, they show that the presence or absence of contextual information interacts with the type of information to be encoded. Tian et al. conducted an EEG study to investigate neural markers for the discrimination of low and high performers in a repeated memory task. They found that fronto-central spectral entropy during the retention period of working memory was predictive for intra-session and inter-session differences in performance. Toppi et al. use graph-theoretic connectivity features to discriminate different phases of working memory processing in 60-channel EEG data from a study in which 17 participants performed a Sternberg task. The authors could identify topological differences between the encoding, storage, and retrieval phases and also found a relationship between the neural markers and the participants' behavior.

2.3. Attention

Riccio et al. worked with a P3-BCI and study how attentional processing differs between patients suffering from amyotrophic lateral sclerosis (ALS) and healthy users and whether and how this corresponds to differences in P3-BCI performance. The corresponding study investigates attentional processing via a rapid serial visual presentation task. Results revealed that ALS patients have degraded attentional performances which translates in smaller P3 amplitude and lower P3-BCI classification performances. Using EEG, Baldwin et al. investigated mind wandering during simulated driving. In

their study, they found increased alpha activity and reduced magnitude of P3a components to auditory stimuli in periods of mind wandering. As mind wandering during driving is a potential safety threat, this study can have large impact in future adaptive interfaces. Brouwer et al. presented a study combining EEG and eye tracking to investigate target encoding during visual search. They could show that eye tracking features distinguished better between hits and misses (i.e. targets which are never reported by the participant), while EEG features differentiated better between targets and non-targets. These results highlight how the two modalities complement each other.

2.4. Methods & Other Cognitive Processes

Verdière et al. measure engagement of pilots via fNIRS in a flight simulator. In their study, different levels of engagement are induced via manual vs. automatic landing. The authors propose connectivity-based features and show that these can be used for effective classification. Takayoshi et al. used rewards in a simple number-discrimination task to study motivation and apathy tendency. In their study, they found P2 and P3 ERPs in EEG to be modulated by reward dependence and apathy tendency, which is an important step toward an adaptive interface based on motivation. Touryan et al. used a simultaneous visual search tasks and auditory n-back to investigate the neural sources of target detection in the presence of eye movements and during simultaneous task-demand. In their study, the authors used a modified measure-projection approach to separate the neural sources into six regions of interest (ROIs). This enabled efficient target detection from EEG signals, as well as revealing the time course of EEG activity in these ROIs—thus leading to a better understanding of the underlying processes. Nicolae et al. investigate a relatively novel concept by studying depth of processing (on three levels: no processing, shallow processing, or deep processing) from event-related time-domain and frequency domain features in EEG on a single-trial basis. In their study, they demonstrate the feasibility of

this approach in three domains (namely: memory, language, and visual imagination).

3. SUMMARY

The presented articles tackle a large variety of cognitive processes, their neurophysiological correlates and their detection. This research shows great progress in basic understanding of cognition as well as for applied research, especially for the development of adaptive systems. They also highlighted the need for more research into realistic and ecologically valid application environments of cognitive state estimation. If such states could be reliably estimated in real application scenarios–outside the lab—which is not really the case so far—then this could prove very useful for multiple applications such as aeronautics, driving, education or video games, among many others. Future directions of research should also include thorough attempts to define and differentiate the modeled concepts (e.g., how can we discriminate "mental workload" from "memory load") and find a consensus on publicly available data sets for joint evaluations. A role model here could be the field of affective computing. Available common data sets such as DEAP (Koelstra et al., 2012), fNIRS-nback (Herff et al., 2014), EEG (So et al., 2017), or EEG combined with fNIRS (Shin et al., 2017) could be a start for such a program.

AUTHOR CONTRIBUTIONS

FP coordinated the writing process and proposed the structure of the article. All authors contributed to the writing process.

ACKNOWLEDGMENTS

REFERENCES

Fairclough, S. H. (2008). Fundamentals of physiological computing. *Interact. Comput.* 21, 133–145. doi: 10.1016/j.intcom.2008.10.011

Fairclough, S. H., Venables, L., and Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. *Int. J. Psychophysiol.* 56, 171–184. doi: 10.1016/j.ijpsycho.2004.11.003

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Front. Neurosci.* 8:385. doi: 10.3389/fnins.2014.00385

Heger, D., Herff, C., Putze, F., Mutter, R., and Schultz, T. (2014). Continuous affective states recognition using functional near infrared spectroscopy. *Brain Computer Interfaces* 1, 113–125. doi: 10.1080/2326263X.2014.912884

Herff, C., Fortmann, O., Tse, C.-Y., Cheng, X., Putze, F., Heger, D., et al. (2015). "Hybrid fNIRS-EEG based discrimination of 5 levels of memory load," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)* (Montpellier: IEEE), 5–8.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task quantified in the prefrontal cortex using fnirs. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J.-S., Nijholt, A., et al. (2010). "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *International Conference on Brain Informatics* (Toronto, ON: Springer), 89–100.

Kothe, C. A., and Makeig, S. (2011). "Estimation of task workload from EEG data: new and current tools and perspectives," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (Boston, MA: IEEE), 6547–6551.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Mühl, C., Jeunet, C., and Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Front. Neurosci.* 8:114. doi: 10.3389/fnins.2014.00114

Putze, F., Hesslinger, S., Tse, C.-Y., Huang, Y., Herff, C., Guan, C., et al. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Front. Neurosci.* 8:373. doi: 10.3389/fnins.2014.00373

Putze, F., Hild, J., Kärgel, R., Herff, C., Redmann, A., Beyerer, J., et al. (2013). "Locating user attention using eye tracking and EEG for spatio-temporal event selection," in *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, CA: ACM), 129–136.

Shin, J., von Lühmann, A., Blankertz, B., Kim, D.-W., Jeong, J., Hwang, H. -J., et al. (2017). Open access dataset for EEG+NIRS single-trial classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1735–1745. doi: 10.1109/TNSRE.2016.2628057

So, W. K., Wong, S. W., Mak, J. N., and Chan, R. H. (2017). An evaluation of mental workload with frontal EEG. *PLoS ONE* 12:e0174949. doi: 10.1371/journal.pone.0174949

frontiers
in Human Neuroscience

In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI

Thibault Gateau[1], Hasan Ayaz[2] and Frédéric Dehais[1]*

[1] ISAE-SUPAERO, Institut Supérieur de l'Aéronautique et de l'Espace, Université Fédérale de Midi-Pyrénées, Toulouse , France, [2] School of Biomedical Engineering, Science Health Systems, Drexel University, Philadelphia, PA, United States

There is growing interest for implementing tools to monitor cognitive performance in naturalistic work and everyday life settings. The emerging field of research, known as neuroergonomics, promotes the use of wearable and portable brain monitoring sensors such as functional near infrared spectroscopy (fNIRS) to investigate cortical activity in a variety of human tasks out of the laboratory. The objective of this study was to implement an on-line passive fNIRS-based brain computer interface to discriminate two levels of working memory load during highly ecological aircraft piloting tasks. Twenty eight recruited pilots were equally split into two groups (flight simulator vs. real aircraft). In both cases, identical approaches and experimental stimuli were used (serial memorization task, consisting in repeating series of pre-recorded air traffic control instructions, easy vs. hard). The results show pilots in the real flight condition committed more errors and had higher anterior prefrontal cortex activation than pilots in the simulator, when completing cognitively demanding tasks. Nevertheless, evaluation of single trial working memory load classification showed high accuracy (> 76%) across both experimental conditions. The contributions here are two-fold. First, we demonstrate the feasibility of passively monitoring cognitive load in a realistic and complex situation (live piloting of an aircraft). In addition, the differences in performance and brain activity between the two experimental conditions underscore the need for ecologically-valid investigations.

Keywords: fNIRS, BCI, working memory, prefrontal cortex, simulated and real flight, neuroergonomics

1. INTRODUCTION

Neuroergonomics is an emerging field of interdisciplinary research that promotes the understanding of the brain in complex real-life activities. This approach merges knowledge and methods from cognitive psychology, system engineering, and neuroscience (Parasuraman and Wilson, 2008). Accurate and reliable mental state assessment of human operators during use of complex systems is a prime goal of neuroergonomics that aims to measure the "brain at work" (Parasuraman and Rizzo, 2008). Understanding the underlying neurocognitive processes of such interaction could be used to improve safety and efficiency of the overall human-machine pairing. This could be achieved by (i) the augmentation of human performance and its translation to

improved functioning "at work", (ii) informing the design of the complex systems, or (iii) adapting the user interface and task parameters dynamically during use.

Aviation operations constitute an ideal paradigm to implement this approach. Pilots deal with an uncertain environment and face complex interaction with the flightdeck (Causse et al., 2013; Çakır et al., 2016; Reynal et al., 2016). For instance, several studies have emphasized that pilots' working memory (WM) abilities are heavily recruited to handle flightpath, to monitor the flight parameters, and to maintain an up-to-date situation awareness (Causse et al., 2011a,b). WM is also an important component when following air traffic control (ATC) instructions (Morrow et al., 1993). This activity indeed requires mentally storing flight parameters (e.g., heading, altitude, speed) to follow the adequate flight path. However, it is well-known that human working memory is fundamentally limited (Baddeley, 1992; Miller, 1994) and easily overwhelmed when task demand is excessive (Durantin et al., 2014a). Human factor studies emphasized that a variety of environmental stressors may negatively impact pilots' ability to execute ATC clearances (Billings and Cheaney, 1981; Taylor et al., 1994, 2005; Scerbo et al., 2003; Risser et al., 2006; Rome et al., 2012; Dehais et al., 2017). Thus, the implementation of monitoring technology in the cockpit to infer a state of cognitive limitation could represent a promising approach to enhance flight safety (Roy et al., 2017; Verdière et al., 2018).

Indeed, the development of brain computer interface (BCI) technology provides interesting prospects to continuously monitor and take advantage of the brain dynamics and the neural mechanisms underlying cognition. Among the three categories of BCIs (active, reactive, and passive) (Zander and Kothe, 2011; Vecchiato et al., 2016), the first two types are aimed at transforming cerebral activity into messages or commands to voluntarily control distant apparatus (e.g., mouse cursor). Passive BCIs (pBCI) are of particular interest for neuroergonomic applications (Cutrell and Tan, 2008; Frey et al., 2017; Gramann et al., 2017). They allow the use of interpretation of unlabeled brain activity during a task to derive various mental states (Blankertz et al., 2010; Roy et al., 2013; Van Erp et al., 2015; Zander et al., 2017). These mental state-inference systems offer a unique insight into the development of the human-system interactions to overcome cognitive limitations (Zander and Kothe, 2011; Brouwer et al., 2013). While several pBCIs have been successfully implemented in driving (Dijksterhuis et al., 2013) and flight simulator (Gateau et al., 2015; Aricò et al., 2016; Çakır et al., 2016; Callan et al., 2016; Verdière et al., 2018), few have attempted to test these systems under more realistic settings. However, very few studies have attempted to test these adaptive systems under realistic settings (Callan et al., 2015).

Electroencephalography (EEG) is by far the most popular technique (George and Lécuyer, 2010; Borghini et al., 2017) in the BCI community as it has excellent qualities for monitoring cognitive states (Brouwer et al., 2012; Roy et al., 2013) including superior temporal resolution and has been used to monitor working memory (Roy et al., 2013; Mühl et al., 2014). However, the localization of sources from the EEG signal requires higher-density recordings and additional computation to solve the

inverse problem that may not be amenable to critical operational situations such as flying real aircraft. In addition, setup time and susceptibility to motion artifacts should be considered for minimally intrusive deployment. Thus, the use of functional near infrared spectroscopy (fNIRS) has been gaining popularity recently as the sensors have been miniaturized, become portable and wireless (Ayaz et al., 2013; Strait et al., 2014; Naseer and Hong, 2015; Schudlo and Chau, 2015).

This brain activity monitoring technique uses near-infrared light absorption properties of hemoglobin to estimate local variations of cortical oxygenation changes (Villringer and Obrig, 2002; Ayaz et al., 2012). fNIRS has been successfully used to detect working memory solicitation (Li et al., 2005; Schreppel et al., 2008; Hirshfield et al., 2011; Gagnon et al., 2012; Herff et al., 2014; McKendrick et al., 2014; León-Domínguez et al., 2015; Unni et al., 2017). Despite its relatively low temporal resolution, fNIRS poses several advantages compared to more established traditional tools (Kikukawa et al., 2008; Piper et al., 2014; McKendrick et al., 2015; Davranche et al., 2016) such a relatively high spatial resolution (around $1\,cm^2$ depending on the sensor geometry) and high signal-to-noise ratio as sensors are relatively more robust to motion artifacts (Huppert et al., 2009), eye-blinks and facial muscles (Izzetoglu et al., 2004). It is also possible to run experiments with active and mobile subjects and even outdoors (Piper et al., 2014; McKendrick et al., 2016). Specifically, it is less sensitive to noisy electromagnetic environment in the aircraft (radio transmission, radio-navigation beacons, GPS antenna, etc.) than EEG, making it a candidate to measure pilot's brain activity during real flight. As an emerging neuroimaging technique, we believe that it is important to investigate the capabilities of fNIRS and its utility in future applications.

The present study aims to develop an on-line fNIRS based pBCI for the assessment of working memory of aircraft pilot during real flight. Earlier studies demonstrate that fNIRS based measures BCI have been implemented. They rely on oxygenation changes in the prefrontal cortex (PFC) and can be used for measuring WM load (Schreppel et al., 2008; Ayaz et al., 2012; Gagnon et al., 2012; Durantin et al., 2014a,b). Here, a pilot-ATC interaction task, was designed with two contrasted levels of WM load. A Support Vector Machine (SVM) based classifier performing on-line for single trial WM load level discrimination was implemented. This classifier was first tested in a high fidelity flight simulator. The same machine learning approach was then utilized to assess the WM load level in an actual flight condition. To the authors' knowledge, this is the first study to monitor pilot's brain activity on-line under such operational settings and ecological validity. We also compared pilot's WM performance and related PFC activity both in high fidelity simulator and real flight conditions. The objective was to determine wether these two conditions simulated and real operational settings were equivalent or not in terms of task demand (Dahlstrom and Nahlinder, 2009; Batula et al., 2017). As most aviation psychology experiments and pilots' training are conducted with flight simulators, such assessment is critical for future design and development of such approaches (Philip et al., 2005).

2. MATERIALS AND METHODS

2.1. Passive BCI in Flight Simulator

2.1.1. Participants

Fourteen visual flight rules (VFR) pilots (three women; mean group age: 29.25 ± 6.92; mean flight hours 80 ± 50) completed the experiment. Pilots had normal or corrected-to-normal vision, normal hearing, and no psychiatric disorders. They all had medical clearance to fly. After providing written informed consent, they were instructed to complete task training. The data from two participants were rejected due to a high level of fatigue in one case, and data collection issue for the second. Typical total duration of a subject's session (informed consent approval, practice task, and real task) was about two hours. This work was approved by the Institutional Review Board (IRB) of the Inserm Committee of Ethics Evaluation (CEEI: Comité d'Evaluation Ethique de l'Inserm IRB00003888). The methods were carried out in accordance with approved guidelines and participants gave written informed consent approved by the IRB of CEEI.

2.1.2. Neurophysiological Measurements: fNIRS

During this experiment, we recorded hemodynamics of the prefrontal cortex using the functional near-infrared spectrometer fNIR Device Model 100B (Biopac®) equipped with 16 optodes (**Figure 1**). On this continuous-wave system, the optode separation was about 25 mm and two wavelengths were used, 730 and 850 nm. DPF (differential pathlength factor) value was 5.97 which is within the range used by many in literature (Kato et al., 1993; Luo et al., 2002) and accepted by many groups. Four regions of interest (ROI) were defined to allow for explorative statistical comparisons with the data collected during the real flight experiment (see section 3).

Each optode of the device records hemodynamics at a frequency of 2 Hz in terms of oxygenation level variations in comparison to an initial baseline performed prior to the experiment. Changes in the concentrations of oxygenated ($\Delta[HbO_2]$) and deoxygenated hemoglobin ($\Delta[hHb]$) relative to the baseline can be calculated from changes in detected light intensity using the modified Beer-Lambert Law (Delpy et al., 1988). Cognitive Optical Brain Imaging (COBI) Studio® software (Ayaz and Onaral, 2005; Ayaz et al., 2011) was used to collect data. The data stream was available on-line from a TCP/IP interface. Before recording, signals for each optode were carefully checked for saturation with COBI Studio which provides signal quality visual representation. COBI studio was also used to check signal quality and to adjust consequently the headband on the participant's forehead. After this check, a baseline was established, which simply consists of letting the participant rest for 10 s.

2.1.3. Experimental Environment: Flight Simulator

We used the ISAE-SUPAERO (Institut Supérieur de l'Aéronautique et de l'Espace - French Aeronautical University in Toulouse, France) flight simulator to conduct the experiment in an ecological situation. Its user interface is composed of a Primary Flight Display, a Navigation Display, and an Electronic Central Aircraft Monitoring Display (**Figure 2**).

2.1.4. Task Description

This protocol was adapted from a previous study (Gateau et al., 2015). As in real flight operations, pilots heard ATC instructions (pre-recorded for this experiment) to vector them and were asked to read back the instructions. Their answers were recorded for off-line behavioral analysis. The ATC messages were delivered at 78 dB through a Sennheiser® headset. Two levels of difficulty were defined based on the flight parameters that the participant had to read back during the experiment:

- Low WM load: The two first digits were the same for each flight parameter (e.g., 14 for "speed 140, heading 140, altitude 1400, vertical speed +1400").
- High WM load: each flight parameter value was different from the previous one and composed of different number to increase task difficulty (e.g., "speed 172, heading 238, altitude 6400, vertical speed −2800").

The task consisted of 10 repetitions of each difficulty for a total of 20 trials. The task difficulty order was randomly distributed with two constraints:

- the first 10 trials contained both 5 trials of high difficulty, and 5 trials of low difficulty (which is necessary for machine learning purposes, see section 2.1.5);
- the difficulty cannot be the same for more than two successive trials.

Each ATC message started with the airplane call sign (i.e., "Supaero 32"), immediately followed by a sequence of flight parameters and ended with the message "over" (**Figure 3**). Thereafter, pilots had a 18 s response window to repeat the instruction. A practice session was conducted prior to the experiment runs to familiarize them with the experiment protocol and the interface.

During the experiment, the experimenter was collecting the volunteer's ability to read back each message so as to compute the total number of correct responses in the low and hard conditions.

2.1.5. Experimental Time Course

For machine learning purposes, the experiment was split into three successive phases (**Figure 4**):

- Phase A – data gathering phase: 10 instructions with two levels of difficulty were successively presented to the pilot in a random order. During phase A, the correctness of the pilot's response was also checked for further pilot performance analysis. The fNIRS's data were processed and recorded for each trial's response window. The levels of difficulty of the message were also recorded.
- Phase B – classifier training phase: the classifier training process was activated, based on the data gathered during phase A. This phase was not perceived by the pilot and allowed further classification actions. At the end of this phase, the pilot's classifier - the pilot's specific classification model, correctly trained - was available for classification requests.
- Phase C – classifier testing phase: 10 instructions with random levels of difficulty (high WM load or low WM load) were successively presented. The aim of the classification process
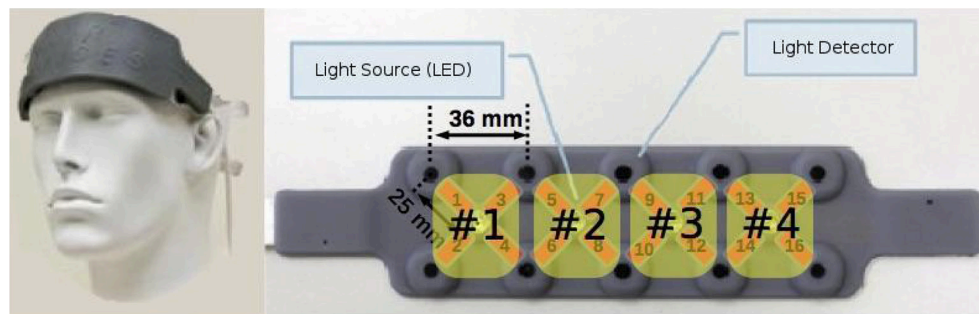
FIGURE 1 | fNIR Device® Model 1200S headband and associated optode numbering. Only the four closest detectors to an emitter constituted optodes. Optodes are represented in red with their associated number. Four regions of interest (ROI) were defined for statistical analyses purposes (#1, #2, #3, #4).
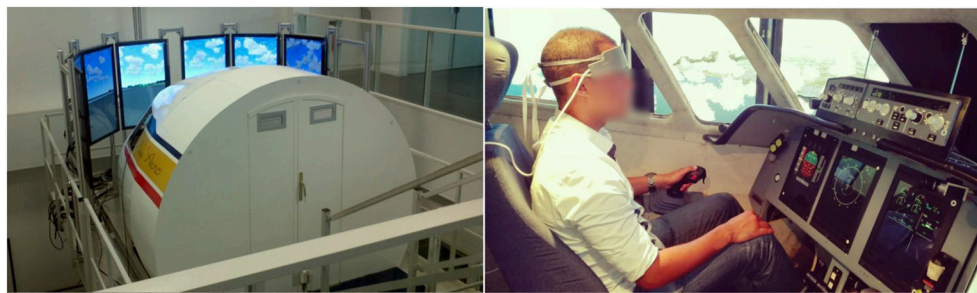
FIGURE 2 | (Left) ISAE-Supaero flight simulator; the closed cabin is visible from rear, and eight screens are used to visualize external environment. (Right) The pilot subject with the fNIRS headband.
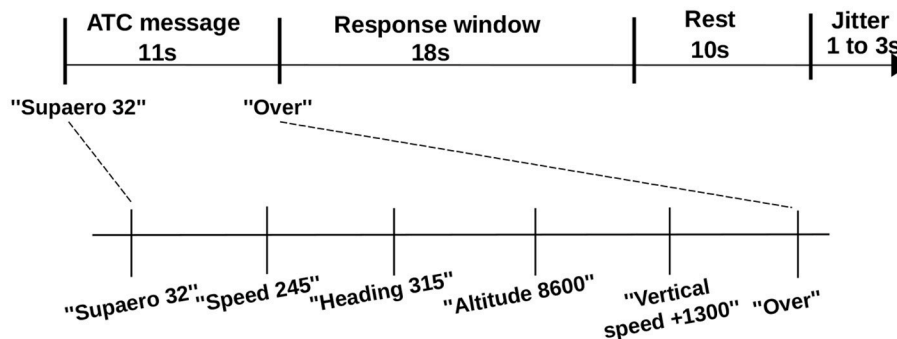
FIGURE 3 | ATC span task trial design (High Load message).

was to discriminate the difficulty of the trial. After each response window of trials, the classifier returned WM load estimation of the trial.

Note that the transition (phase B) from phase A to phase C was not perceptible to the participants.

2.1.6. MACD Filter

Raw fNIRS data were real-time filtered using a MACD filter, commonly used in economic market analysis (Appel, 2005). This filter, based on the difference between a short-term EMA (Exponential Moving Average) and a long-term EMA, implements a second order band-pass filtering to eliminate low-frequency (<0.02 Hz) and high-frequency (>0.33 Hz) components from the raw fNIRS signal (Utsugi et al., 2007). This low order filter has a quasi linear phase in its bandwith and is particularly suited for real-time applications. For the experiment, we proceeded to an on-line filtering of $\Delta[HbO_2]$ and $\Delta[hHb]$ on 16 optodes.

N represents the number of time points defining the EMA window:

$$y = EMA_N(x) \Leftrightarrow y_n = \frac{2}{N+1}x_n + \frac{N-1}{N+1}y_{n-1} \quad (1)$$

$$MACD_{N_{short}, N_{long}}(x) = EMA_{N_{short}}(x) - EMA_{N_{long}}(x) \quad (2)$$
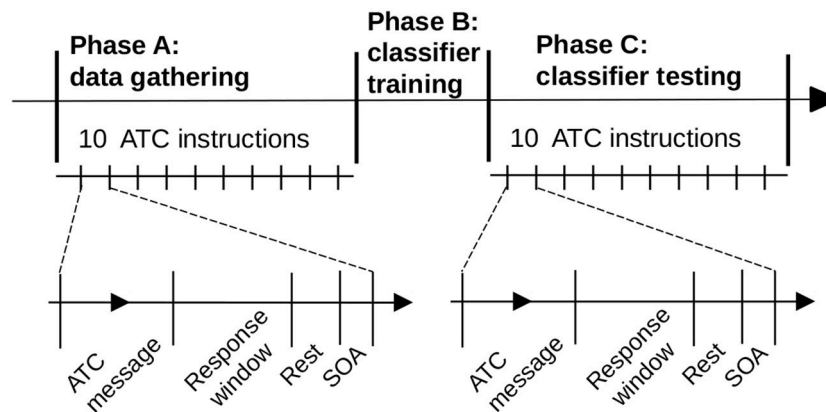
FIGURE 4 | The experiment was split into three successive phases. Data gathering (*phase A*) and classifier testing (*phase C*) consisted of 10 ATC instructions each. The pilot's classifier was trained between these two phases (*phase B*). The time scale of the figure is illustrative.

We chose a 6 s short-term EMA and a 13 s long-term EMA according to previous work (Durantin et al., 2014b) for MACD filtering, to get the desired bandwidth.

2.1.7. Single Trial SVM-Based WM Load Estimation

The classification's goal was to discriminate on-line whether the last trial was a high WM load trial or a low WM load trial. For each pilot, we used the first 10 trials to train the pilot's classifier (phase A and B, see section 2.1.5). From trial 11 to 20, we used the pilot's classifier to discriminate trial difficulty, without any further training. An accuracy score (sum of correct predictions divided by total number of predictions) of the pilot's classifier was provided at the end of the experimental session.

Sixteen optodes of $\Delta[HbO_2]$ and $\Delta[hHb]$ filtered signals were segmented into trials, in real-time, according to the task synchronization module (**Figure 5**). Each trial starts when an ATC message is played, and lasts 30 s. All data points of a trial – 2 different inputs per optode, 16 optodes, 30 s of data with a 2 Hz sampling corresponding to 1920 features – were considered as the input of the machine learning process. A 30 s sliding window was chosen to be consistent with literature regarding inter-subject variability (Jasdzewski et al., 2003; Sato et al., 2005). Note that the transition from the "Response" phase to the "Rest" one was unnoticeable, as it was anticipated that participants started to rest as soon as they completed the instruction.

As our number of features was large compared to the training sample, we used a linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995). The principle of the SVM is to find the separating hyperplane that maximizes the distance between the hyperplane and the closest training points in each class. To avoid over-fitting, we chose to customize the SVM regularization parameter for each pilot's classifier. In a linear SVM, the regularization parameter C controls the trade-off between errors of the SVM on training data and margin maximization. During the training process of each participant, the parameter C is incrementally changed over a large range of values (from 10^{-3} to 10^4) with a 10-step factor.

Hence, a five-fold cross-validation on the first 10 trials with scikit-learn (Pedregosa et al., 2011) packages (sklearn.svm and sklearn.cross_validation) was ran to select the C parameter with the highest performance in terms of accuracy. The classifier training (phase B) was performed as soon as the data of the first 10 trials were available for online purposes (Aricò et al., 2016).

2.1.8. Experimental Components' Architecture

We implemented a WM load estimator that integrated different components (**Figure 5**):

- a simulated ATC which broadcasts a list of chosen messages to the pilot;
- the ISAE flight simulator (**Figure 2**);
- a fNIR sensor which measures the prefrontal oxygenation (**Figure 1**);
- a MACD filter for artifact removal (see section 2.1.6);
- a synchronization module that also formats filtered data for the classification process: filtered fNIRS output must be synchronized with the pilot's state, according to the instant of the arrival of that incoming message and according to the pilot's response window;
- a classifier (see section 2.1.7) which evaluates in real-time whether the last ATC instruction was a high WM load trial or a low WM load trial. Results were logged into a file, while a real time feedback is provided through a system terminal.

Task monitoring, data acquisition, and computation were conducted on the same computer (core i5-3210M, 2.50 GHz, 4 GB RAM). During the experiment, the classifier training (phase B) duration was short (800 ms) and remained unnoticeable for the participant. The classifier testing phase lasted 10 ms and was also unnoticeable for the participant" (**Figure 6**).

2.2. Passive BCI in Real Flight
2.2.1. Participants

Fourteen VFR pilots (1 women; mean group age: 23.07 ± 5.35; mean flight hours 44.07 ± 37.52), completed the
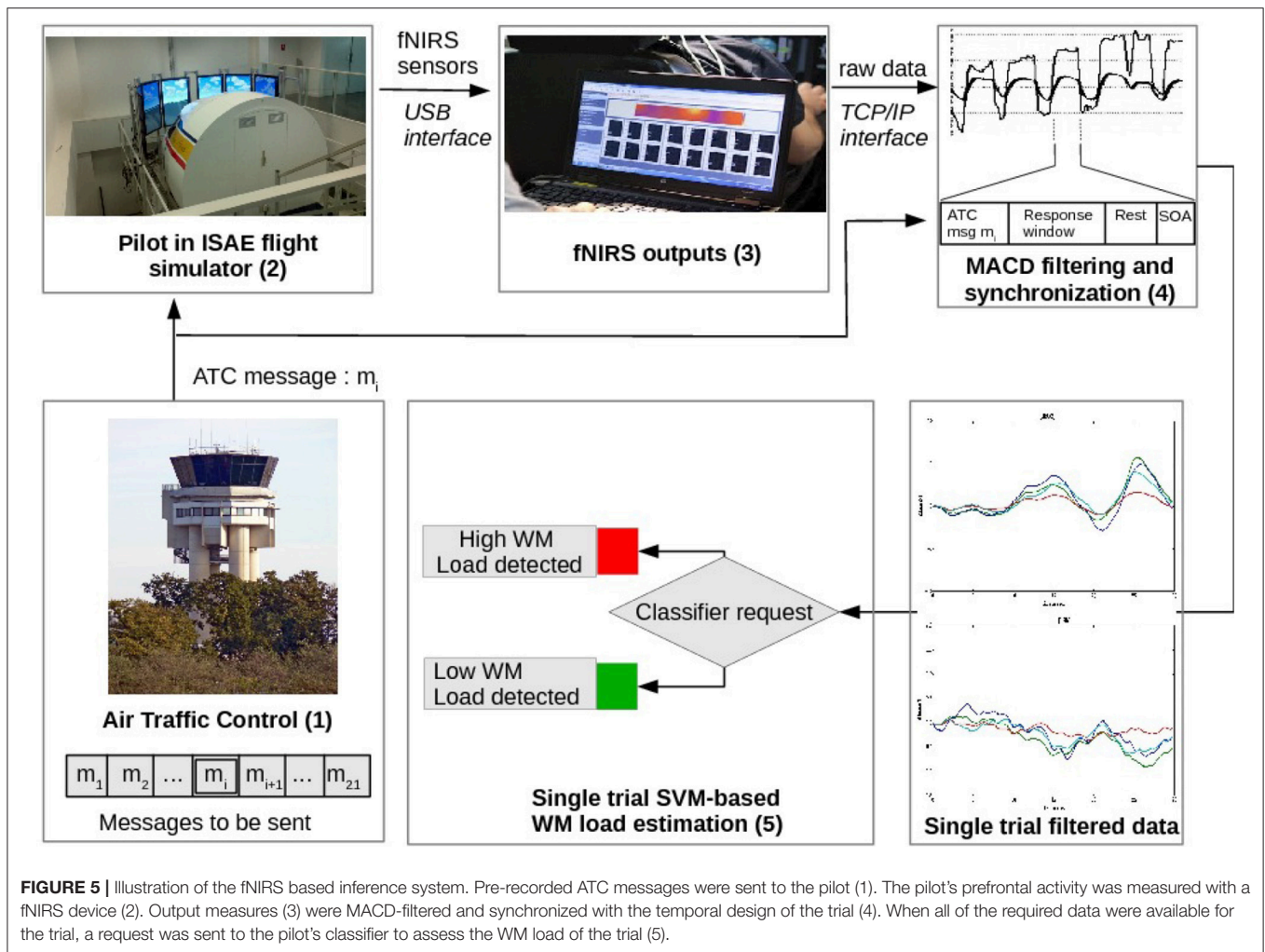
FIGURE 5 | Illustration of the fNIRS based inference system. Pre-recorded ATC messages were sent to the pilot (1). The pilot's prefrontal activity was measured with a fNIRS device (2). Output measures (3) were MACD-filtered and synchronized with the temporal design of the trial (4). When all of the required data were available for the trial, a request was sent to the pilot's classifier to assess the WM load of the trial (5).
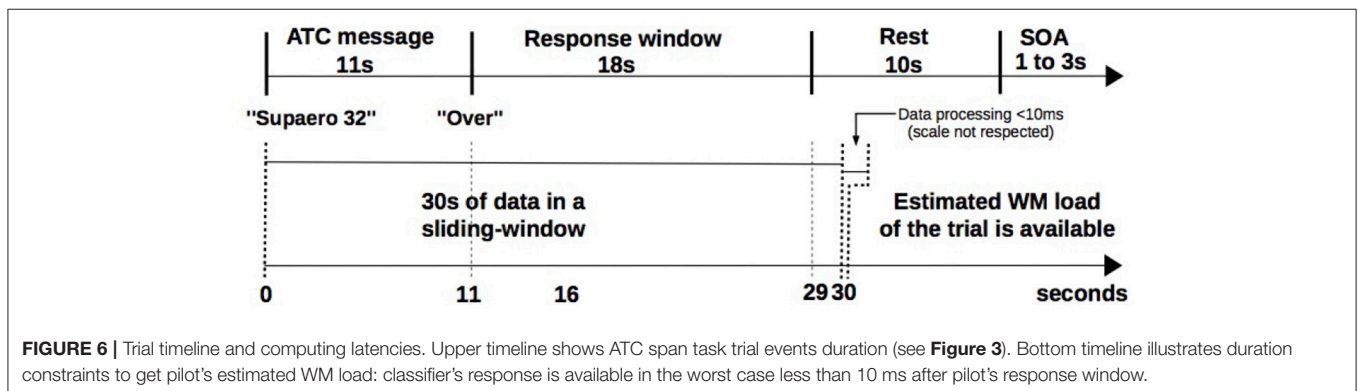
FIGURE 6 | Trial timeline and computing latencies. Upper timeline shows ATC span task trial events duration (see **Figure 3**). Bottom timeline illustrates duration constraints to get pilot's estimated WM load: classifier's response is available in the worst case less than 10 ms after pilot's response window.

experiment. Please note that these volunteers were different from the ones who participated to the flight simulator experiment. The data from two participants were rejected due to light saturation issues and a device synchronization issue. After providing informed consent, they were instructed to complete task training on the ground. None of the recruited subjects had neurological or psychiatric history or was on medication. Each of them gave written informed consent for the experiment. The experimental protocol was approved by the committee of the European Aviation Safety Agency (EASA permit to fly approval number : 60049235). The methods were carried out in accordance with approved guidelines and participants gave written informed consent approved by the EASA.

2.2.2. Neurophysiological Measurements: Mini-fNIRS

We used the miniaturized and wireless fNIR Device Model 1200W (Biopac®) portable system (Ayaz et al., 2013) to record the pilots' hemodynamics of the prefrontal cortex (**Figure 7**). This device was chosen as it was wireless (i.e., the pilot's head was not attached to any cables) and did not require external power supply as the Model 1200S. This was a prerequisite to facilitate its implementation and use in the aircraft for our experiment. This device had the same hardware design, and exactly same LED light source components and detectors than the fNIRS Model 1200S used in the flight simulator. Consistent with the previous device, on this continuous-wave system, the optode separation was about 25 mm and two wavelengths were used, 730 and 850 nm. The DPF value was 5.97. This four-optode device records hemodynamics at a frequency of 4 Hz in terms of oxygenation level variations in comparison to a baseline same as the 1200S desktop version. With flexible circuit board and separation-adjustable split pads, the sensors were positioned to aim monitoring brain areas similar to the ROIs extracted from 1200S sensor. Changes in the concentrations of oxygenated ($\Delta[HbO_2]$) and deoxygenated hemoglobin ($\Delta[hHb]$) can be calculated from changes in detected light intensity using the modified Beer-Lambert Law (Delpy et al., 1988).

Cognitive Optical Brain Imaging (COBI) Studio® software (Ayaz and Onaral, 2005; Ayaz et al., 2011) was used to collect data. The data stream was available on-line from a TCP/IP interface. Before recording, signals for each optode was carefully checked for saturation with COBI Studio which provided a visual representation of signal quality. An aluminum foil attached to a dark ski band band and a cap were placed over the mini-fNIRS to shield against ambient sunlight infrared.

Data was MACD filtered and we used a similar on-line Experimental Components' Architecture with the exception that we used a real plane instead of the flight simulator.

2.2.3. Experimental Environment: DR400 Aircraft

The ISAE Supaero DR400 light aircraft was used for the purpose of the experiment (**Figure 8**). It was powered by a 180HP Lycoming engine and was equipped with classical gauges, radio and radio navigation equipment, and actuators such as rudder, stick, thrust, and switches to control the flight. The participant was placed on the left seat and was equipped with the mini fnirs system. The participant wore a Clarify Aloft® that was used to trigger task-related auditory stimuli from a PC via an audio cable. The participant could still communicate with the other crew members, real ATC and when he received auditory stimuli (emulated ATC).The safety pilot was an ISAE flight instructor. He was right seated and had the authority to stopping the experiment and taking over the control of the aircraft for any safety reason. The backseater was the experimenter: his role was to set the sensor, to trigger the experimental scenario and to supervise data collection.

2.2.4. Task Description

The experimental task and audio messages were similar to the previous protocol (see section 2.1), with the same experimental time course and the same instructions for the participant.

A practice session on the ground was conducted prior to the experiment runs to familiarize them with the experiment protocol and the interface. After training was completed on the ground, the mini-fNIRS system was placed over the participants forehead. The participant then took off from Lasbordes (LFCL, Toulouse, France) airfield and began a local flight. The experimental task *per se* started when the pilot left the Lasbordes traffic pattern and was stabilized at an altitude of 2500 feet. The participant was asked to fly as straight and stable as possible and to only perform slight avoidance maneuvers as necessary. Once stabilized, the baseline of ten seconds was recorded. After the completion of the WM task, the participant was heading back to land at Lasbordes airfield. The total flight lasted one hour including the WM task.

As in simulated condition, the backseater was collecting the volunteer's ability to read back each message in order to compute the total number of correct responses in the low and hard conditions. These data allowed to compare the WM peformance accross the conditions (i.e., low vs. high; simulated vs. real flight).

2.2.5. Experimental Components' Architecture and WM Load Estimation

We implemented a similar WM load estimator in the airplane as in the flight simulator. Machine learning inputs were lightly adjusted to fit the data flow available with the mini-fNIRS wireless portable device. The four (instead of 16) available optodes of $\Delta[HbO_2]$ and $\Delta[hHb]$ filtered signals were segmented into trials, in real-time, according to the task synchronization module (see section 2.1.8). Each trial starts when an ATC message is played, and lasts 30 s. All data points of a trial - two different inputs per optode (i.e., $\Delta[HbO_2]$ and $\Delta[hHb]$), four optodes, 30 s of data with a 4 Hz sampling corresponding to 960 features - were considered as the input of the machine learning process.

2.3. Statistical Analyses

Off-line statistical analyses were performed with "R" (R Core Team, 2013) software and the "EzANOVA" (Anderson, 2001) package to compare WM performance and prefrontal cortex activations in the flight simulator and in the real flight conditions during the 20 trials. Two-tailed unpaired t-tests were performed to compare the WM performance in the high and low load conditions across the two flight conditions (simulator and real flight). As the number of optodes was not equivalent between the two fNIRS devices (16 vs. 4), we defined four regions of interests (ROIs) for the fNIR100 device that was used in the simulator condition to allow for explorative comparisons with the real flight condition. ROI1, ROI2, ROI3, and ROI4 were derived respectively from the spatial averaging of optodes 1 to 4, 5 to 8, 9 to 12, and 13 to 16 (see **Figure 1**). The mean frontal $\Delta[HbO_2]$ peak response and the mean frontal $\Delta[hHb]$ peak response (peak value within 30 s post-trial onset minus 2 s average pre-trial onset) over the four ROIs of the PFC for each trial and each pilot using the MACD-filtered data in both flight conditions (i.e., simulator and real flight) were computed. A multivariate analyses for repeated measures (MANOVA) was conducted over the mean $\Delta[HbO_2]$ data with between factor flight condition (simulator vs. real flight) and within subject
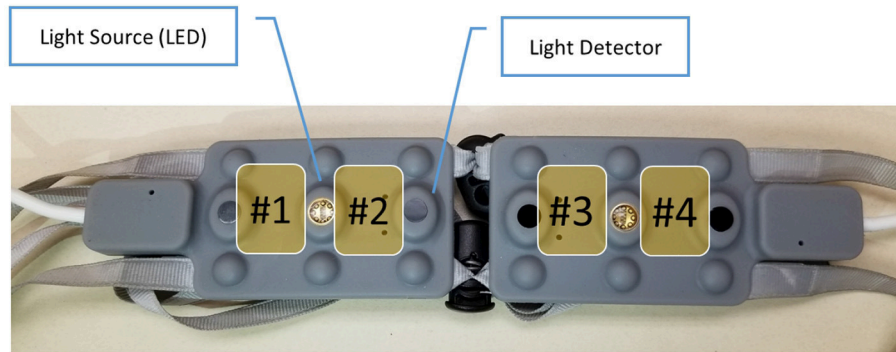
FIGURE 7 | Miniaturized and wireless fNIR Device® Model 1200W headband and associated regions of interest (optodes). Optodes are represented in yellow with their associated number.

FIGURE 8 | (Left) ISAE Supaero DR400 plane used for the experiment. **(Right)** The pilot subject with the mini-fNIRS headband (on the left) and the safety pilot (on the right).

factors WM Load (High vs. Low) and ROI (#1, #2, #3 & #4); see **Figure 1** was led. A similar MANOVA over the mean $\Delta[hHb]$ was then conducted. We then ran a two-tailed unpaired t-test to compare the classification accuracy in the two experimental conditions. The Tukey's Honestly Significant Difference (HSD) test was used for all *post-hoc* comparisons.

3. RESULTS

3.1. Real Flight vs. Flight Simulator: Off-Line Behavioral and Neurophysiological Analyses

Participants committed on average 5.33 errors ($SD = 1.95$) in the WM task in the simulator condition and on average 8.25 ($SD = 2.42$) errors in the real flight condition, all occurring during the high load trials (see **Figure 9**). As no error was committed in the low WM load condition, we performed a statistical analysis to compare the effect of the flight conditions on WM performance in the high load conditions. An unpaired t-test revealed that the real flight condition led to significantly higher number of errors on the WM task in the high load condition ($p < 0.001$, *Cohen's d* = 1.34). The MANOVA over the mean $\Delta[HbO_2]$ data disclosed a significant WM load × Flight condition ×

ROI interaction [$F_{(3,66)} = 3.36$; $p = 0.039$; see **Figure 10**]. *Post-hoc* analyses revealed that high load trials performed in real flight condition led to higher $\Delta[HbO_2]$ in ROI #2 than their counterparts performed in simulator ($p = 0.0001$). The MANOVA over the mean $\Delta[hHb]$ data did not disclose any significant WM load × Flight condition × ROI interaction [$F_{(3,66)} = 0.69$; $p = 0.56$].

3.2. Single Trial SVM-Based WM Load Estimation Results

3.2.1. Simulator

During the testing phase, a mean of 76.66% ($SD : 16.14\%$) of the trials were accurately classified (discriminated into on-line low WM load trials and high WM load trials). We obtained a 85.60% mean precision ($SD : 19.36\%$) and a 73.33% mean recall ($SD = 24.62\%$). Individual classifiers' accuracies are shown in **Table 1**.

3.2.2. Real Flight

During the testing phase, a mean of 78.33% ($SD : 11.93\%$) of the trials were accurately classified (discriminated into on-line low WM load trials and high WM load trials). We obtained a 84.14% mean precision ($SD : 18.56\%$) and a 76.67% mean recall

(SD = 22.29%). Individual classifiers' accuracies are shown in **Table 2**.

3.2.3. Real Flight vs. Flight Simulator: Statistical Analysis

A t-test disclosed no statistical differences of the classification accuracy in the two experimental conditions (p = 0.67, *Cohen's d* = 0.17).

FIGURE 9 | Behavioral performance in the high load WM condition. No errors occurred during the low load WM condition.
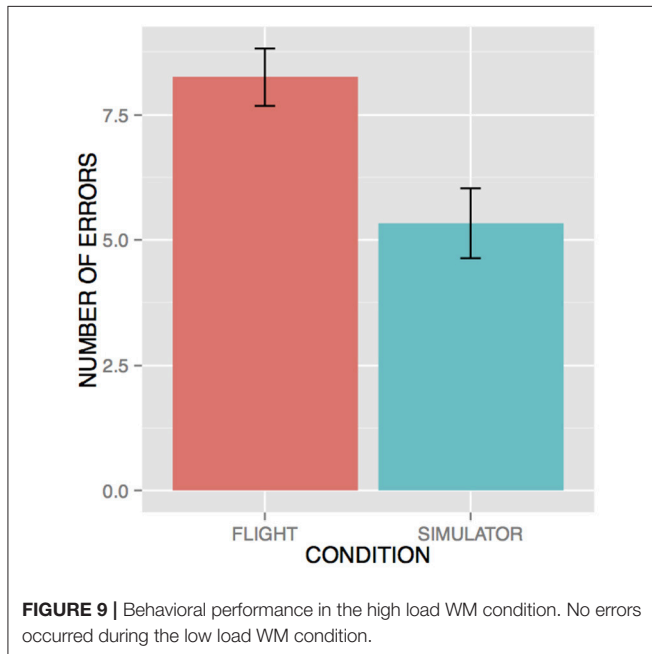
4. DISCUSSION

The motivation of this study was to develop on-line tools to monitor pilots' cognitive performance under realistic settings. We followed a two-step methodological approach as we first implemented and tested an inference system in a flight simulator and then in a real aircraft. We designed a task known to elicit WM (Durantin et al., 2015; Gateau et al., 2015) as this executive function is particularly engaged when operating aircraft (Causse et al., 2011a,b).

4.1. Summary of Findings

The behavioral results confirmed that these two levels of WM load were well contrasted, as the participants exhibited lower performance during the higher difficulty level. This result is in line with Taylor et al. study (Taylor et al., 2005; Durantin et al., 2015) and previous experiments (Gateau et al., 2015) showing that pilots' WM performance decline when four different ATC instructions have to be read back. Moreover, this drop in performance was most significant for the participants under actual flight conditions. Consistent with this finding, the real flight condition yielded to higher PFC activation than the simulated one only when the pilots had to execute the difficult WM load task. Taken together, these findings suggest that the mental demand was higher when operating the actual aircraft as the participants had not only to perform the WM task but also to monitor the flight path, the aircraft status and the airspace in a much more careful fashion than in the simulated condition.

Whereas this multitasking aspect of the real flight was not detrimental from a behavioral and neurophysiological point of view when performing the low WM stimuli it became

FIGURE 10 | The mean frontal $\Delta[HbO_2]$ peak response (peak value within 30 s post-trial onset minus 2 s average pre-trial onset) over the four ROIs of the PFC (from 1 - left to 4 - right) across WM loads and flight conditions.
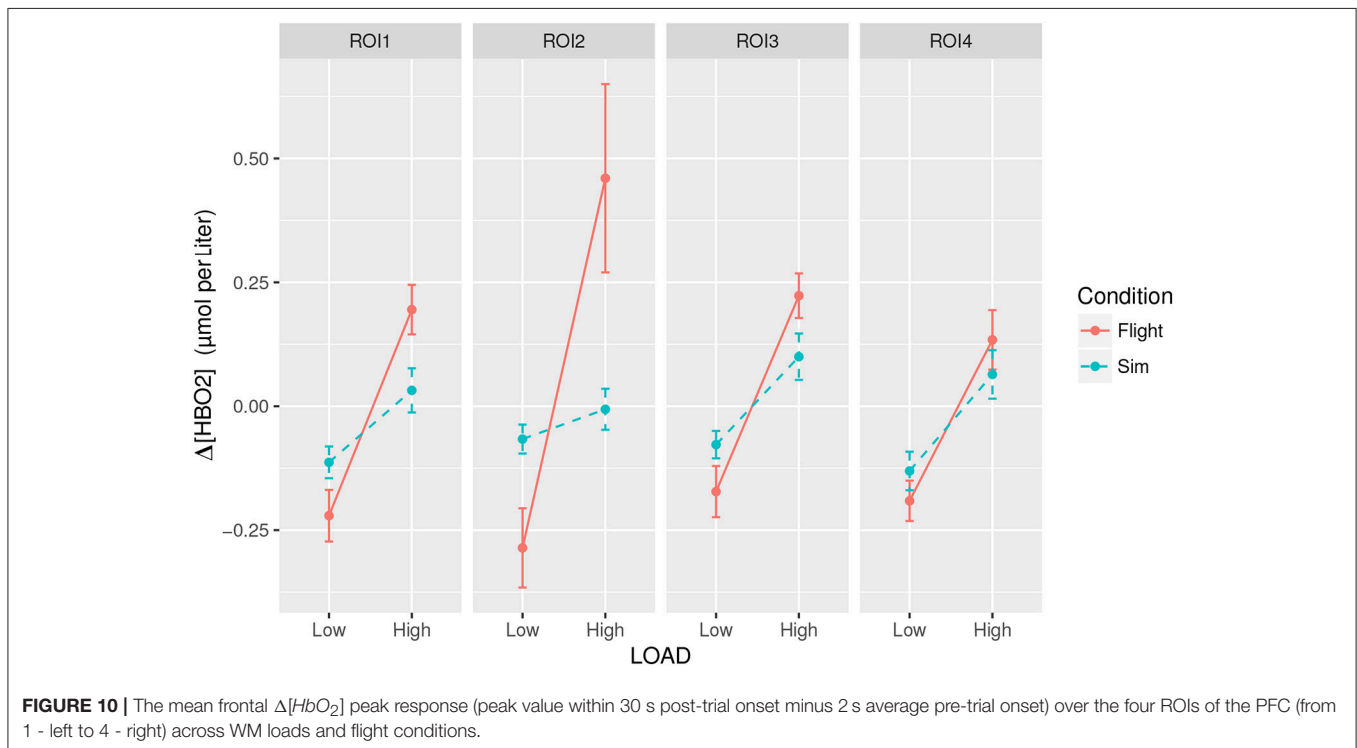
TABLE 1 | Simulator experiment: machine learning results.

Subject	Accuracy training	Accuracy testing	Precision	Recall	f1 score
01	90.00	100.00	100.00	100.00	100.00
02	100.00	60.00	66.67	40.00	50.00
03	100.00	70.00	100.00	40.00	57.14
04	100.00	100.00	100.00	100.00	100.00
05	100.00	70.00	75.00	60.00	66.67
06	100.00	80.00	100.00	60.00	75.00
07	100.00	90.00	100.00	80.00	88.89
08	100.00	80.00	80.00	80.00	80.00
09	100.00	90.00	100.00	80.00	88.89
10	80.00	70.00	100.00	40.00	57.14
11	70.00	50.00	50.00	100.00	66.67
12	90.00	60.00	55.56	100.00	71.43

TABLE 2 | Real flight experiment: machine learning results.

Subject	Accuracy training	Accuracy testing	Precision	Recall	f1 score
01	70.0	80.0	71.4	100	83.3
02	100	90.0	100	80.0	88.9
03	100	70.0	62.5	100	76.9
04	100	80.0	100	60.0	75.0
05	100	90.0	100	80.0	88.9
06	100	90.0	100	80.0	88.9
07	90.0	50.0	50.0	40.0	44.4
08	100	70.0	62.5	100	76.9
09	100	70.0	100	40.0	57.1
10	100	80.0	100	60.0	75.0
11	80.0	80.0	80.0	80.0	80.0
12	100	90.0	83.3	100	90.9

critical when engaged in the high WM one. One could suspect prioritization issue leading the pilots to focus more on flying the aircraft thus leaving few resources available to face the demand of the high WM stimuli. This could be one explanation for the higher levels of activation observed in fNIRS measurements that reflect the higher load of concurrent cognitive tasks induced by the real flying task compared to the simulated. Unfortunately, our aircraft was not equipped with a flight data recorder preventing us from analyzing the flight performance and investigating these prioritization and multi-tasking issues. Despite this limit, our study is consistent with Dahlstrom and Nahlinder (2009) who found evidence of higher cardiac activity when flying under realistic settings than in flight simulator. These results raise the question of the ecological validity of simulators. Their use is of undeniable interest (e.g., understanding cognitive performance, training pilots, assessing cockpit design) and they present several advantages in terms of economical costs and reproducibility issues. However, our findings and others (Philip et al., 2005; Dahlstrom and Nahlinder, 2009) suggest that the simulators may

need to be calibrated against real flying conditions to be more engaging.

Several field studies have demonstrated the potential of fNIRS to measure cortical activity while walking outdoors (McKendrick et al., 2016), facing prolonged stay at high altitude (Davranche et al., 2016), riding bikes (Piper et al., 2014), motorcycles (Kawashima et al., 2014), and flying helicopters (Kikukawa et al., 2008). Our study was conducted in accordance with the recent neuroergonomics approach to measure brain activity out of the laboratory. Indeed, beyond the offline analyses, we used machine learning techniques to perform single trial discrimination of the low WM load versus high WM load trials. The results of the classification process were available and displayed in a terminal to the experimenter after each: as soon as data of the trial were available, SVM discrimination process never required more than 10 ms to provide its result. The mean accuracy to classify low vs. high WM trials in the two experimental conditions exceeded the threshold of 70%, defined as a sufficient rate for pBCI (Kubler et al., 2006; Tai and Chau, 2009). These results compare well to the rare on-line studies such as the ones conducted by Naseer et al. (2014) (14 participants: 82.14% accuracy), Girouard et al. (2013) (9 participants: 83.5% accuracy), and (Schudlo and Chau, 2014) (10 participants: 77.4% accuracy). However, these and other (Kanoh et al., 2009; Hu et al., 2012; Power et al., 2012; Robinson et al., 2016) fNIRS-based BCI were not implemented under realistic settings and describe experiments in controlled lab settings.

4.2. Limitations and Avenues for Future Research

Despite the promising results presented in this paper for development of fNIRS based pBCI in ecologically valid environment, one could argue that the translation of the fNIRS-based pBCI in real cockpit to day-to-day flight operations might not be applicable. First, the addition of machine learning and this on-line classifier approach to standard procedures of aviation still remains a challenge as the reliability of the classifier does not meet aviation certification criteria (10^{-3} allowable failure probability). One approach to overcome this reliability problem would be to integrate complementary measurements such as EEG that could significantly enhance classification performance when combined with fNIRS as suggested by Khan et al. (2014).

Also, the accuracy score per subject must be interpreted with caution. In a two classes and five testing trials per class to fit with experimental constraints, classification performance should be higher than 75% to be statistically significant ($p < 0.05$) (Müller-Putz et al., 2008; Combrisson and Jerbi, 2015). Considering both groups in this study, 17 of 24 subjects were already above this threshold with our online classifier. Further improvements with machine learning methodologies would be needed to improve and optimize the classifier performance.

Secondly, availability of the information about WM level estimation is a key preoccupation. One criteria to evaluate on-line inference system is related to the delay of single trial classification. In our study, the diagnosis of the WM lasted less than 1.01 s after each pilot's response window. It could allow, for instance,

to automatically give a feedback to ATC that the pilot is currently facing a high workload situation and may have misinterpreted the last communication. This timing was comparable with results from other on-line fNIRS-based BCI latency (for a review of on-line fNIRS-based BCI latency, please see Strait et al., 2014). However, solutions have to be explored to speed up response detection on fNIRS signal that can drastically reduce latency in detecting change in a mental state (Cui et al., 2010; Hong and Naseer, 2016). Thirdly, our study was limited to monitor WM load in a binary and discrete fashion. Further studies have to be conducted to continuously discriminate a gradient of WM levels from underload to overload (Unni et al., 2017). Eventually, lingering issues remain regarding the effect of accelerations and headband motion on fNIRS signal (Mackey et al., 2013). In other scenarios accelerometer data with special processing could be used to eliminate any systemic effect of blood pooling.

Also, one should consider that fNIRS based pBCI could be first used for civilian application as highly automated modern aircraft prevents pilots from exceeding 1g maneuvers for passenger comfort and to avoid going against the *flight envelope protection*. Despite these limits, one can propose a progressive framework for the introduction of fNIRS in aviation. A first step is to consider the use of fNIRS based BCI to improve training via neurofeedback (Pope et al., 2014) and to tailor the flight sessions to the trainee (Chad et al., 2018). A second step is to use such inference system to monitor pilot's brain activity during each operational flight for *quantified self* purpose. These daily measures can be used to assess pilot's cognitive workload state and mental fatigue thus providing airlines with analyses tools for crew rostering. A third step is to stream the fNIRS data to the flight data recorder for accident analyses. These logged neurophysiological data would provide additional insights on the crew's cognition during these critical events and help accident investigators. A last step, when the reliability of the fNIRS-based inference system will meet the standard, would be to adapt the flight deck depending on the crew's changing WM load level. As previously demonstrated, stochastic decisional systems could be implemented to infer that human operators are engaged in demanding WM task and dynamically adapt interactions to prevent them from distraction (Gateau et al., 2016). The objective is to improve task allocation to enable better task switching, interruption management, and multi-tasking (Kohlmorgen et al., 2007; Solovey et al., 2011). Eventually, one should consider that such fNIRS based system could be applied to variety of contexts

whereby human operators interact with complex and critical systems (e.g., nuclear powerplant, train).

In summary, this study is the first report of the use of an online fNIRS based pBCI both in simulation (in silico) and in aircraft during flight (over the clouds) to measure pilot's WM . The implementation of this pBCI led to address several technical constraints, adapting and testing for instance a new wireless fNIRS that can be used by pilots and that has been approved for use during real flight. It also led to identify solutions to address potential sources of noise in signals such as the sunlight infrared shielding using aluminum based cover. Moreover, it provides important albeit preliminary information about fNIRS measures of the PFC hemodynamic response and its relationship to working memory workload, and in both simulation and actual flight environment. Level of immersion or realistic aspect of flight environment does appear to influence the performance as well as hemodynamic response in the anterior prefrontal cortex, at least for the air traffic control related working memory task. The measurements in simulator had larger fNIRS sensor coverage and future studies may compare simulation vs. actual flight or level of realistic aspect of environment with larger cortical coverage within the actual flight environment, for a more granular detailed comparison. Since fNIRS technology allows the development of mobile, non-intrusive and miniaturized devices, it has the potential to be deployed in future operational environments to monitor the pilot, adapt the complex system interface, and/or to assess the training of operators.

AUTHOR CONTRIBUTIONS

Study conception and design: FD, TG, and HA. Data acquisition : TG and FD. Data analysis : TG, FD, and HA. Data interpretation and writing FD, HA, and TG.

ACKNOWLEDGMENTS

REFERENCES

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Appel, G. (2005). *Technical Analysis: Power Tools for Active Investors*. FT Press.

Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Pozzi, S., and Babiloni, F. (2016). A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. *Prog. Brain Res.* 228, 295–328. doi: 10.1016/bs.pbr.2016.04.021

Ayaz, H. and Onaral, B. (2005). *Analytical Software and Stimulus-Presentation Platform to Utilize, Visualize and Analyze Near-Infrared Spectroscopy Measures.* Master's thesis, Master's Degree Thesis, Drexel University.

Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P. A., McKendrick, R., and Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Front. Hum. Neurosci.* 7:871. doi: 10.3389/fnhum.2013.00871

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Ayaz, H., Shewokis, P. A., Curtin, A., Izzetoglu, M., Izzetoglu, K., and Onaral, B. (2011). Using mazesuite and functional near infrared spectroscopy to study learning in spatial navigation. *J. Vis. Exp.* e3443. doi: 10.3791/3443

Baddeley, A. (1992). Working memory. *Science* 255, 556–559.

Batula, A. M., Kim, Y. E., and Ayaz, H. (2017). Virtual and actual humanoid robot control with four-class motor-imagery-based optical brain-computer interface. *BioMed. Res. Int.* 2017:1463512. doi: 10.1155/2017/1463512

Billings, C. E. and Cheaney, E. S. (1981). *Information Transfer Problems in the Aviation System.* NASA Technical Report 1875, 89–90.

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., et al. (2010). The berlin brain–computer interface: non-medical uses of bci technology. *Front. Neurosci.* 4:198. doi: 10.3389/fnins.2010.00198

Borghini, G., Aricò, P., Di Flumeri, G., Cartocci, G., Colosimo, A., Bonelli, S., et al. (2017). EEG-based cognitive control behaviour assessment: an ecological study with professional air traffic controllers. *Sci. Rep.* 7:547. doi: 10.1038/s41598-017-00633-7

Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using eeg spectral power and erps in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Brouwer, A.-M., Van Erp, J., Heylen, D., Jensen, O., and Poel, M. (2013). "Effortless passive bcis for healthy users," in *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion, Vol. 8009* (Berlin; Heidelberg; Vancouver: Springer), 615–622.

Çakır, M. P., Vural, M., Koç, S. Ö., and Toktaş, A. (2016). "Real-time monitoring of cognitive workload of airline pilots in a flight simulator with fnir optical brain imaging technology," in *International Conference on Augmented Cognition* (Springer), 147–158.

Callan, D. E., Durantin, G., and Terzibas, C. (2015). Classification of single-trial auditory events using dry-wireless eeg during real and motion simulated flight. *Front. Syst. Neurosci.* 9:11. doi: 10.3389/fnsys.2015.00011

Callan, D. E., Terzibas, C., Cassel, D. B., Sato, M.-A., and Parasuraman, R. (2016). The brain is faster than the hand in split-second intentions to respond to an impending hazard: a simulation of neuroadaptive automation to speed recovery to perturbation in flight attitude. *Front. Hum. Neurosci.* 10:187. doi: 10.3389/fnhum.2016.00187

Causse, M., Dehais, F., Arexis, M., and Pastor, J. (2011a). Cognitive aging and flight performances in general aviation pilots. *Aging Neuropsychol. Cogn.* 18, 544–561. doi: 10.1080/13825585.2011.586018

Causse, M., Dehais, F., and Pastor, J. (2011b). Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *Int. J. Aviat. Psychol.* 21, 217–234. doi: 10.1080/10508414.2011.582441

Causse, M., Dehais, F., Péran, P., Sabatini, U., and Pastor, J. (2013). The effects of emotion on pilot decision-making: a neuroergonomic approach to aviation safety. *Transport. Res. Part C Emerg. Technol.* 33, 272–281. doi: 10.1016/j.trc.2012.04.005

Chad, S., Dehais, F., Roy, N. R., Harrivel, A., Last, M. C., Kennedy, K., et al. (2018). *Biocybernetic Adaptation Strategies: Machine Awareness of Human Engagement for Improved Operational Performance, in HCI Conference* (Las Vegas).

Combrisson, E. and Jerbi, K. (2015). Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J. Neurosci. Methods* 250, 126–136. doi: 10.1016/j.jneumeth.2015.01.010

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297.

Cui, X., Bray, S., and Reiss, A. L. (2010). Speeded near infrared spectroscopy (NIRS) response detection. *PLoS ONE* 5:e15474. doi: 10.1371/journal.pone.0015474

Cutrell, E. and Tan, D. (2008). BCI for passive input in HCI. in *Proc. CHI* 8, 1–3.

Dahlstrom, N. and Nahlinder, S. (2009). Mental workload in aircraft and simulator during basic civil aviation training. *Int. J. Aviat. Psychol.* 19, 309–325. doi: 10.1080/10508410903187547

Davranche, K., Casini, L., Arnal, P. J., Rupp, T., Perrey, S., and Verges, S. (2016). Cognitive functions and cerebral oxygenation changes during acute and prolonged hypoxic exposure. *Physiol. Behav.* 164, 189–197. doi: 10.1016/j.physbeh.2016.06.001

Dehais, F., Behrend, J., Peysakhovich, V., Causse, M., and Wickens, C. D. (2017). Pilot flying and pilot monitoring aircraft state awareness during go-around execution in aviation: a behavioral and eye tracking study. *Int. J. Aerospace Psychol.* 27, 15–28. doi: 10.1080/10508414.2017.1366269

Delpy, D. T., Cope, M., Van der Zee, P., Arridge, S., Wray, S., and Wyatt, J. (1988). Estimation of optical pathlength through tissue from direct time of flight measurement. *Phys. Med. Biol.* 33:1433.

Dijksterhuis, C., de Waard, D., Brookhuis, K., Mulder, B., and de Jong, R. (2013). Classifying visuomotor workload in a driving simulator using subject specific spatial brain patterns. *Front. Neurosci.* 7:149. doi: 10.3389/fnins.2013.00149

Durantin, G., Gagnon, J.-F., Tremblay, S., and Dehais, F. (2014a). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* 259, 16–23. doi: 10.1016/j.bbr.2013.10.042

Durantin, G., Scannella, S., Gateau, T., Delorme, A., and Dehais, F. (2014b). "Moving average convergence divergence filter preprocessing for real-time event-related peak activity onset detection: Application to fnirs signals," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE), 2107–2110.

Durantin, G., Scannella, S., Gateau, T., Delorme, A., and Dehais, F. (2015). Processing functional near infrared spectroscopy signal with a kalman filter to assess working memory during simulated flight. *Front. Hum. Neurosci.* 9:707. doi: 10.3389/fnhum.2015.00707

Frey, J., Hachet, M., and Lotte, F. (2017). EEG-based neuroergonomics for 3D user interfaces: opportunities and challenges. *Le travail humain.* 80, 73–92.

Gagnon, J.-F., Durantin, G., Vachon, F., Causse, M., Tremblay, S., and Dehais, F. (2012). "Anticipating human error before it happens: towards a psychophysiological model for online prediction of mental workload," in *Proceedings of the Human Factors and Ergonomics Society Chapter Europe.*

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., and Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PLoS ONE* 10:e0121279. doi: 10.1371/journal.pone.0121279

Gateau, T., P. Carvalho Chanel, C., Le, M.-H., and Dehais, F. (2016). "Considering human's non-deterministic behavior and his availability state when designing a collaborative human-robots system," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Daejeon: IEEE).

George, L. and Lécuyer, A. (2010). "An overview of research on" passive" brain-computer interfaces for implicit human-computer interaction," in *International Conference on Applied Bionics and Biomechanics ICABB 2010-Workshop W1" Brain-Computer Interfacing and Virtual Reality* (Venice).

Girouard, A., Solovey, E. T., and Jacob, R. J. (2013). Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *Inte. J. Auton. Adapt. Commun. Syst.* 6, 26–44. doi: 10.1504/IJAACS.2013.050689

Gramann, K., Fairclough, S. H., Zander, T. O., and Ayaz, H. (2017). Editorial: trends in neuroergonomics. *Front. Hum. Neurosci.* 11:165. doi: 10.3389/fnhum.2017.00165

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task?quantified in the prefrontal cortex using fNIRS. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935.

Hirshfield, L. M., Gulotta, R., Hirshfield, S., Hincks, S., Russell, M., Ward, R., et al. (2011). "This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (ACM), 373–382.

Hong, K.-S. and Naseer, N. (2016). Reduction of delay in detecting initial dips from functional near-infrared spectroscopy signals using vector-based phase analysis. *Int. J. Neural syst.* 26:1650012. doi: 10.1142/S012906571650012X

Hu, X.-S., Hong, K.-S., and Ge, S. S. (2012). fNIRS-based online deception decoding. *J. Neural Eng.* 9:026012. doi: 10.1088/1741-2560/9/2/026012

Huppert, T. J., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2009). Homer: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Appl. Opt.* 48, D280–D298. doi: 10.1364/AO.48.00D280

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., and Chance, B. (2004). Functional optical brain imaging using near-infrared during cognitive tasks. *Int. J. Hum. Comput. Interact.* 17, 211–227. doi: 10.1207/s15327590ijhc1702_6

Jasdzewski, G., Strangman, G., Wagner, J., Kwong, K., Poldrack, R., and Boas, D. (2003). Differences in the hemodynamic response to event-related motor and visual paradigms as measured by near-infrared spectroscopy. *Neuroimage* 20, 479–488. doi: 10.1016/S1053-8119(03)00311-2

Kanoh, S., Murayama, Y.-M., Miyamoto, K.-I., Yoshinobu, T., and Kawashima, R. (2009). "A NIRS-based brain-computer interface system during motor imagery: System development and online feedback training," in *2009 Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society (Minneapolis: IEEE), 594–597.

Kato, T., Kamei, A., Takashima, S., and Ozaki, T. (1993). Human visual cortical function during photic stimulation monitoring by means of near-infrared spectroscopy. *J. Cereb. Blood Flow Metab.* 13, 516–520.

Kawashima, R., Matsumoto, T., and Tanimoto, Y. (2014). Cortical activity while riding motorcycles measured with a wearable near infrared topography system. *Int. J. Automot. Eng.* 5, 77–83. doi: 10.20485/jsaeijae.5.2_77

Khan, M. J., Hong, M. J., and Hong, K.-S. (2014). Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface. *Front. Hum. Neurosci.* 8:244. doi: 10.3389/fnhum.2014.00244

Kikukawa, A., Kobayashi, A., and Miyamoto, Y. (2008). Monitoring of pre-frontal oxygen status in helicopter pilots using near-infrared spectrophotometers. *Dyn. Med.* 7:10. doi: 10.1186/1476-5918-7-10

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., et al. (2007). "Improving human performance in a real operating environment through real-time mental workload detection," in *Toward Brain-Computer Interfacing* (Cambridge, MA: MIT press), 409–422.

Kubler, A., Mushahwar, V., Hochberg, L. R., and Donoghue, J. P. (2006). BCI meeting 2005-workshop on clinical issues and applications. *IEEE Trans. Neural Syst. Rehabil. Eng.* 14, 131–134. doi: 10.1109/TNSRE.2006.875585

León-Domínguez, U., Martín-Rodríguez, J. F., and León-Carrión, J. (2015). Executive n-back tasks for the neuropsychological assessment of working memory. *Behav. Brain Res.* 292, 167–173. doi: 10.1016/j.bbr.2015.06.002

Li, C., Gong, H., Zeng, S., and Luo, Q. (2005). *Verbal Working Memory Load Affects Prefrontal Cortices Activation: Evidence from a Functional Nirs Study in Humans* (San Diego, CA: In SPIE Conference on Optics and Photonics), 33–40.

Luo, Q., Zeng, S., Chance, B., and Nioka, S. (2002). "Monitoring of brain activity with near-infrared spectroscopy," in *Handbook of Optical Biomedical Diagnostics* (Bellingham, WA: SPIE Press), 455–486. doi: 10.1117/3.2219603.ch8

Mackey, J. R., Harrivel, A. R., Adamovsky, G., Lewandowski, B. E., Gotti, D. J., Tin, P., et al. (2013). "Effects of varying gravity levels on fnirs headgear performance and signal recovery," in *The American Institute of Aeronautics and Astronautics* (Boston, MA).

McKendrick, R., Ayaz, H., Olmstead, R., and Parasuraman, R. (2014). Enhancing dual-task performance with verbal and spatial working memory training: continuous monitoring of cerebral hemodynamics with NIRS. *Neuroimage* 85, 1014–1026. doi: 10.1016/j.neuroimage.2013.05.103

McKendrick, R., Parasuraman, R., and Ayaz, H. (2015). Wearable functional near infrared spectroscopy (fnirs) and transcranial direct current stimulation (tdcs): expanding vistas for neurocognitive augmentation. *Front. Syst. Neurosci.* 9:27. doi: 10.3389/fnsys.2015.00027

McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., et al. (2016). Into the wild: neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy. *Front. Hum. Neurosci.* 10:216. doi: 10.3389/fnhum.2016.00216

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* 101:343.

Morrow, D., Lee, A., and Rodvold, M. (1993). Analysis of problems in routine controller-pilot communication. *Int. J. Aviat. Psychol.* 3, 285–302.

Mühl, C., Jeunet, C., and Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Front. Neurosci.* 8:114. doi: 10.3389/fnins.2014.00114

Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., and Pfurtscheller, G. (2008). Better than random: a closer look on bci results. *Int. J. Bioelectromagnet.* 10, 52–55.

Naseer, N. and Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: a review. *Front. Hum. Neurosci.* 9:3. doi: 10.3389/fnhum.2015.00003

Naseer, N., Hong, M. J., and Hong, K.-S. (2014). Online binary decision decoding using functional near-infrared spectroscopy for the development of brain–computer interface. *Exp. Brain Res.* 232, 555–564. doi: 10.1007/s00221-013-3764-1

Parasuraman, R. and Rizzo, M. (2008). *Neuroergonomics: The Brain at Work.* New York, NY: Oxford University Press.

Parasuraman, R. and Wilson, G. F. (2008). Putting the brain to work: Neuroergonomics past, present, and future. *Hum. Factors* 50, 468–474. doi: 10.1518/001872008X288349

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.1016/j.patcog.2011.04.006

Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Akerstedt, T., et al. (2005). Fatigue, sleepiness, and performance in simulated versus real driving conditions. *Sleep* 28:1511. doi: 10.1093/sleep/28.12.1511

Piper, S. K., Krueger, A., Koch, S. P., Mehnert, J., Habermehl, C., Steinbrink, J., et al. (2014). A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *Neuroimage* 85, 64–71. doi: 10.1016/j.neuroimage.2013.06.062

Pope, A. T., Stephens, C. L., and Gilleade, K. (2014). "Biocybernetic adaptation as biofeedback training method," in *Advances in Physiological Computing*, eds S. Fairclough and K. Gilleade (London: Springer), 91–115.

Power, S. D., Kushki, A., and Chau, T. (2012). Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by NIRS. *PLoS ONE* 7:e37791. doi: 10.1371/journal.pone.0037791

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Reynal, M., Colineaux, Y., Vernay, A., and Dehais, F. (2016). "Pilot flying vs. pilot monitoring during the approach phase: an eye–tracking study," in *HCI-Aero 2016, International Conference on Human-Computer Interaction in Aerospace* (Paris).

Risser, M. R., Scerbo, M. W., Baldwin, C. L., and McNamara, D. S. (2006). "Interference timing and acknowledgement response with voice and datalink atc commands," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50 (San Francisco, CA: Sage Publications), 11–15.

Robinson, N., Zaidi, A. D., Rana, M., Prasad, V. A., Guan, C., Birbaumer, N., et al. (2016). Real-time subject-independent pattern classification of overt and covert movements from fNIRS signals. *PLoS ONE* 11:e0159959. doi: 10.1371/journal.pone.0159959

Rome, F., Adam, G., Condette, J., Causse, M., and Dehais, F. (2012). "Go-around manoeuver: a simulation study," in *Proceedings of the European Association For Aviation Psychology Conference* (Lisboa: SAGE Publications).

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2013). "Mental fatigue and working memory load estimation: interaction and implications for eeg-based passive bci," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka: IEEE), 6607–6610.

Roy, R. N., Verdiere, K., Scannella, S., and Dehais, F. (2017). "Passive BCI tools for mental state estimation in aerospace applications," in *The First Biannual Neuroadaptive Technology Conference* (Berlin), 79.

Sato, H., Fuchino, Y., Kiguchi, M., Katura, T., Maki, A., Yoro, T., et al. (2005). Intersubject variability of near-infrared spectroscopy signals during sensorimotor cortex activation. *J. Biomed. Opt.* 10:044001. doi: 10.1117/1.1960907

Scerbo, M. W., Risser, M. R., and Baldwin, C. L. (2003). "Implementing speech and simulated data link commands: the role of task interference and message length," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47 (Denver: SAGE Publications), 95–99.

Schreppel, T., Egetemeir, J., Schecklmann, M., Plichta, M. M., Pauli, P., Ellgring, H., Fallgatter, A. J., and Herrmann, M. J. (2008). Activation of the prefrontal cortex in working memory and interference resolution processes assessed with near-infrared spectroscopy. *Neuropsychobiology*, 57, 188–193.

Schudlo, L. C. and Chau, T. (2014). Dynamic topographical pattern classification of multichannel prefrontal NIRS signals: II. online differentiation of mental arithmetic and rest. *J. Neural Eng.* 11:016003. doi: 10.1088/1741-2560/11/1/016003

Schudlo, L. C. and Chau, T. (2015). Towards a ternary NIRS-bci: single-trial classification of verbal fluency task, stroop task and unconstrained rest. *J. Neural Eng.* 12:066008. doi: 10.1088/1741-2560/12/6/066008

Solovey, E. T., Lalooses, F., Chauncey, K., Weaver, D., Parasi, M., Scheutz, M., et al. (2011). "Sensing cognitive multitasking for a brain-based adaptive user interface," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC: ACM), 383–392.

Strait, M., Canning, C., and Scheutz, M. (2014). What we can and cannot do with near infrared spectroscopy. *Front. Neurosci.* 8:117. doi: 10.3389/fnins.2014.00117

Tai, K. and Chau, T. (2009). Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface. *J. Neuroeng. Rehabil.* 6:39. doi: 10.1186/1743-0003-6-39

Taylor, J. L., O'Hara, R., Mumenthaler, M. S., Rosen, A. C., and Yesavage, J. A. (2005). Cognitive ability, expertise, and age differences in following air-traffic control instructions. *Psychol. Aging* 20, 117–133. doi: 10.1037/0882-7974.20.1.117

Taylor, J. L., Yesavage, J. A., Morrow, D. G., Dolhert, N., Brooks, J. O., and Poon, L. W. (1994). The effects of information load and speech rate on younger and older aircraft pilots' ability to execute simulated air-traffic controller instructions. *J. Gerontol.* 49, 191–200.

Unni, A., Ihme, K., Jipp, M., and Rieger, J. W. (2017). Assessing the driver?s current level of working memory load with high density functional near-infrared spectroscopy: a realistic driving simulator study. *Front. Hum. Neurosci.* 11:167. doi: 10.3389/fnhum.2017.00167

Utsugi, K., Obata, A., Sato, H., Katsura, T., Sagara, K., Maki, A., et al. (2007). "Development of an optical brain-machine interface," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (IEEE), 5338–5341.

Van Erp, J., Brouwer, A., and Zander, T. (2015). Editorial: Using neurophysiological signals that reflect cognitive or affective state. *Front. Neurosci.* 9:193. doi: 10.3389/fnins.2015.00193

Vecchiato, G., Borghini, G., Aricò, P., Graziani, I., Maglione, A. G., Cherubino, P., et al. (2016). Investigation of the effect of eeg-bci on the simultaneous execution of flight simulation and attentional tasks. *Med. Biol. Eng. Comput.* 54, 1503–1513. doi: 10.1007/s11517-015-1420-6

Verdière, K. J., Roy, R. N., and Dehais, F. (2018). Detecting pilot's engagement using fnirs connectivity features in an automated vs manual landing scenario. *Front. Hum. Neurosci.* 12:6. doi: 10.3389/fnhum.2018.00006

Villringer, A., and Obrig, H. (2002). "Near infrared spectroscopy and imaging," in *Brain Mapping: The Methods, 2nd edn.* (Sendai: Elsevier), 141–158.

Zander, T. O. and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Zander, T. O., Shetty, K., Lorenz, R., Leff, D. R., Krol, L. R., Darzi, A. W., Gramann, K., and Yang, G.-Z. (2017). Automated task load detection with electroencephalography: Towards passive brain–computer interfacing in robotic surgery. *J. Med. Robot. Res.* 2:1750003. doi: 10.1142/S2424905X17500039

frontiers
in Human Neuroscience

Electroencephalography Based Analysis of Working Memory Load and Affective Valence in an N-back Task with Emotional Stimuli

Sebastian Grissmann[1]*, Josef Faller[2], Christian Scharinger[3], Martin Spüler[4] and Peter Gerjets[1,3]

[1]LEAD Graduate School, University of Tübingen, Tübingen, Germany, [2]Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, NY, United States, [3]Leibniz-Institut für Wissensmedien, Multimodal Interaction Lab, Tübingen, Germany, [4]Wilhelm-Schickard-Institute for Computer Science, University of Tübingen, Tübingen, Germany

Most brain-based measures of the electroencephalogram (EEG) are used in highly controlled lab environments and only focus on narrow mental states (e.g., working memory load). However, we assume that outside the lab complex multidimensional mental states are evoked. This could potentially create interference between EEG signatures used for identification of specific mental states. In this study, we aimed to investigate more realistic conditions and therefore induced a combination of working memory load and affective valence to reveal potential interferences in EEG measures. To induce changes in working memory load and affective valence, we used a paradigm which combines an N-back task (for working memory load manipulation) with a standard method to induce affect (affective pictures taken from the International Affective Picture System (IAPS) database). Subjective ratings showed that the experimental task was successful in inducing working memory load as well as affective valence. Additionally, performance measures were analyzed and it was found that behavioral performance decreased with increasing workload as well as negative valence, showing that affective valence can have an effect on cognitive processing. These findings are supported by changes in frontal theta and parietal alpha power, parameters used for measuring of working memory load in the EEG. However, these EEG measures are influenced by the negative valence condition as well and thereby show that detection of working memory load is sensitive to affective contexts. Unexpectedly, we did not find any effects for EEG measures typically used for affective valence detection (Frontal Alpha Asymmetry (FAA)). Therefore we assume that the FAA measure might not be usable if cognitive workload is induced simultaneously. We conclude that future studies should account for potential context-specificity of EEG measures.

Keywords: electroencephalography (EEG), working memory, affective valence, emoback, IAPS

INTRODUCTION

Investigating Complex User States with the Electroencephalogram

In recent years, there has been increased interest to use the electroencephalogram (EEG) in the context of human-machine interaction (Frey et al., 2013). However, most studies using the EEG to measure mental states focus on very specific states like working memory (Klimesch, 1999) or affective valence (Ahern and Schwartz, 1985), which are investigated in well controlled lab environments. Therefore, the indicators used in such studies might not provide robust measurements outside the lab, since real world environments tend to evoke much more complex and multidimensional mental states that involve different cognitive, emotional and motivational components (Gerjets et al., 2014). Furthermore, many measures used to infer mental states from the ongoing EEG are known to have many-to-many relations, meaning that several physiological variables are associated with multiple psychological elements (Fairclough, 2009). Hence, it is necessary to systematically investigate the relationship between different mental states and the different EEG measures that are widely used in neuroscientific studies, to investigate if such measures can be used outside the lab. In this article we systematically investigate the relation between two types of mental states that are widely used in the context of human-machine interaction, namely working memory load (e.g., Spüler et al., 2016) and affective valence (e.g., De Smedt and Menschaert, 2012), to study the interaction of the brain responses typically associated to these mental states.

Working Memory Load

There are many different ways to induce mental states characterized by high levels of working memory load (Wilhelm et al., 2013). One way to induce working memory load that has been widely used in the context of cognitive neuroscience is the N-back task (Kirchner, 1958; Scharinger et al., 2015). The N-back task is a continuous performance task where subjects are presented with a series of stimuli and have to indicate whether the current stimulus is identical to the stimulus presented N-steps before, or nor. Hence, the load factor "N" allows to adjust the difficulty of the task, thereby manipulating working memory load.

Working memory load is commonly defined as the interplay of controlled attentional processes and short term memory structures that handle different representational codes via various temporal storage components (Baddeley, 2003, 2012). The central attentional control system is assumed to be mainly located in frontal regions of the brain like the dorsolateral prefrontal cortex, while content in short term memory is thought to be maintained via parietal brain areas like the intraparietal sulcus (Klingberg, 2009; Scharinger et al., 2015). Accordingly, increases in mental workload usually result in increased frontal theta activity as well as decreased parietal alpha activity (Gevins et al., 1995; Klimesch, 1999; Smith and Gevins, 2005). However, previous research has shown that other mental states can also have an effect on measures used for workload detection. Roy et al. (2013), for example, induced fatigue and found that with increasing time on task the discriminability of working memory load was decreased.

Affective Valence

There are multiple ways to induce affective experiences in lab settings (Gerrards-Hesse et al., 1994). One effective way is the use of affective picture stimuli. The most prominent picture database is the International Affective Picture System (IAPS; Lang et al., 1997) which comprises a large set of standardized and emotionally evocative color photographs. All stimuli of the database have been rated along the dimensions of valence and arousal as described in the two-dimensional circumplex model of emotion (Russell and Pratt, 1980).

The valence dimension reflects the pleasantness of a situation and ranges from sadness to happiness. The arousal dimension reflects the responsiveness of the organism and ranges from sleep to frenzied excitement. Several studies have used the EEG to study affective states in the past Olofsson et al. (2008) and Kim et al. (2013). While there are many different approaches to infer affective valence in the EEG, some even using right hemispheric activity in the beta (Rowland et al., 1985) or gamma band (Müller et al., 1999), the two most widely used EEG measures to infer affective states are the late positive potential (LPP) and the so-called Frontal Alpha Asymmetry (FAA). The LPP is an EEG feature in the time domain and represents a positive deflection in the ERP-curve, reflecting the activity related to the arousal dimension (Schupp et al., 2000). The FAA represents the individual hemispheric contributions which is related to the affective valence dimension (Ahern and Schwartz, 1985; Tomarken et al., 1990; Huang et al., 2012). Increased right frontal activity is an indicator of a mental state characterized by negative affective valence. Usually, this measure is used during resting (with closed eyes) or passive viewing conditions (Davidson, 1992). Interestingly, previous work has shown that working memory load can result in lateralized activity as well. For instance, a study by Baldwin and Penaranda (2012) used several tasks to induce mental workload and found more left hemispheric activity during increased cognitive load. This might result in potential interferences between affective valence and working memory load in the FAA measure. Analyzing this type of potential interference between working memory load and affective valence is the main goal of this article.

Investigating Interactions between Working Memory Load and Affective Valence: The Emoback Task

Previous research on the use of neural signatures of mental states has largely ignored the problem of potential interference between working memory load and affective experiences. Only one recent study addressed the effect of an affective experience on the automatic identification of working memory load. In this study, Mühl et al. (2014) used an N-back task to manipulate working memory load while social stress was induced with a stress-induction protocol based on the Trier Social Stress Task (Kirschbaum et al., 1993). The authors attempted to detect mental workload during stress using features from the frequency

domain as well as the time domain. They concluded that it is possible to transfer methods across affective contexts, but only with diminished performance. However, these results are limited to one specific affective context (social stress). Furthermore, the authors were focusing on the identification of working memory load alone. The current article wants to extend these findings using a more general affective response and also account for the potential interference of cognitive and affective components.

In order to collect suitable data for this objective, we used a combination of an N-back task with a standard affect induction, called the emoback task. Interactions between cognitive and affective processes in the EEG have been previously investigated with the affective flanker task (Alguacil et al., 2013). However, the affective flanker task can be seen as a simple stimulus-response task that only requires perceptual inhibition and therefore does not represent a genuine working memory task. The emoback task does involve memory components and, to our knowledge, has only been used twice in combination with the EEG. First, a study by MacNamara et al. (2011) used the emoback task with affective pictures as distractors and found that the LPP was modulated not only by affective responses towards emotional pictures, but also by working memory load. Second, a study by Kopf et al. (2013) used a N-back task with emotional words and recorded data from EEG and fNIRS. They found more errors during the negative condition, especially for high task difficulty. An ERP-analysis also revealed that the LPP is influenced by the difficulty in the working memory task and that this influence is further modulated by affective valence. However, both these studies used the LPP, as feature commonly associated with the arousal dimension. We want to investigate affective responses related to the valence dimension, which allows to make more general discriminations of affective states into positive and negative states. Moreover, the study by MacNamara et al. (2011) only used affective stimuli as distractors, while we want to use a paradigm that inherently activates cognitive as well as affective processes. Finally, while the study by Kopf et al. (2013) used emotional words to induce affective reactions, we assume that affective pictures can elicit stronger affective reactions. We therefore decided to use an emoback task with affective pictures from the IAPS database to investigate potential interferences between working memory load and affective valence using frontal theta activity, parietal alpha activity as well as the FAA.

In previous analyses of the same dataset (Grissmann et al., in press), we found that classification of working memory load under affective valence can result in classification accuracies above 70%, which can be further improved via data integration over time. However, we also found that positive as well as negative valenced affective contexts led to decreased classification accuracies, when compared to a neutral affective context. Additionally, classifiers failed to generalize across affective contexts, which highlighted the need to better understand the interactions between working memory load and affective valence in such a context.

Research Questions and Hypotheses

We investigated the influence of working memory load and affective valence on subjective measures, behavioral measures

as well as the corresponding EEG measures. Furthermore, we investigated potential interferences between cognitive processes and affective processes as reflected in EEG measures used to infer working memory load and EEG measures used to infer affective valence.

More specifically, we investigated potential effects of load levels in the emoback task on subjective ratings, accuracies and reaction times, which might also be reflected in the corresponding EEG measures.

Additionally, we investigated potential effects of the affective valence inductions in the emoback task on subjective measures, behavioral measures as well as EEG measures.

Beyond the main effects of working memory load and affective valence on behavioral measures and EEG measures, we also analyzed whether EEG measures used for mental workload detection are sensitive to different affective contexts and whether EEG measures used to infer affective valence are sensitive to working memory load.

MATERIALS AND METHODS

Sample

For this study, we collected data from 27 female subjects. Female subjects were used in this study because they tend to show stronger reactions toward affective stimuli (Lang et al., 1993) and also exhibit more stable responses (Ahern and Schwartz, 1985). Three subject were removed due to the low quality of the signals. All of the participants were university students, aged above 18 years (mean: 23.0 years; range: 19–32 years), right handed and had no blood phobia to avoid extreme responses toward the experimental stimuli. All subjects provided written informed consent and were paid 20 € for participation in the experiment. The study was approved by the ethic committee of the Knowledge Media Research Center Tuebingen.

Recording of Physiological Data

Sixty channels of EEG were recorded using an ActiCHamp amplifier and active Ag/Cl-electrodes (Brainproducts GmbH, Gilching, Germany). Electrodes were placed according to the extended 10-20 system. The electrooculogram (EOG) was recorded with four EEG electrodes located at the left and right canthi as well as above (channel Fp1) and below the left eye. All channels were referenced to channel FCz. Impedances were kept below 10 kΩ. The data were sampled at 1 kHz. This electrode layout was chosen to allow for potential source localization approaches. For later processing the data was downsampled to 250 Hz. During the recordings subjects were instructed to sit in a relaxed posture to avoid artifact contamination of the data.

Preprocessing

To automatically reject time windows that are contaminated by artifacts, a time window was removed if channel power at more than six channels exceeded five times the standard deviation. After computing independent components using the CUDAICA implementation (Raimondo et al., 2012) of the Infomax independent component analysis (Bell and Sejnowski, 1995), eye

movement artifacts were removed (reduced) using the ADJUST approach (Mognon et al., 2011) which is implemented as plug-in in the EEGLAB toolbox (Delorme et al., 2011). Finally the EEG signal was bandpass filtered between 1 Hz and 45 Hz using two-way least-squares finite impulse response filtering.

Stimuli

For the positive, neutral and negative valence conditions we selected 96 stimuli each. The stimuli were taken from the IAPS database (Lang et al., 1997) and selected based on the valence and arousal ratings provided with the database. Valence and arousal ratings are usually confounded, meaning that stimuli with strong (positive or negative) valence ratings usually also have high arousal ratings. To improve discriminability between affective conditions we made sure that stimuli for the positive condition had the highest valence ratings while stimuli for the negative conditions had the lowest valence ratings. Furthermore, stimuli for the neutral condition were selected based on the lowest arousal ratings. All selected stimuli had a quadratic shape and were centrally presented on a standard 23 inch display. To avoid unnecessary eye-movements the size of the stimuli was scaled to fill 60% of the height of the display.

Study Design and Block Structure

The affective valence conditions were presented in groups of four trial blocks with identical affective valence. The affective valence conditions being either positive, neutral or negative. Load levels of the emoback task were alternated block wise. The load factor of the emoback task had two levels, one-back and two-back. We avoided the use of a zero-back condition, because this would require to repeatedly use the same affective target stimuli. This repetition of the same stimuli might have resulted in affective habituation with regard to the target stimuli, thereby diminishing the affect induction (Leventhal et al., 2007). We also avoided a 3-back condition because some studies showed that this load level can lead to task disengagement (Ayaz et al., 2007). Target response hand as well as starting load level of the emoback task were balanced across the subject sample. See **Figure 1** for an illustration of the study design.

All participants performed 12 blocks in total. Each block started with a 10 s baseline where subjects were instructed to relax and visually fixate a centrally presented light gray fixation cross. After the baseline the first trial started with a stimulus presentation. There were 72 trials in one block, consisting of 24 target stimuli and 48 distractors. The 24 target stimuli were randomly selected for each subject from 96 unique stimuli (four blocks with 24 target stimuli per affective condition). Targets and distractors were sometimes interleaved in the 2-back condition, meaning that two target stimuli could appear right after each other. Here is an example: Distractor (e.g., table), distractor (e.g., chair), target (table) and target (chair). After the last trial, another baseline phase, also with a duration of 10 s, was recorded. Between blocks there were short brakes between 1 min and 3 min. See **Figure 2** for an illustration of the block structure.

Subjective Measures

After each run, subjects were asked to rate their subjective experience of the last run. Subjective experience of working memory load was measured using one (modified) item taken from the NASA task load index (Hart and Staveland, 1988). The item asked participants how cognitively demanding they experienced the last experimental run. The rating scale ranged from 0 (absolutely no mental demand) to 100 (highest possible mental demand).

Emotional experiences are commonly judged with the help of rating scales. The most widely used is a visual analog scale called the Self-Assessment Manikin (SAM; Bradley and Lang, 1994). It enables fast and reliable judgments about the current emotional state.

Trial Structure

Each trial started with a stimulus presentation phase of 1500 ms. The duration of the stimulus presentation was selected based on tests using pilot subjects and should ensure that enough trials can be recorded to get reliable estimates for the EEG measures used and the desired affective responses are evoked. During the stimulus presentation phase the subjects had to indicate if the current stimuli was a target (i.e., identical to the stimulus one step before in the 1-back condition and two steps before in the 2-back condition) or if the current stimulus was a distractor. Left and right control keys of a standard keyboard were used as inputs. The subjects were instructed to react as quickly as possible to ensure an effect in performance measures. Between stimuli there were interstimulus intervals (ISI) of 1500 ms duration including one to 500 ms of jitter at the end of the ISI to avoid periodic responses in the EEG data. During the ISI the same gray crosshair as in the baseline conditions was presented on the screen.

FIGURE 1 | Illustration of EMOBACK study design: the three affective valence conditions were presented grouped and permutated across the whole sample. Each affective condition consisted of four workload blocks, resulting in 12 blocks for each subject. Workload blocks were presented in an alternating fashion and balanced across the whole sample.

FIGURE 2 | Schematic representation of EMOBACK block structure. **(A)** 1-back condition. **(B)** 2-back condition. Hand symbols indicate key press. Left hand press was used for target stimuli and right hand press for distractors in this example. The simple shapes in this illustration only serve as an example.

From psychological perspective, the N-back task may conceptually be divided into two phases. The first phase starts with the stimulus presentation and ends after the stimulus response. In this phase we assume that subjects had to perform a simple matching task by comparing the current stimulus with the stimulus stored in short term memory. In the second phase, ranging from the stimulus response to the next stimulus onset we assume that multiple executive functions are required for a correct response. Inhibition, shifting and updating represent the core executive functions as identified by Miyake et al. (2000). Inhibition is seen as the ability to suppress automatic responses that might arise during task processing. Shifting is referring to the concept of cognitive flexibility, making us capable to switch between different tasks. Updating refers to the continuous monitoring and changing of content in working memory. Content in short term memory needs to be updated via inhibition of the last stimuli in the stimuli list. Furthermore, the participants need to switch from a simple matching task to a working memory task and back. We therefore assumed that the second phase was more relevant for the measurement of working memory load. Concerning the affective reaction we first assumed that it would be stronger right after stimulus onset, since the stimulus was fresh. However, pilot measurements revealed that subjects tend to focus first on the response toward the stimuli and only then direct their attention toward the affective content of the presented stimuli. We therefore decided to focus our analyses on the second phase. Additionally, we wanted to avoid contamination of the data due to muscular artifacts which originate from the keyboard input.

Analysis

The window of analysis was 1400 ms wide. It started 1100 ms after stimulus onset to exclude post-motor responses in the EEG and ended 2500 ms after stimulus onset. The time window also included 1000 ms of the ISI.

All spectra for the EEG analysis were computed using the Welch method (Welch, 1967) implemented in the EEGLAB toolbox (Delorme and Makeig, 2004). Theta bands were computed between 4 Hz and 7 Hz and alpha bands were computed between 8 Hz and 12 Hz. Power of the frontal (AFz, Fz) and parietal (CPz, Pz, POz) electrodes was averaged. Power for FAA computation was averaged across channels AF3, F3 and FC1 for the left hemisphere and across channels AF4, F4 and FC2 for the right hemisphere. We followed the approach from Allen et al. (2004) and computed FAA as difference score (see Equation 1).

Equation 1: Frontal Alpha Asymmetry Index (Allen et al., 2004).

$$FAA = right\ alpha\ power - left\ alpha\ power$$

To evaluate the influence of affective valence and working memory load on the EEG, we performed repeated measures analysis of variances (ANOVAs) with two factors. The first factor was affective valence with three levels (positive, neutral and

negative). The second factor was working memory load with two levels (1-back and 2-back). All analyses were conducted at the group level.

RESULTS

Subjective Measures—Working Memory Load

As expected, using workload ratings as dependent variable, we found a significant main effect for working memory load, $F_{(1,23)} = 22.7$, $p < 0.001$, $\eta^2 = 0.50$. Higher working memory load resulted in the increased subjective experience of working memory load. Additionally, there was a significant main effect for affective valence, $F_{(2,46)} = 11.6$, $p < 0.001$, $\eta^2 = 0.34$. A *Post hoc* test using the Šidák correction revealed that the negative valence condition resulted in an increase of subjective working memory load when compared to the positive valence condition ($p < 0.001$) as well as the neutral valence condition ($p < 0.02$). There was no significant interaction between affective valence and working memory load, $F_{(2,46)} = 0.8$, n.s. Box plots showing the different conditions can be seen in **Figure 3A**.

Subjective Measures—Affective Valence

Since, this manuscript was focusing on the effect of the experimental manipulation on the EEG measure used to infer affective valence, we do not report results from the arousal dimension here. As anticipated, using the affective valence ratings as dependent variable, there was a significant main effect for affective valence, $F_{(1.4,33)} = 55.7$, $p < 0.001$, $\eta^2 = 0.71$. Decreased affective valence due to the emotion induction resulted in decreased subjective valence ratings. The neutral condition resulted in more positive valenced ratings than the

negative condition ($p > 0.001$) and more negative valenced ratings than the positive condition ($p > 0.01$). However, there was neither a significant main effect for working memory load, $F_{(1,23)} = 0.1$, n.s., nor was there a significant interaction between affective valence and working memory load, $F_{(2,46)} = 1.0$, n.s. Median values for all conditions are shown via box plots in **Figure 3B**.

Behavioral Performance Measures—Accuracies

Using accuracies as dependent variable we found a significant main effect for working memory load, $F_{(1,23)} = 32.70$, $p < 0.001$, $\eta^2 = 0.59$. Higher working memory load resulted in decreased accuracies. Additionally, there was a significant main effect for affective valence, $F_{(2,46)} = 3.73$, $p < 0.035$, $\eta^2 = 0.14$. A *Post hoc* test using the Šidák correction revealed that the negative condition resulted in a decreased accuracy when compared to the positive condition ($p < 0.03$), but not when compared to the neutral condition (n.s.). Interestingly, there was a significant interaction between affective valence and working memory load, $F_{(2,46)} = 3.68$, $p < 0.035$, $\eta^2 = 0.14$. The negative condition had a negative impact on accuracy, but only under high working memory load. Box plots showing the different conditions can be seen in **Figure 4A**.

Behavioral Performance Measures—Reaction Times

There was a significant main effect for working memory load, $F_{(1,23)} = 60.53$, $p < 0.001$, $\eta^2 = 0.73$. Higher load resulted in longer reaction times. Moreover, there was a significant main effect for affective valence, $F_{(2,46)} = 10.94$, $p < 0.001$, $\eta^2 = 0.32$. The negative condition resulted in longer reaction times than the
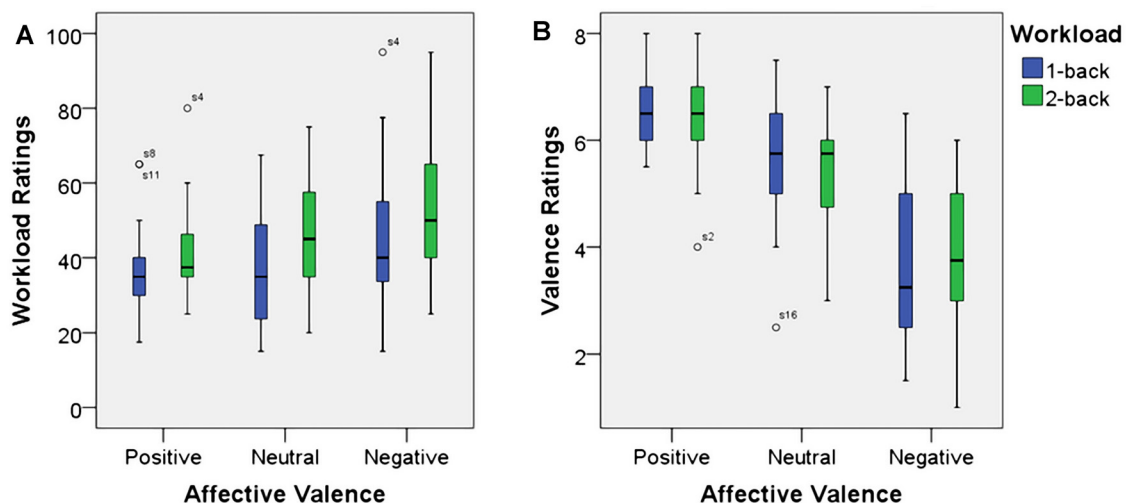
FIGURE 3 | Box plots showing median values for workload ratings **(A)** and affective valence ratings **(B)**. Blue boxes show working memory load ratings **(A)** and affective valence ratings **(B)** for the 1-back conditions. Green boxes show working memory load ratings **(A)** and affective valence ratings **(B)** for the 2-back conditions. Median values are indicated by black horizontal lines within the boxes. Top and bottom borders of the boxes represent the middle 50% of the data. Whiskers represent the smallest and largest values not classified as outliers (between 1.5 and 3 times the height of the boxes) or extreme values (more than three times the height of the boxes). Circles indicate outliers.

FIGURE 4 | Box plots showing median values for accuracies and reaction times. Blue boxes show accuracies **(A)** and reaction times **(B)** for the 1-back conditions. Green boxes show accuracies **(A)** and reaction times **(B)** for the 2-back conditions. Median values are indicated by black horizontal lines within the boxes. Top and bottom borders of the boxes represent the middle 50% of the data. Whiskers represent the smallest and largest values not classified as outliers (between 1.5 and 3 times the height of the boxes) or extreme values (more than three times the height of the boxes). Circles indicate outliers and stars show extreme values.

FIGURE 5 | Topographic plots showing difference in electroencephalography (EEG) workload measures between 2-back and 1-back conditions. **(A)** Frontal theta for the post-motor time window showing increased frontal theta activity for the 2-back conditions. **(B)** Parietal alpha for the Post motor time window showing decreased in parietal alpha power for the 2-back conditions. Nose is at the top. Values are averaged across all three affective conditions. Electrodes used for analysis are marked with black rectangles.

positive condition ($p < 0.01$) as well as the neutral ($p < 0.005$) condition. However, there was no significant interaction between affective valence and working memory load, $F_{(2,46)} = 1.02$, n.s. Median values for all conditions are shown via box plots in **Figure 4B**.

EEG Working Memory Load Measures

Frontal Theta

Using frontal theta activity as dependent variable there was a significant main effect for working memory load, $F_{(1,23)} = 9.00$, $p < 0.01$, $\eta^2 = 0.28$. Frontal theta power was higher in the 2-back
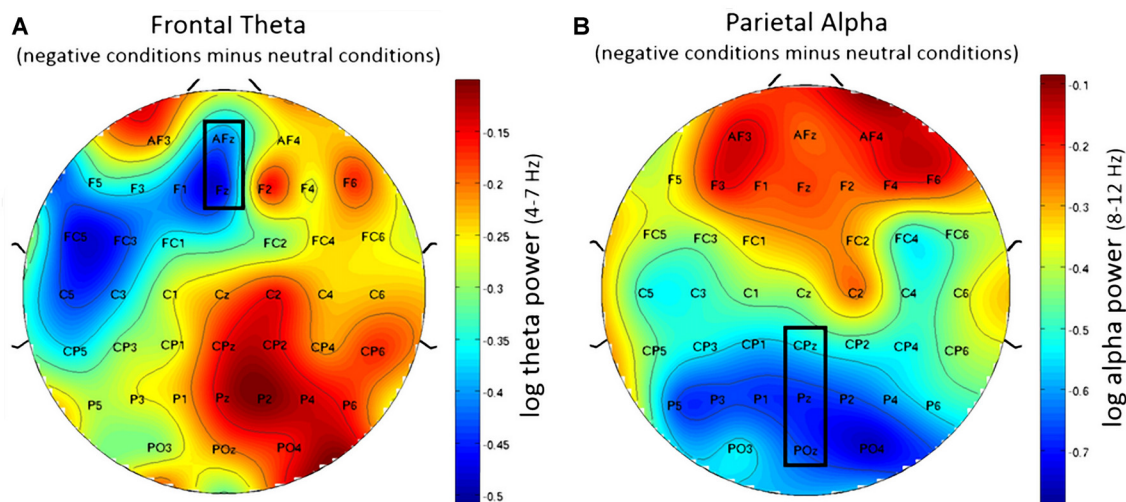
FIGURE 6 | Topographic plots showing difference in EEG workload measures between negative and neutral conditions. **(A)** Frontal theta showing decreased frontal theta activity for the negative conditions when compared to the neutral conditions. **(B)** Parietal alpha showing widespread decrease in parietal alpha power for the negative conditions when compared to the neutral conditions. Nose is at the top. Values are averaged across both N-back levels. Electrodes used for analysis are marked with black rectangles.
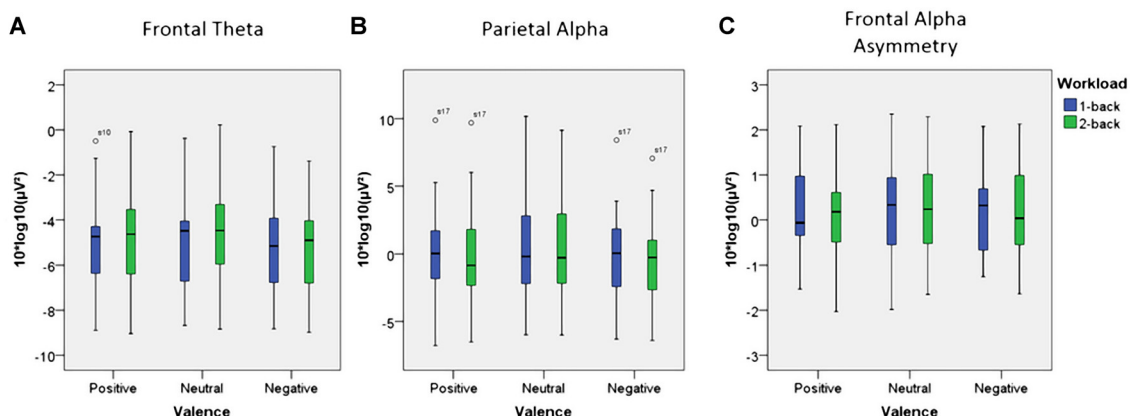
FIGURE 7 | Box plots showing median values for frontal theta activity, parietal alpha activity and Frontal Alpha Asymmetry (FAA). Blue boxes show frontal theta activity **(A)**, parietal alpha activity **(B)** and FAA **(C)** for the 1-back conditions. Green boxes show frontal theta activity **(A)**, parietal alpha activity **(B)** and FAA **(C)** for the 2-back conditions. Median values are indicated by black horizontal lines within the boxes. Top and bottom borders of the boxes represent the middle 50% of the data. Whiskers represent the smallest and largest values not classified as outliers (between 1.5 and 3 times the height of the boxes) or extreme values (more than three times the height of the boxes). Circles indicate outliers. Please note the strong variability in the data due to large inter-individual differences.

conditions. **Figure 5A** illustrates this effect. Interestingly, there was a significant main effect for affective valence, $F_{(2,46)} = 8.28$, $p < 0.001$, $\eta^2 = 0.27$. A *Post hoc* test using the Šidák correction revealed that frontal theta power was lower in the negative condition, when compared to the neutral condition ($p < 0.005$) as well as the positive condition ($p < 0.01$). **Figure 6A** shows a topographic plot displaying this frontal theta effect. However, there was no significant interaction between affective valence and working memory load, $F_{(2,46)} = 0.87$, n.s. Median values for all conditions can be seen in **Figure 7A**.

Parietal Alpha

Using parietal alpha activity as dependent variable we found a significant main effect for working memory load, $F_{(1,23)} = 4.22$, $p = 0.05$, $\eta^2 = 0.16$. Higher working memory load resulted in decreased parietal alpha power. This effect can be seen in **Figure 5B**. Interestingly, there was also a significant main effect for the factor affective valence in the post-motor window, $F_{(2,46)} = 5.27$, $p < 0.01$, $\eta^2 = 0.19$. The negative condition exhibited lower parietal alpha power than the neutral condition ($p < 0.03$), but not lower than the positive condition (n.s.).

Figure 6B shows a topographic plot displaying this parietal alpha effect. There was no significant interaction between affective valence and working memory load, $F_{(2,46)} = 1.52$, n.s. **Figure 7B** shows box plots for all conditions.

EEG Affective Valence Measures—Frontal Alpha Asymmetry

Unexpectedly, there were no significant main effects with regard to FAA. Neither for affective valence, $F_{(2,46)} = 1.87$, n.s. nor for working memory load, $F_{(1,23)} = 0.96$, n.s. Finally, there was also no significant interaction between affective valence and working memory, $F_{(2,46)} = 0.08$, n.s. Median values for all conditions are summarized in **Figure 7C**.

DISCUSSION

Using the emoback paradigm, we found that increased working memory load had a negative impact on task performance, as reflected in decreased accuracies and increased reaction times. These effects were also reflected in corresponding EEG measures. Increased working memory load was accompanied by increases in frontal theta activity as well as decreases in parietal alpha activity.

We found that negative affective valence had a negative impact on accuracies as well as reaction times. Interestingly, measures used for working memory load estimate also appeared sensitive to changes in affective valence. In contrast to that, FAA, a measure typically used to infer affective valence, did not show any effects in our paradigm, neither with regard to affective valence, nor with regard to increases in working memory load. In the following sections we will discuss these results in detail.

Subjective Measures

As expected, we found that increased working memory load, due to an increase in the load factor of the emoback task, resulted in higher subjective working memory load ratings. This finding is reassuring, especially since some subjects reported difficulties judging the working memory load with regard to the difficulty of the emoback task. Some subjects even reported that they experienced the 1-back conditions as more demanding, due to their monotonous nature.

As hypothesized, we also found a significant effect of affective valence on ratings from the valence dimension of the SAM. Notably, the successful induction of positive affective valence shows that the emoback paradigm worked in the intended way. The induction of positive emotions can be very difficult to achieve in laboratory environments, since positive emotions usually arise in a specific context which is difficult to represent in a single picture (Kim and Hamann, 2007).

Behavioral Performance Measures

Our analyses revealed that increased working memory load induced via the emoback task did result in decreased performance. The 2-back conditions had reduced accuracy as well as increased reaction times. We also found that inducing

negative affective valence had detrimental effects on accuracies as well as reaction times. Similar results have been found in a study by Passarotti et al. (2011). The authors used face stimuli in an affective N-back task and found slower reaction times for angry faces. These negative effects of affective stimuli on cognitive processing might be explained within the theory of hot and cold cognition, concepts related to executive function (Zelazo and Müller, 2002). Hot cognition refers to a process where a persons' thinking is influenced by their affective state (Brand, 1986), while cold cognition is more based on rational thinking and critical analysis (Roiser and Sahakian, 2013). Hot cognition seems to be able to overpower cold cognition in certain situations and has been shown to impair decision making (Huijbregts et al., 2008). Processing of negative stimuli can divert cognitive resources from the primary task, and thereby lead to decreased performance. A study by Levens and Gotlib (2010) found that strongly valenced stimuli tend to stay longer active in working memory, which might interfere with different core executive functions necessary during the emoback task (Miyake et al., 2000). In the emoback task, subjects constantly needed to update content in working memory by replacing (inhibition) previous stimuli. Additionally, the subjects seemed to need to shift between these updating tasks and a rather simple identity-matching task. Negative stimuli might catch the attention, slowing down reorganization of stimuli in working memory and thereby impairing task performance.

Interestingly, we also found an interaction for accuracy between working memory load and affective valence. Negative affective valence resulted in decreased accuracies, but only during increased working memory load. These results are in line with findings from the study by Kopf et al. (2013) who found more errors for the difficult task during the negative condition using an affective word N-back task. These findings seem to indicate that both, cognitive and affective processing, compete for limited cognitive resources and this dual strain on mental resources leads to decreased performance. Similar results have also been found in the study by MacNamara et al. (2011). The authors used emotional pictures as distractors in an emoback task and found that the emotional content of negative valenced stimuli increased the negative impact of working memory load on performance. These findings might be explained using the capacity model from Ellis and Ashbrook (1989). Their model assumes that there is a limited pool of attentional resources that can be used to complete a certain task. Affective states can influence the allocation of available attentional resources toward the task, thereby potentially impairing task performance. However, based on performance measures alone, we can only vaguely assume how negative valence impaired task performance.

EEG Working Memory Load Measures
Frontal Theta Activity

Increased working memory load was reflected in EEG measures as well. Frontal theta activity was increased during high working memory load. Interestingly, the negative affective

valence condition exhibited decreased frontal theta after the motor response. If negative affective valence would impair performance via the production of additional working memory load, one would expect to find increased frontal theta activity during negative affective valence. However, the decreased frontal theta activity under negative affective valence indicates that the detrimental effect on performance likely has different reasons than the performance decrease under high working memory load. Accordingly, we assume that the negative affective valenced stimuli interfered with task processing through a reduction of activity in the frontal attentional control network.

Parietal Alpha Activity

Our analyses revealed that parietal alpha activity was decreased under high working memory load (2-back) condition. We also found that the parietal alpha activity was reduced for the negative condition. An fMRI study by Rämä et al. (2001) used different affective voices and found similar results. The authors discovered that the parietal cortex is bilaterally involved in active maintenance of emotional content. Affective stimuli are more salient and usually carry more relevant information (Carretié, 2014). We therefore assume that negative valenced stimuli result in stronger processing in the storage areas of the parietal cortex, which seems to be reflected through increased parietal activity (i.e., reduced alpha power).

EEG Affective Valence Measures—Frontal Alpha Asymmetry

Unexpectedly, we did not find any effect using the FAA measure. We used the FAA measure during the induction of working memory load, while most other studies used brain laterality responses during rest or passive viewing (Ahern and Schwartz, 1985; Tomarken et al., 1992; Lin et al., 2009; Huang et al., 2012; Ramirez and Vamvakousis, 2012). Use of the FAA in combination with other tasks is further complicated because the FAA is known to be also influenced by other factors like seating position and working memory load (Briesemeister et al., 2013). A study by Baldwin and Penaranda (2012) found that increased task difficulty resulted in increased left frontal activity. It appears that the negative condition increased task difficulty, since it did impair task performance. Negative affective valence in passive conditions usually results in increased right frontal activity. It is therefore conceivable that both, affective and workload related, processes acted on the same FAA measure, but in opposing directions. This could mean that both effects canceled each other out and thereby masked any potential effects.

Limitations and Outlook

We did not find any effects using the FAA measure. We assume that was due to the motor response required in our paradigm. It would be interesting to test this assumption by developing a study design that contrasts an active condition with a passive viewing condition. However, it could also be possible, that the paradigm did not induce sufficient stress to

induce changes in the FAA measure. A study by Goodman et al. (2013) used a working memory task and concurrently induced stress. The authors found that changes in the FAA measure (as indicated by left frontal activity) can be used to infer emotional processing, but only when the emotional induction reaches a certain intensity. Future studies could try to use other ways to induce emotional reactions along the valence dimension, like short video clips.

Additionally, our intent to induce the strongest possible valence effect had the drawback that the affective conditions also differed concerning the arousal dimension. This meant that the negative conditions were also experienced as more arousing. Since even this valence induction did not produce a valence effect that could be measured with the FAA, we recommend that future studies should control for arousal as well as other dimensions like affective dominance.

Furthermore, we based our choice of the window used for analysis on our understanding of the existent literature. However our assumptions need further support and therefore future studies should investigate the different sub-processes that contribute to the n-back task and their temporal evolution in greater detail.

While we decided to focus on the use of a frequency based approach, future approaches could also include different features for the measurement of affective valence and working memory load. One example would be the use of the ERP, since previous studies have already demonstrated that the ERP has some potential in this context (see Olofsson et al., 2008; Brouwer et al., 2012). Another approach could be the use of connectivity measures, which have already been shown to be of use in similar contexts (e.g., Lee and Hsieh, 2014). A study by Martini et al. (2012) has even combined frequency measures, ERP measures and connectivity measures to differentiate neutral from negative pictures.

Future studies should also explore how well these findings generalize to male individuals as well as other stimulus modalities.

Finally, in this study we focused on the interaction between working memory load and affective valence, using established measures that were already used in the context of human-machine interaction. Future studies should further investigate this issue to gain more insight about the processes underlying the results that were extracted from the EEG-recordings. One example for such an integrative approach is the review by Schwabe et al. (2012) that integrates multiple findings into a framework that allows to create hypotheses with regard to specific neural structures. Something which is beyond the methodology used in this study.

CONCLUSION

We demonstrated that increased working memory load and negative valenced stimuli can have an effect on performance measures. Additionally, when using established EEG measures we found that increased working memory load can be detected in the EEG, even when affective valence is induced at the same time. However, EEG measures used to infer working

memory load were still influenced by negative affective valence. Furthermore, FAA did not prove useful for identification of emotional states when working memory load is induced at the same time. Therefore, future studies should further investigate the context sensitivity and applicability of EEG measures in various contexts to identify the ramifications in which such measures can be used to identify different states and processes.

AUTHOR CONTRIBUTIONS

SG is the main author and responsible for all tasks. JF was involved in the planning of the experiment and also provided technical advice. CS and MS helped with general advice. PG is the main coordinator and also involved in all related tasks.

REFERENCES

Ahern, G. L., and Schwartz, G. E. (1985). Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis. *Neuropsychologia* 23, 745–755. doi: 10.1016/0028-3932(85)90081-8

Alguacil, S., Tudela, P., and Ruz, M. (2013). Cognitive and affective control in a flanker word task: common and dissociable brain mechanisms. *Neuropsychologia* 51, 1663–1672. doi: 10.1016/j.neuropsychologia.2013.05.020

Allen, J. J. B., Coan, J. A., and Nazarian, M. (2004). Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biol. Psychol.* 67, 183–218. doi: 10.1016/j.biopsycho.2004.03.007

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., and Onaral, B. (2007). "Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy," in *Proceedings of the 3rd International IEEE/EMBS Conference on Neural Engineering* (Hohala Coast, Hawaii), 342–345. doi: 10.1109/CNE.2007.369680

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Baldwin, C. L., and Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *Neuroimage* 59, 48–56. doi: 10.1016/j.neuroimage.2011.07.047

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* 25, 49–59. doi: 10.1016/0005-7916(94)90063-9

Brand, A. G. (1986). Hot cognition: emotions and writing behavior. *J. Adv. Compos.* 6, 5–15.

Briesemeister, B. B., Tamm, S., Heine, A., and Jacobs, A. M. (2013). Approach the good, withdraw from the bad—a review on frontal alpha asymmetry measures in applied psychological research. *Psychology* 4, 261–267. doi: 10.4236/psych.2013.43a039

Brouwer, A. M., Hogervorst, M. A., van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Carretié, L. (2014). Exogenous (automatic) attention to emotional stimuli: a review. *Cogn. Affect. Behav. Neurosci.* 14, 1228–1258. doi: 10.3758/s13415-014-0270-2

Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain Cogn.* 20, 125–151. doi: 10.1016/0278-2626(92)90065-t

De Smedt, T., and Menschaert, L. (2012). VALENCE: affective visualisation using EEG. *Digit. Creat.* 23, 272–277. doi: 10.1080/14626268.2012.719240

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component

anlaysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Delorme, A., Mullen, T., Kothe, C., Akalin Acar, Z., Bigdely-Shamlo, N., Vankov, A., et al. (2011). EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing. *Comput. Intell. Neurosci.* 2011:130714. doi: 10.1155/2011/130714

Ellis, H. C., and Ashbrook, P. W. (1989). The "state" of mood and memory research: a selective review. *J. Soc. Behav. Pers.* 4, 1–21.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interact. Comput.* 21, 133–145. doi: 10.1016/j.intcom.2008.10.011

Frey, J., Mühl, C., Lotte, F., and Hachet, M. (2013). Review of the use of electroencephalography as an evaluation method for human-computer interaction. *arXiv* preprint arXiv:1311.2222.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Front. Neurosci.* 8:385. doi: 10.3389/fnins.2014.00385

Gerrards-Hesse, A., Spies, K., and Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: a review. *Br. J. Psychol.* 85, 55–78. doi: 10.1111/j.2044-8295.1994.tb02508.x

Gevins, A., Leong, H., Du, R., Smith, M. E., Le, J., DuRousseau, D., et al. (1995). Towards measurement of brain function in operational environments. *Biol. Psychol.* 40, 169–186. doi: 10.1016/0301-0511(95)05105-8

Goodman, R. N., Rietschel, J. C., Lo, L. C., Costanzo, M. E., and Hatfield, B. D. (2013). Stress, emotion regulation and cognitive performance: the predictive contributions of trait and state relative frontal EEG alpha asymmetry. *Int. J. Psychophysiol.* 87, 115–123. doi: 10.1016/j.ijpsycho.2012.09.008

Grissmann, S., Spüler, M., Faller, J., Krumpe, T., Zander, T. O., Kelava, A., et al. (in press). "Context sensitivity of EEG-based workload classification under different affective valence," in *IEEE Transactions on Affective Computing*.

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/s0166-4115(08)62386-9

Huang, D., Zhang, H., Ang, K., Guan, C., Pan, Y., Wang, C., et al. (2012). "Fast emotion detection from EEG using asymmetric spatial filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Kyoto, Japan: IEEE), 589–592. doi: 10.1109/ICASSP.2012.6287952

Huijbregts, S. C. J., Warren, A. J., de Sonneville, L. M. J., and Swaab-Barneveld, H. (2008). Hot and cool forms of inhibitory control and externalizing behavior in children of mothers who smoked during pregnancy: an exploratory study. *J. Abnorm. Child Psychol.* 36, 323–333. doi: 10.1007/s10802-007-9180-x

Kim, S. H., and Hamann, S. (2007). Neural correlates of positive and negative emotion regulation. *J. Cogn. Neurosci.* 19, 776–798. doi: 10.1162/jocn.2007.19.5.776

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Comput. Math. Methods Med.* 2013:573734. doi: 10.1155/2013/573734

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Kirschbaum, C., Pirke, K. M., and Hellhammer, D. H. (1993). The 'trier social stress test'—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 76–81. doi: 10.1159/000119004

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/s0165-0173(98)00056-3

Klingberg, T. (2009). *The Overflowing Brain Information Overload and the Limits of Working Memory.* Oxford: Oxford University Press.

Kopf, J., Dresler, T., Reicherts, P., Herrmann, M. J., and Reif, A. (2013). The effect of emotional content on brain activation and the late positive potential in a word N-back task. *PLoS One* 8:e75598. doi: 10.1371/journal.pone.0075598

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). "International affective picture system (IAPS): technical manual and affective ratings," in *NIMH Center for the Study of Emotion and Attention* (Gainesville, FL), 39–58. doi: 10.1027/0269-8803/a000147

Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 261–273. doi: 10.1111/j.1469-8986.1993.tb03352.x

Lee, Y. Y., and Hsieh, S. (2014). Classifying different emotional states by means of EEG-based functional connectivity patterns. *PLoS One* 9:e95415. doi: 10.1371/journal.pone.0095415

Levens, S. M., and Gotlib, I. H. (2010). Updating positive and negative stimuli in working memory in depression. *J. Exp. Psychol. Gen.* 139, 654–664. doi: 10.1037/a0020283

Leventhal, A. M., Martin, R. L., Seals, R. W., Tapia, E., and Rehm, L. P. (2007). Investigating the dynamics of affect: psychological mechanisms of affective habituation to pleasurable stimuli. *Motiv. Emot.* 31, 145–157. doi: 10.1007/s11031-007-9059-8

Lin, Y. P., Wang, C.-H., Wu, T. L., Jeng, S.-K., and Chen, J.-H. (2009). "EEG-based emotion recognition in music listening: a comparison of schemes for multiclass support vector machine," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Taipei, Taiwan: IEEE), 489–492. doi: 10.1109/ICASSP.2009.4959627

MacNamara, A., Ferri, J., and Hajcak, G. (2011). Working memory load reduces the late positive potential and this effect is attenuated with increasing anxiety. *Cogn. Affect. Behav. Neurosci.* 11, 321–331. doi: 10.3758/s13415-011-0036-z

Martini, N., Menicucci, D., Sebastiani, L., Bedini, R., Pingitore, A., Vanello, N., et al. (2012). The dynamics of EEG gamma responses to unpleasant visual stimuli: from local activity to functional connectivity. *Neuroimage* 60, 922–932. doi: 10.1016/j.neuroimage.2012.01.060

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48, 229–240. doi: 10.1111/j.1469-8986.2010.01061.x

Mühl, C., Jeunet, C., and Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Front. Neurosci.* 8:114. doi: 10.3389/fnins.2014.00114

Müller, M. M., Keil, A., Gruber, T., and Elbert, T. (1999). Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clin. Neurophysiol.* 110, 1913–1920. doi: 10.1016/s1388-2457(99)00151-0

Olofsson, J. K., Nordin, S., Sequeira, H., and Polich, J. (2008). Affective picture processing: an integrative review of ERP findings. *Biol. Psychol.* 77, 247–265. doi: 10.1016/j.biopsycho.2007.11.006

Passarotti, A. M., Sweeney, J. A., and Pavuluri, M. N. (2011). Fronto-limbic dysfunction in mania pre-treatment and persistent amygdala over-activity post-treatment in pediatric bipolar disorder. *Psychopharmacology* 216, 485–499. doi: 10.1007/s00213-011-2243-2

Raimondo, F., Kamienkowski, J. E., Sigman, M., and Fernandez Slezak, D. (2012). CUDAICA: GPU optimization of infomax-ICA EEG analysis. *Comput. Intell. Neurosci.* 2012:206972. doi: 10.1155/2012/206972

Rämä, P., Martinkauppi, S., Linnankoski, I., Koivisto, J., Aronen, H. J., and Carlson, S. (2001). Working memory of identification of emotional vocal expressions: an fMRI study. *Neuroimage* 13, 1090–1101. doi: 10.1006/nimg.2001.0777

Ramirez, R., and Vamvakousis, Z. (2012). "Detecting emotion from EEG signals using the emotive epoc device," in *Brain Informatics*, eds F. M. Zanzotto, S. Tsumoto, N. Taatgen and Y. Yao (Berlin, Heidelberg: Springer), 175–184. doi: 10.1007/978-3-642-35139-6_17

Roiser, J. P., and Sahakian, B. J. (2013). Hot and cold cognition in depression. *CNS Spectr.* 18, 139–149. doi: 10.1017/s1092852913000072

Rowland, N., Meile, M. J., and Nicolaidis, S. (1985). EEG alpha activity reflects attentional demands and beta activity reflects emotional and cognitive processes. *Science* 228, 750–752. doi: 10.1126/science.3992243

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2013). "Mental fatigue and working memory load estimation: interaction and implications for EEG-based passive BCI," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* (Osaka, Japan: IEEE), 6607–6610. doi: 10.1109/EMBC.2013.6611070

Russell, J. A., and Pratt, G. (1980). A description of the affective quality attributed to environments. *J. Pers. Soc. Psychol.* 38, 311–322. doi: 10.1037/0022-3514.38.2.311

Scharinger, C., Soutschek, A., Schubert, T., and Gerjets, P. (2015). When flanker meets the N-back: what eeg and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology* 52, 1293–1304. doi: 10.1111/psyp.12500

Schupp, H. T., Cuthbert, B. N., Bradley, M. M., Cacioppo, J. T., Ito, T., and Lang, P. J. (2000). Affective picture processing: the late positive potential is modulated by motivational relevance. *Psychophysiology* 37, 257–261. doi: 10.1017/s0048577200001530

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., and Oitzl, M. S. (2012). Stress effects on memory: an update and integration. *Neurosci. Biobehav. Rev.* 36, 1740–1749. doi: 10.1016/j.neubiorev.2011.07.002

Smith, M. E., and Gevins, A. (2005). "Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator," in *Proceedings of the SPIE 5797 Defense and Security, Symposium Biomonitoring for Physiological and Cognitive Performance during Military Operations*, eds J. A. Caldwell and N. J. Wesensten (Orlando: SPIE), 116–126. doi: 10.1117/12.602181

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., and Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 267–278. doi: 10.1007/s11858-015-0754-8

Tomarken, A. J., Davidson, R. J., and Henriques, J. B. (1990). Resting frontal brain asymmetry predicts affective responses to films. *J. Pers. Soc. Psychol.* 59, 791–801. doi: 10.1037/0022-3514.59.4.791

Tomarken, A. J., Davidson, R. J., Wheeler, R. E., and Doss, R. C. (1992). Individual differences in anterior brain asymmetry and fundamental dimensions of emotion. *J. Pers. Soc. Psychol.* 62, 676–687. doi: 10.1037/0022-3514.62.4.676

Welch, P. D. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electro.* 15, 70–73. doi: 10.1109/tau.1967.1161901

Wilhelm, O., Hildebrandt, A., and Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Front. Psychol.* 4:433. doi: 10.3389/fpsyg.2013.00433

Zelazo, P. D., and Müller, U. (2002). "Executive function in typical and atypical development," in *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami (Oxford: Wiley-Blackwell), 574–603. doi: 10.1002/9780470996652.ch20

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Measuring Mental Workload with EEG+fNIRS

Haleh Aghajani[1]*, Marc Garbey[2] and Ahmet Omurtag[1]†

[1] Department of Biomedical Engineering, University of Houston, Houston, TX, United States, [2] Center for Computational Surgery, Department of Surgery, Research Institute, Houston Methodist, Houston, TX, United States

We studied the capability of a Hybrid functional neuroimaging technique to quantify human mental workload (MWL). We have used electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) as imaging modalities with 17 healthy subjects performing the letter *n*-back task, a standard experimental paradigm related to working memory (WM). The level of MWL was parametrically changed by variation of *n* from 0 to 3. Nineteen EEG channels were covering the whole-head and 19 fNIRS channels were located on the forehead to cover the most dominant brain region involved in WM. Grand block averaging of recorded signals revealed specific behaviors of oxygenated-hemoglobin level during changes in the level of MWL. A machine learning approach has been utilized for detection of the level of MWL. We extracted different features from EEG, fNIRS, and EEG+fNIRS signals as the biomarkers of MWL and fed them to a linear support vector machine (SVM) as train and test sets. These features were selected based on their sensitivity to the changes in the level of MWL according to the literature. We introduced a new category of features within fNIRS and EEG+fNIRS systems. In addition, the performance level of each feature category was systematically assessed. We also assessed the effect of number of features and window size in classification performance. SVM classifier used in order to discriminate between different combinations of cognitive states from binary- and multi-class states. In addition to the cross-validated performance level of the classifier other metrics such as sensitivity, specificity, and predictive values were calculated for a comprehensive assessment of the classification system. The Hybrid (EEG+fNIRS) system had an accuracy that was significantly higher than that of either EEG or fNIRS. Our results suggest that EEG+fNIRS features combined with a classifier are capable of robustly discriminating among various levels of MWL. Results suggest that EEG+fNIRS should be preferred to only EEG or fNIRS, in developing passive BCIs and other applications which need to monitor users' MWL.

Keywords: functional near-infrared spectroscopy (fNIRS), electroencephalography (EEG), human mental workload, cognitive state monitoring, *n*-back, multi-modal brain recording, machine learning

INTRODUCTION

Mental workload (MWL) affects people who are interacting with computers and other devices. The use of technology in everyday life may impose high cognitive demands as users navigate complex interfaces. Mental overload may compromise users' performance and sometimes safety, by increasing error rates and engendering fatigue, decline in motivation, higher reaction times

(Xie and Salvendy, 2000; Young and Stanton, 2002), and neglect of critical information, known as cognitive tunneling (Thomas and Wickens, 2001; Dixon et al., 2013; Dehais et al., 2014). Taking into account the users' cognitive characteristics and limitations are thus critical in improving the design of human-machine interfaces (BMI) and for operating them efficiently by installing adaptive features that can respond to changes in the MWL (Kaber et al., 2000; Parasuraman and Wilson, 2008; Gagnon et al., 2012).

MWL has been defined as the proportion of the human operator's mental capabilities that is occupied during the performance of a given task (Boff et al., 1994). According to the prevalent Multiple Resources theory (Navon and Gopher, 1979; Wickens, 2002), performing different tasks requires a subject to tap into a set of separate resources, which are limited in capacity and distributable among tasks (Horrey and Wickens, 2003). In general, these resources can be categorized among four dimensions: processing stage (perception or cognition vs. response), perceptual modality (visual vs. auditory), visual channel (focal vs. ambient), and processing code (verbal vs. spatial; Wickens, 2002; Horrey and Wickens, 2003). Based on the Multiple Resources theory, equal resource demands between two tasks that both recruit one level of a given dimension will interfere with each other more than two tasks that recruit separate levels on the dimension (Wickens, 2002), and may create bottlenecks and consequent decrements in performance. Similar conclusions have been reached in the areas of aviation (Stanney and Hale, 2012; Causse and Matton, 2014; Durantin et al., 2014), education (Palmer and Kobus, 2007; Spüler et al., 2016), and a variety of clinical situations (Carswell et al., 2005; Stefanidis et al., 2007; Prabhu et al., 2010; Yurko et al., 2010; Byrne, 2013; Guru et al., 2015). In the case of driving while having a phone conversation, in addition to the interference of resources the "engagement phenomenon" also controls the outcome of multi-tasking scenario. This happens when one of the tasks attracts so much attention that the advantage of separate resource demand would be eliminated (Strayer and Johnston, 2001; Strayer and Drews, 2007).

MWL is a construct that arises from the interaction of the properties of a task, the environment in which it is performed, and the characteristics of the human operator performing it (Longo, 2016). Task properties include the difficulty and monotony of the task and the types of resources that it engages. The environment may contain various degrees of distraction and noise. The subject characteristics involve training and expertise as well as changing levels of fatigue, motivation, and vigilance. Thus, the MWL can be systematically adjusted by tuning a subset of these variables while controlling for the rest.

Methods of determining MWL fall into three broad categories: (1) Self-reporting and subjective ratings using standard questionnaires such as the NASA-TLX (Hart and Staveland, 1988); (2) Behavioral measures, such as primary- and secondary-task performance; and (3) Measures based on the physiology of the user, including heart rate variability, oculomotor activity, pupillometry, electromyography, galvanic skin response, and brain activity (Xiao et al., 2005; Wickens, 2008; Sahayadhas et al., 2012). Self-reporting and behavioral based information tends to be delayed, sporadic, and intrusive

to obtain. Performance based information, in addition, can be misleading since multiple degrees of MWL may accompany the same level of performance (Yurko et al., 2010). Physiological measures, on the other hand, do not require overt behavior, can be arranged to have little or no interference with task execution, and can supply information continuously without significant delay. Progresses in miniaturization and wireless technology have amplified these advantages of physiological measures (Sahayadhas et al., 2012).

Most studies of MWL based on brain function have utilized electroencephalography (EEG), following a large number of studies using EEG for developing BMI (Wolpaw and Wolpaw, 2012). Functional near-infrared spectroscopy (fNIRS) as a newer modality have shown promising capabilities in BMI applications for discrimination of different motor tasks (Naseer and Hong, 2013) or decoding subjects' binary decisions (Naseer et al., 2014). The relationship between MWL and central nervous system activity is well-established (McBride and Schmorrow, 2005). BMIs that do not attempt to directly control a device but modulate its user interface based on real-time user status are referred as passive BMIs (Gateau et al., 2015). In such applications, of paramount importance. Recently multi-modal techniques utilizing concurrent EEG and fNIRS have gained popularity due to their relatively richer information content (Hirshfield et al., 2009; Liu Y. et al., 2013; Liu T. et al., 2013; Keles et al., 2014b; Aghajani and Omurtag, 2016; Buccino et al., 2016; Omurtag et al., 2017). Available evidence indicates that brain activity measures of MWL are more informative than ocular or peripheral physiology measures (Hogervorst et al., 2014).

Concurrent EEG and fNIRS, which we refer to as EEG+fNIRS, is promising as a practical technique that is more accurate than the individual modalities alone. fNIRS provides information that is complementary to EEG, by measuring the changes in cerebral blood flow (CBF) and related hemoglobin concentrations through near-infrared light source/detectors on the scalp. It is comparable to EEG in portability. fNIRS does not have electromyographic (EMG) and blink artifacts and its signal closely correlates with the blood oxygen level dependent (BOLD) signal from functional magnetic resonance imaging (fMRI; Strangman et al., 2002; Huppert et al., 2006), which is a gold-standard for measuring cerebral hemodynamics. In addition to the advantages of pooling different types of signals, EEG+fNIRS offers new types of features, ultimately based on neurovascular coupling (NVC), the cascade of processes by which neural activity modulates local blood flow and oxygenation, and NVC related features are not resolvable by a uni-modal signal sensitive to only neural activity (e.g., EEG) or only hemodynamics (e.g., BOLD).

Working memory (WM) is a brain system that provides transient holding and processing of the information necessary for complex cognitive tasks (Baddeley, 2003). It has been investigated in previous functional neuroimaging studies, which identified the prefrontal cortex (PFC) as the most relevant area of activation (Cohen et al., 1997; Smith and Jonides, 1997; Hoshi et al., 2003; Owen et al., 2005). MWL detection using WM load as an experimental paradigm has been studied using EEG (Berka et al., 2007; Dornhege et al., 2007; Grimes et al., 2008; Brouwer

et al., 2012), fNIRS (Hoshi et al., 2003; Izzetoglu et al., 2003; Ayaz et al., 2012; Durantin et al., 2014; Herff et al., 2014), and concurrent EEG and fNIRS (Hirshfield et al., 2009; Coffey et al., 2012). We have previously shown that aspects of NVC are characterizable by EEG+fNIRS, by taking advantage of the synergistic interaction between the modalities (Keles et al., 2016). The potential of EEG+fNIRS for active BCIs has recently been investigated (Fazli et al., 2012; Liu Y. et al., 2013; Tomita et al., 2014; Buccino et al., 2016). In this study, we built on this work to explore the unique properties of EEG+fNIRS for MWL detection.

The n-back task was introduced by Kirchner (1958). n-back is a continues-performance task for measurement of WM capacity, which has been used frequently in the field of cognitive neuroscience. Gevins (Gevins et al., 1997, 1998; Smith and Gevins, 2005a) and Smith (Smith and Gevins, 2005b) showed during high task-load conditions EEG theta activity increases in the frontal midline and alpha activity attenuates during the performance of an n-back task. In addition, fNIRS revealed WM load while performing n-back task activates PFC (Ayaz et al., 2012; Sato et al., 2013; Fishburn et al., 2014; Herff et al., 2014; Mandrick et al., 2016b). The n-back task engages WM and becomes more demanding as the value of n increases. We have therefore used the n-back task as our experimental paradigm with n ranging from 0 to 3, allowing us to tune the task difficulty. Our study maintained all other conditions constant by employing only healthy adult volunteers with no previous experience with this task, all performing under the same laboratory conditions with no distractions or additional activities. Our experimental design is consistent with numerous BMI studies that take WM load as a proxy for MWL (Berka et al., 2007; Grimes et al., 2008; Coffey et al., 2012; Herff et al., 2014).

The aim of this study was, first, to introduce and validate a state of the art EEG+fNIRS set up in a single headpiece. Since PFC is the main region of interest in WM load detection (Owen et al., 2005), our design had the advantage of frontal lobe coverage by fNIRS. We had whole-head coverage by EEG. We used the term whole-head to refer to the fact that we placed EEG electrodes at all (except frontopolar) standard 10–20 sites bilaterally covering the frontal, central, temporal, parietal, and occipital areas. Having fNIRS optodes on the forehead not only improved the quality of acquired signal but also reduced the preparation time. The second aim of this study was to develop EEG+fNIRS measures that discriminate among levels of MWL and show that they are promising for the practical and accurate quantification of MWL in real-world settings. We developed such measures by extracting EEG, fNIRS, and Hybrid (EEG+fNIRS) based features from the full set of signals. Most discriminating features were selected and fed into support vector machines (SVM) to perform binary or multi-class classification. The handful of EEG+fNIRS studies currently available (Coffey et al., 2012; Putze et al., 2014; Buccino et al., 2016) have not systematically quantified the performance of subsets of a Hybrid system and its features contribute to the accuracy of classification. Therefore, the third aim of this study was to rigorously compare the performance of uni-modal and Hybrid systems.

METHODS

Subjects

Seventeen healthy volunteers (16 males, 1 female) with a mean age of 26.2 and standard deviation of 7.7 years from University of Houston students or employees participated in the experiment. The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board of the University of Houston. The participants gave written informed consent prior to the experiments and were compensated for their effort by being given a gift card from a major retailer. During the performance of the verbal n-back task, target letters should be detected by the operator by means of pressing Space button on the keyboard. All subjects were right-handed and used their dominant hand for performing the experiment. This will reduce the variability of brain signals based on the motor function through all subjects. None of the subjects had ever taken part in an n-back study, thus no training effects were expected.

Experimental Design

One of the most common WM paradigm for MWL assessment is the n-back task, which was first introduced by Kirchner (1958). In the letter n-back task, participants observe a sequence of single letters separated by a certain amount of time each; for each letter they decide whether it is a target, i.e., identical to the item that appeared n items back in the sequence. The value of n is kept constant during a segment of the experiment referred to as a session. As n increases the difficulty of the task becomes higher. In the literature usually 0-back task has been used as a control state. **Figure 1** illustrates how letter n-back task works when n is 0, 1, 2, or 3. Depending on n, subject should find the target letter and interact with the user-interface.

In each experiment, we had a total number of 40 sessions. These sessions were presented in pseudorandom order, 10 sessions per n. Each session started with an instruction block that is displayed for 5 s on the screen and informed the subject about which type of the n-back tasks was about to start (instruction block). Then 22 randomly selected letters (out of 10 candidate pool of letters) appeared in sequence on the screen (task block). Each letter stayed on the screen for 500 ms and the subject had 1,500 ms to press the space button in case that the letter was a target according to the type of session. At the end of each session there was a 25 s resting block. During this block the subject remained relaxed and fixated at a cross on the screen to let the brain activation return to its baseline and get ready for the next n-back session (Herff et al., 2014). **Figure 2** shows one sample session. Total recording time was 50 min. The program for implementation of this task was written using Presentation software (Neurobehavioral Systems, Inc.). All the information about the appearance time of each letter, session type, subject's response time, and also whether the presented letter was a target or not was recorded by this software and stored as a text file for later processing. The objective performance of the subjects within each session was computed from this information. Subjects who had too low accuracy (<90% in the 0- or <80% in the 1-back) were deemed insufficiently focused on the task. Performance
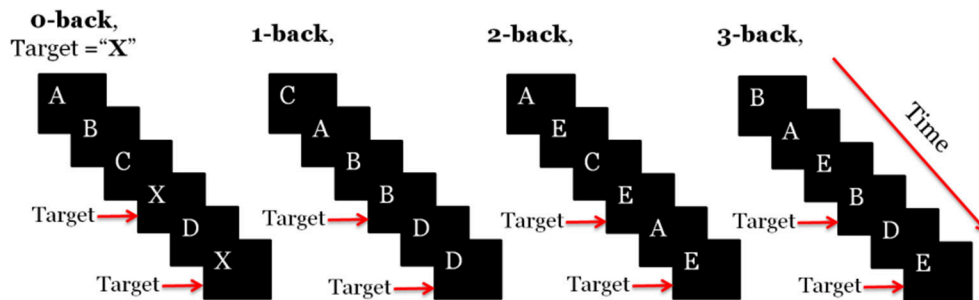
FIGURE 1 | Schematic illustration of the letter *n*-back task for *n* ϵ {0, 1, 2, 3}.
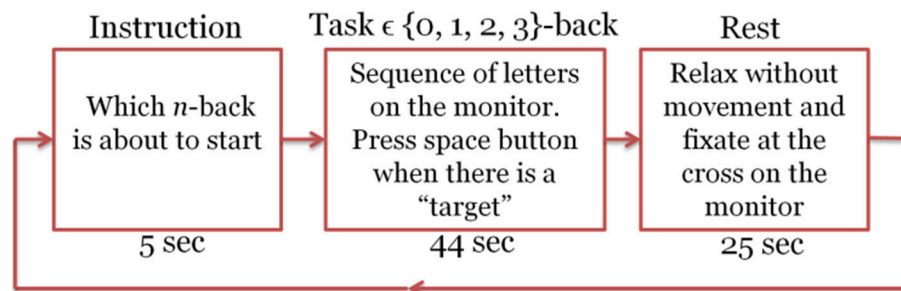
FIGURE 2 | Experimental design for the letter *n*-back task. Each session includes the Instruction, task, and rest blocks.

level was measured by computing the accuracy defined as the fraction of correct responses. We considered a missed target as an incorrect response.

Data Acquisition and Preprocessing

A quantitative meta-analysis has found the cortical regions that were activated robustly during letter *n*-back task (Broadman Areas 6, 7, 8, 9, 10, 32, 40, 45, 46, 47, and supplementary motor area; Owen et al., 2005). We used this information together with the results of previous EEG studies to choose the optimum locations for our 19 EEG electrodes (F7, F8, F3, F4, Fz, Fc1, Fc2, T3, T4, C3, C4, Cp1, Cp2, P3, P4, Pz, Poz, O1, O2). We used Fpz as the ground and Cz as the reference electrode. In the literature, several different reference electrode positioning is indicated, which have their own set of strengths and weaknesses. Among them, linked ears and vertex (Cz) are the most common. Cz reference is advantageous when it is located in the middle among active electrodes, however, for close points, it may result in poor resolution (Teplan, 2002). Based on the previous studies, central brain region in not majorly involved during the performance of a WM task compared to the frontal and parietal lobes and choosing Cz as the reference may be more appropriate rather than any other electrode in the 10–20 system. microEEG (a portable device made by Bio-Signal Group Inc., Brooklyn, New York) was used to sample EEG at 250 Hz (**Figure 3a**). Electrode impedances were kept below 10 kΩ. A 128-channel electrode cap with Ag/AgCl electrodes (EasyCap, Germany) was used to physically stabilize the sensors and provide uniform scalp coverage. We located the fNIRS optodes on the subject's forehead to fully cover the PFC,

which plays a significant role in WM (Fitzgibbon et al., 2013). Seven sources and seven detectors were located on the forehead resulting in 19 optical channels, each consisting of a source–detector (S–D) pair separated by a distance of 3 cm. The 19 optical channels used in this study are shown in **Figure 3c**. The S–D placement starts from the left hemisphere and ends on the right hemisphere. S4 and D4 are located at the center of forehead, where D4 is located at the AFz location and channel 10 is located at the Fpz location according to the standard international 10–20 system (**Figures 3b,c**). We used our triplet-holders (Keles et al., 2014a) on the forehead to keep each EEG electrode in the middle of an S–D pair and fix the distances between the sensors. fNIRS signals were acquired at 8.93 Hz via NIRScout extended (NIRx Medical Technologies, New York) device, which was synchronized with the EEG data by means of common event triggers (**Figure 3a**). NIRScout is a dual wavelength continuous wave system. The EEG signal was band-pass filtered (0.5–80 Hz), and a 60 Hz notch filter was used to reduce the power line noise.

The spatial Laplacian transform is generally effective in muscle artifact removal from EEG signal (Fitzgibbon et al., 2013). We subtracted the mean EEG voltage of the neighbor electrodes from each EEG signal. **Figure 4** shows the configuration of neighbor electrodes for 19 EEG channels. Each detector in NIRScout device records the signal from each separate source in two different wavelengths (760 and 850 nm). Oxy- and deoxyhemoglobin concentration changes (HbO and HbR) were computed using the modified Beer-Lambert law (Sassaroli and Fantini, 2004) using standard values for the chromophore extinction coefficients and differential path-length factor (Keles et al., 2016). fNIRS might be
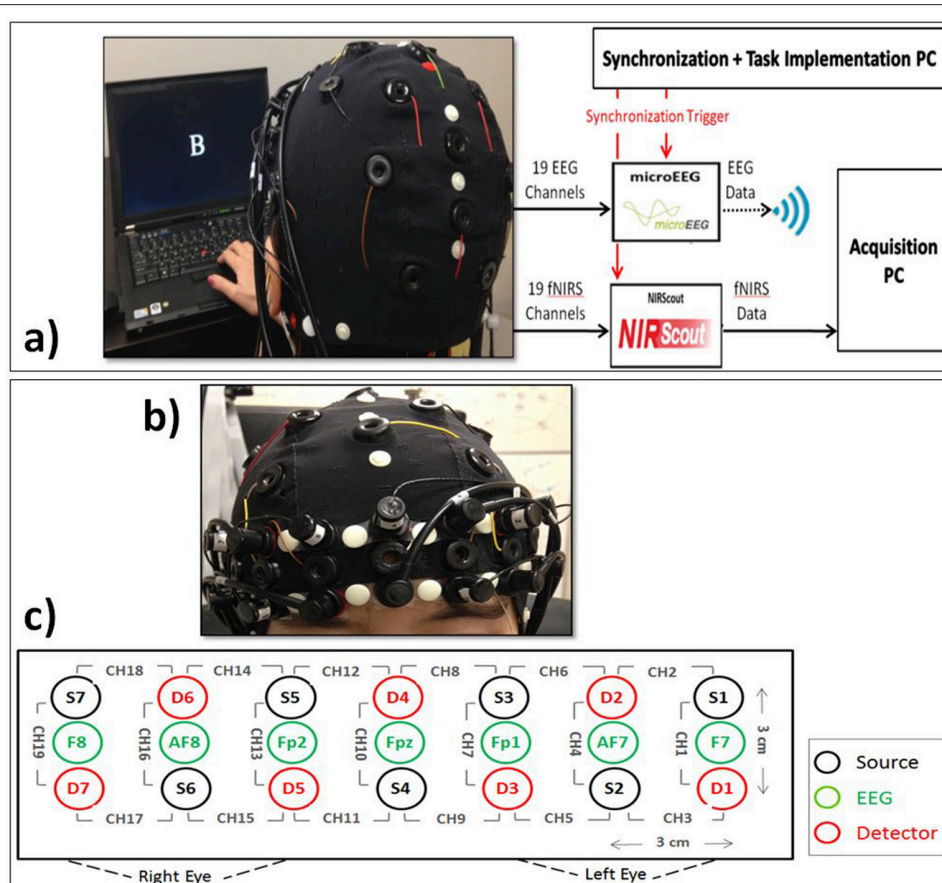
FIGURE 3 | (a) EEG+fNIRS recording setup. Subject interaction with the computer, synchronization of EEG and fNIRS signal, recording of EEG and fNIRS signals, and data transmission to the acquisition platform. **(b)** Coronal view of the subject showing the close view of the placement fNIRS optodes and EEG electrodes. **(c)** Topographical view of fNIRS sources (S_i, black) and detectors (D_i, red) and EEG electrodes (green). Each pair of source and detector separated by 3 cm creates a channel (CH_i). We used the signals from F7, Fpz, and F8.

contaminated with the movement, heart rate, and Mayer wave artifacts. In order to reduce these artifacts while retaining the maximum possible amount of information, a band pass filter of 0.01–0.5 Hz was applied to fNIRS signals. After the preprocessing step, two subjects were excluded from the rest of analysis due to the poor quality of the signal and excessive noise. The processed signals were inspected visually for the presence of muscle and motion, eye movements, and other artifacts. The recordings that were contaminated in excess of 10% by artifact were excluded as a whole (Keles et al., 2016). In addition, one subject was excluded since he was not sufficiently focused in the experiment according to 0-back low accuracy cut-off. **Figure 5** shows a segment of preprocessed data for one of the subjects. The figure indicates the temporal variations in the fNIRS signals and the EEG frequency bands, which are utilized in feature extraction. First and second rows are HbO and HbR of fNIRS channel 17, respectively. Third row is the EEG time-frequency map for channel O2.

After preprocessing, each task block ({0, 1, 2, 3}-back) and rest block was divided into 5, 10, 20, or 25 s epochs in order to assess the effect of window size on classification results. **Figure 6** shows four different epoch type with window size from 5 to 25 s.

In most of the cases there is an overlap between adjacent epochs (half size of epoch's length). This overlap was considered in order to capture the unique temporal response for each individual, as there could be inter-subject variability in the time required for the hemodynamic response to peak, and/or in the number of peaks (Power et al., 2012). In addition, during the classification phase, an imbalance in the number of features within each class biases the training procedure in favor of the class with a higher number of training features (He and Garcia, 2009). In our experiment design we have 40 rest blocks and 10 blocks from each n-back task type. From each task block, 16, 8, 4, and 2 features were extracted when we changed the size of the window from 5 to 25 s, respectively. From each rest block 5, 4, 2, and 1 features were extracted when we changed the size of the window from 5 to 25 s, respectively.

Feature Extraction

We extracted from each window three main categories of features for all 19 EEG electrodes and 19 fNIRS channels: EEG (uni-modal), fNIRS (uni-modal), and EEG+fNIRS (multi-modal or Hybrid).
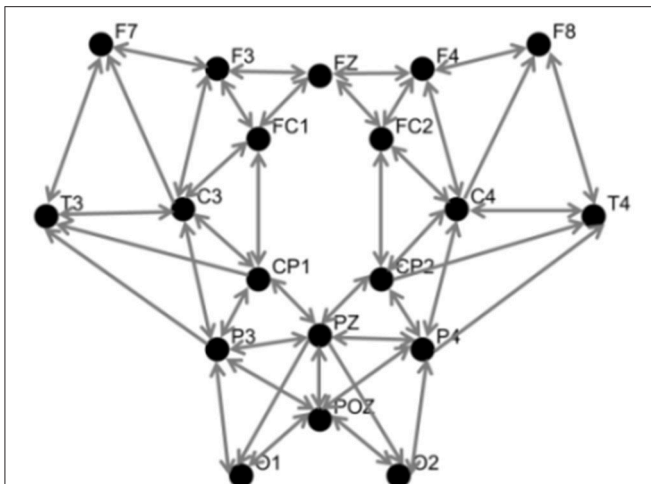
FIGURE 4 | Topographic view of EEG electrodes showing neighborhood pattern for Laplacian spatial filtering. Inward arrows to each node indicate the corresponding neighbors used for spatial filtering.

EEG-based features were computed from the frequency band power (PSD), phase locking value (PLV), phase-amplitude coupling (PAC), and the asymmetry of frequency band power between right and left hemispheres (Asym_PSD). Initially, the spectrogram was calculated using short-time Fourier transform method with windows of 1 s and half window size overlapping and frequency resolution of 1 Hz. The power was calculated in eight frequency bands each with a width of 4 Hz in the range 0 to 32 Hz. The ranges are referred to by their conventional labels: delta (0–4 Hz), theta (4–8), alpha (8–12), followed by five intervals ranging from low beta (12–16) to high beta (28–32). We also used the labels f1, f2,..., f8 for these frequency bands. EEG frequency band power for each epoch was extracted by integration of the corresponding power over each frequency band. We imposed the 32 Hz cutoff since higher frequencies in scalp EEG are generally not considered informative about cortical activity (Goncharova et al., 2003; Muthukumaraswamy, 2013). PLV is a measure of phase synchrony between two distinct neuronal populations, which is computed between two selected EEG electrodes as an estimate the inter-area synchrony (Vinck et al., 2011). PLV was estimated between electrode pairs that were selected to assess three different types of synchrony: intra-hemispheric (F3-P3, F4-P4, Fc1-Cp1, Fc2-Cp2, Fz-Poz), symmetric inter-hemispheric (F7-F8, F3-F4, Fc1-Fc2, C3-C4, T3-T4, Cp1-Cp2, P3-P4, O1-O2), and asymmetric inter-hemispheric (F3-P4, F4-P3, Fc1-Cp2, Fc2-Cp1). PLV was computed for four band of interest ([3–5], [9–11], [19–21], [39–41] Hz). PAC measures coupling between the phase of a low frequency (here [4–7], [9–13] Hz) oscillations and the amplitude of a high frequency ([15–35], [30–60] Hz) oscillation (Radwan et al., 2016). It provides an estimate of local, multi-frequency organization of neuronal activity (Dvorak and Fenton, 2014). We chose 8 EEG channel pairs between right and left hemispheres (F8-F7, F4-F3, Fc2-Fc1, T4-T3, C4-C3, Cp2-Cp1, P4-P3, and O2-O1) for Asym_PSD feature.

fNIRS features were based on HbO and HbR amplitude (HbO/R Amp.), slope of HbO and HbR (HbO/R slope), standard deviation of HbO and HbR (HbO/R Std.), skewness of HbO and HbR (HbO/R Skew.), and kurtosis of HbO and HbR (HbO/R Kurt.). The statistics of HbO and HbR are commonly used as features in fNIRS studies of MWL and BMIs (Naseer and Hong, 2015; Naseer et al., 2016a,b). Our inspection of the fNIRS data revealed patterns of correlation between HbO and HbR that were time and area dependent. Hence, we also included the zero-lagged correlation between HbO and HbR (HbO-HbR Corr.) as an additional feature. Hybrid features were based on EEG and fNIRS features in addition to specifically Hybrid quantities that depend simultaneously on both systems.

We chose to focus on a straightforward quantity, which can be easily calculated within the time windows of interest: the zero-lagged correlation between the Hb (HbO or HbR) amplitude and the EEG frequency band power (in eight separate bands described above). These neurovascular features based on HbO and HbR were denoted NVO (oxygenated neurovascular coupling) and NVR (deoxygenated neurovascular coupling), respectively. To calculate NVO/R for the left hemisphere, the correlation between each fNIRS channel (CH1 to CH9) and each frequency band of F7 EEG channel was calculated (band-passed filter within the specific frequency range). For the right hemisphere, the correlation between each fNIRS channel (CH11–CH19) and each frequency band of F8 EEG channel was calculated. For the fNIRS channel 10, which is located at the center we used the average of F7 and F8 channels to find NVO and NVR. This resulted in 152 (19 × 8) NVO and 152 NVR features from each window. Each set of features extracted from one subject's data were dc-shifted and scaled in order to have a mean value of zero and standard deviation of one.

Classification and Validation

Following feature extraction, we implemented SVM classification and k-fold cross-validation with $k = 10$. SVM is a non-parametric supervised classification method, which already showed promising results in the medical diagnostics, optical character recognition, electric load forecasting, and other fields. SVM can be a useful tool in the case of non-regularity in the data, for example when the data are not regularly distributed or have an unknown distribution (Auria and Moro, 2008). Linear SVM constructs an optimal hyperplane creating a decision surface maximizing the margin of separation between the closest data points belonging to different classes (Aghajani et al., 2013). The observations were randomly partitioned into k groups of approximately the same size. One group was selected as the testing and the rest as the training data. Principal component analysis (PCA) was applied to the training set. We applied PCA separately on each feature subgroup (11 subgroups, with the subgroups divided further by frequency bands as described in our feature extraction methods). For example, the EEG alpha frequency band power (8–12 Hz) consisted of the time series from 19 EEG channels. After PCA, these signals yielded 19 principal components (PC) and their associated time series as the new set of features. A similar PCA was applied to each feature subgroup. The resulting PCs contained a set of weights (for
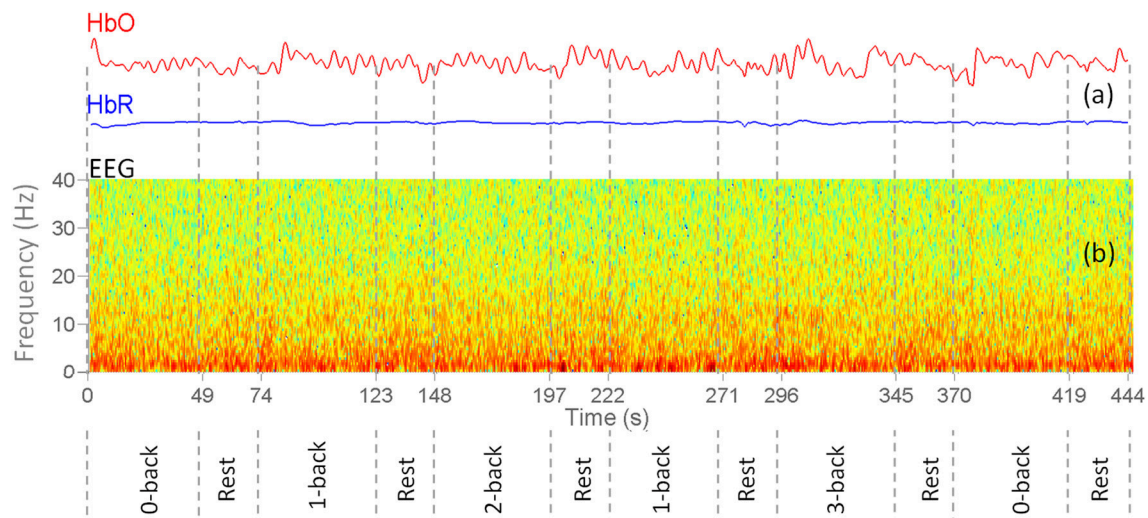
FIGURE 5 | Sample preprocessed EEG+fNIRS data for one of the subjects. Vertical dashes separate different *n*-back task and rest blocks. **(a)** Concentration changes of oxy-hemoglobin (red curve) and deoxy-hemoglobin (blue) for channel 17. **(b)** EEG Time-frequency map of the channel O2.
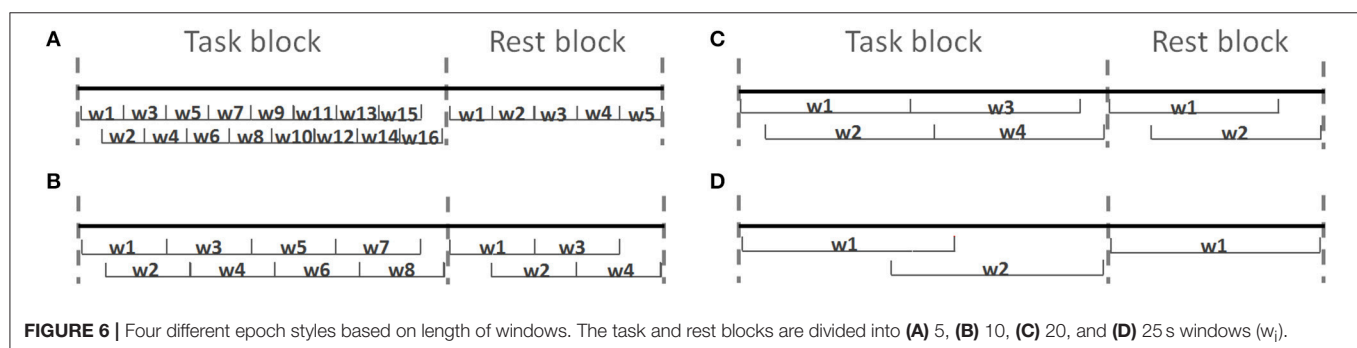
FIGURE 6 | Four different epoch styles based on length of windows. The task and rest blocks are divided into **(A)** 5, **(B)** 10, **(C)** 20, and **(D)** 25 s windows (w_i).

the EEG channels), which could be interpreted as an activation map. PCA therefore allowed us to interpret the topographic distribution of activation associated with every feature. PCA also yielded an eigenvalue corresponding to the variance of that feature. The eigenvalues typically decrease sharply, the sum of the first few accounting for almost all of the total energy of the 19 PCs. However, the most energetic PCs are not necessarily the most informative, as shown in Results (**Table 1**). In order to estimate the features' discriminating ability, we used the Pearson correlation coefficient method (Mwangi et al., 2014). A reference time series was constructed by labeling each window by a distinct integer that represented the rest or the task difficulty level ({0 (rest), 1 (0-back), 2 (1-back), 3 (2-back), 4 (3-back)}). We used R^2, the square of the Pearson correlation between the time series and the reference signal, to rank the set of features. The testing data were projected into the PC space that was obtained from the training data and the testing features were ranked by using the same method. In part of our analysis, we have chosen to reduce the number of features of the systems (EEG, fNIRS, and Hybrid) by truncating all systems at the same fixed size, eliminating the lowest ranked features. The labeled training examples were fed into a binary linear SVM. Each training

example contained a vector of feature values in a given window and its label that denoted one of the two classes of interest. The SVM constructed an optimal hyper-plane creating a decision surface maximizing the margin of separation between the closest data points belonging to different classes (Aghajani et al., 2013). In this study there were 10 possible pairs of binary classifications corresponding to our five distinct classes. In order to investigate the ability to discriminate WM loading against a baseline, we have chosen the pairs {1-back v rest}, {2-back v rest}, {3-back v rest}, {1-back v 0-back}, {2-back v 0-back}, and {3-back v 0-back} and performed binary classifications on them. We also investigated the ability to discriminate between degrees of MWL by using a multi-class scheme. For this purpose we utilized the error-correcting output code multiclass model (ECOC), which employs a set of binary classifiers. We adopted an all-pairs ECOC model to train a binary classifier on the pairs of classes in the training data and, for each new instance in the testing data, assigned the label that minimizes the aggregate Hamming loss from the predictions of all binary classifiers (Dietterich and Bakiri, 1995). In comparison to its alternatives, this approach has been shown to enhance accuracy while maintaining a low run-time complexity (Fürnkranz, 2002). We investigated four groups of multi-class

TABLE 1 | The top R^2 ranked features for three representative subjects for the binary rest v 3-back classification.

	Subject 1				Subject 2				Subject 3			
Rank	R^2	Descr.	Freq. (Hz)	PC or Chan. Pair	R^2	Descr.	Freq. (Hz)	PC or Chan. Pair	R^2	Descr.	Freq. (Hz)	PC or Chan. Pair
1	0.37	PSD	4–8	PC3	0.26	PSD	8–12	PC1	0.19	NVR	28–32	PC2
2	0.30	PSD	12–16	PC5	0.23	PSD	4–8	PC1	0.18	NVR	12–16	PC2
3	0.26	PSD	8–12	PC1	0.21	PSD	0–4	PC1	0.18	NVR	24–28	PC2
4	0.18	PLV	3–5	O1-O2	0.18	PSD	12–16	PC1	0.18	NVR	20–24	PC2
5	0.17	NVR	20–24	PC4	0.16	HbO	–	PC4	0.18	NVR	16–20	PC2
6	0.16	NVR	8–12	PC4	0.13	COR	–	PC2	0.18	HbR	–	PC2
7	0.16	NVR	16–20	PC4	0.13	HbR	–	PC1	0.17	NVR	4–8	PC2
8	0.16	NVR	12–16	PC4	0.12	NVR	24–28	PC1	0.16	PLV	9–11	Fz-Poz
9	0.16	NVR	24–28	PC4	0.11	NVR	12–16	PC1	0.15	PSD	12–16	PC3
10	0.15	NVR	28–32	PC4	0.11	NVR	16–20	PC1	0.14	PSD	8–12	PC4

Columns indicate the feature description (Descr.), corresponding frequency range if applicable, and PC index in order of descending energy (or the channel pair for PLV). The HbO-HbR corr. has been abbreviated as COR.

sets that contained narrow gradations of MWL: {3-back v 2-back v 1-back}, {3-backv 2-back v 1-back v 0-back}, {3-back v 2-back v 1-back v rest}, and {3-back v 2-back v 1-back v 0-back v rest}. The accuracy was computed as the fraction of labels in the testing data that were correctly identified by the SVM. Finally, the cross-validation was repeated k times with each group of observations being used exactly once as the testing data. The overall accuracy was calculated as the mean of the repetitions. In addition to overall accuracy, confusion matrices yield a very detailed overview of a classifier's performance. Usually, the confusion matrix is further summarized by some proportions extracted from the confusion matrix. The main metrics are (a) sensitivity of class A (Sens.$_A$) which describes how well the classifier recognize observations of class A, (b) specificity of class A (Spec.$_A$) which describes how well the classifier recognizes that an observation does not belong to class A, (c) positive predictive value of class A (PPV$_A$) tells us given the prediction is class A, what is the probability that the observation truly belongs to class A?, (d) negative predictive value of class A (NPV$_A$) tells us given a prediction does not belong to class A, what is the probability that the sample truly does not belong to class A (Beleites et al., 2013)? We pooled all the k confusion matrices of the k-fold cross validation to calculate Sens.$_A$, Spec.$_A$, PPV$_A$, and NPV$_A$. For all the calculations described in this paper we used Matlab v.8.6.0.267246 (R2015b) (The MathWorks, Inc., Natick, Massachusetts, United States).

RESULTS

We initially investigated the relationship between the subjects' performance and task difficulty, in order to insure that it was consistent with expectations. **Figure 7** shows the accuracy and response time of all subjects with error bars showing the standard deviation of inter-subject variability. The figure shows that the fraction of accurate responses decreased with increasing task difficulty. There was little or no accuracy decrement between 0- and 1-back tasks, as expected (Jonides et al., 1997).

FIGURE 7 | Behavioral performance of the subjects during task conditions of increasing difficulty, showing response accuracy (red) and response time (black). Error bars indicate the standard deviation of inter-subject variability. Asterisks indicate statistical significance derived from a two-way ANOVA comparison of each two response accuracy (red) or response time (black) (*p < 0.05, **p < 0.001, ***p < 0.0001).

Furthermore, the time it took subjects to produce a correct response increased (and eventually more than doubled) with task difficulty (**Figure 7**).

We next examined the HbO and HbR patterns during changes in mental load. **Figures 8 (a–e)** shows the grand block average of HbO (red) and HbR (blue) amplitude. The shaded area shows the standard deviation of the inter-subject variability. In this paper, the term grand block averaging denotes the average over the blocks of the same class and over all channels and subjects. Following neural activation, local blood flow and volume typically increase on a time scale of seconds and, at the beginning of the task, there is a localized rise in oxygenation in
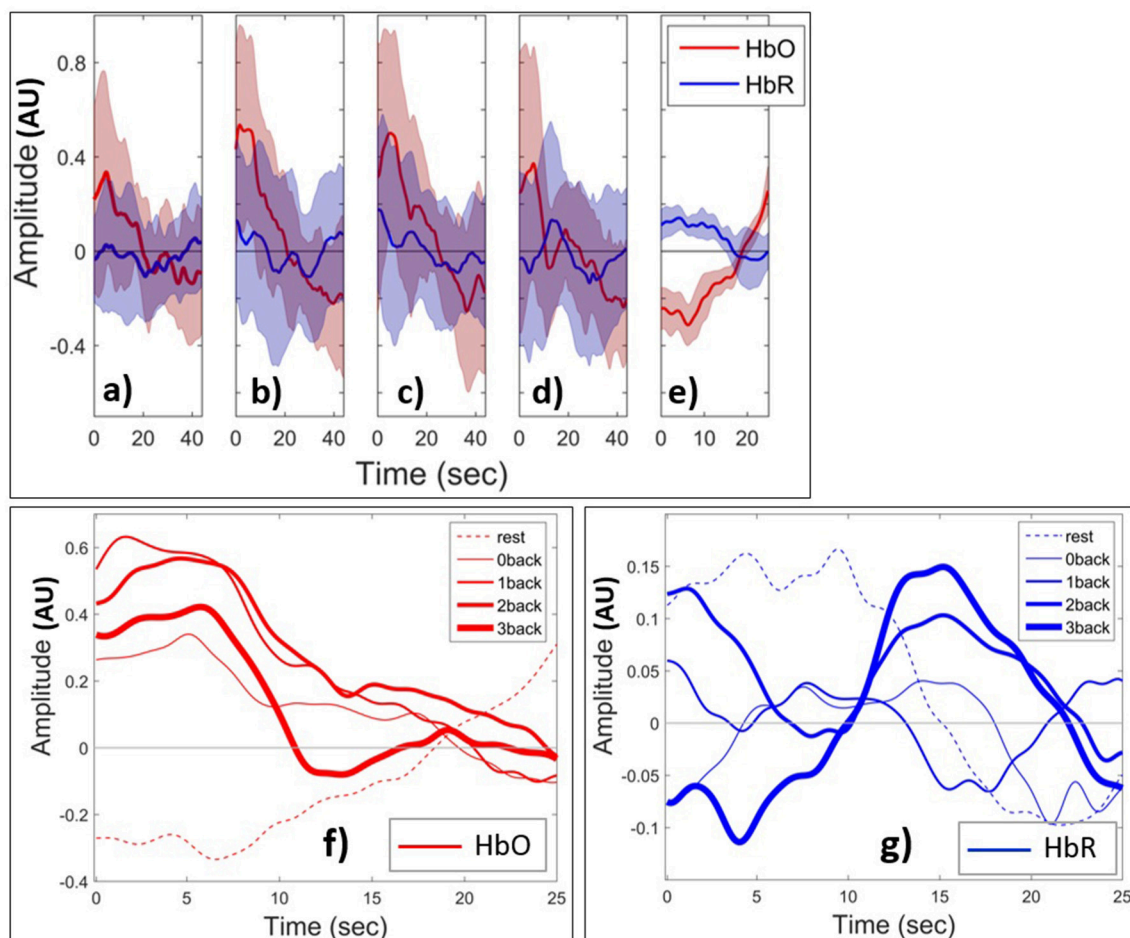
FIGURE 8 | Grand block average of normalized HbO (red) and HbR (blue) during **(a)** 0-back, **(b)** 1-back, **(c)** 2-back, **(d)** 3-back, **(e)** rest. The thick curves show the average over all channels and subjects. The shaded area indicates the standard deviation of inter-subject variability. Grand block average of HbO **(f)** and HbR **(g)** for rest (dashed curves) and task (solid). Increasing thickness of solid curves corresponds to increasing task difficulty from 0- to 3-back. AU means arbitrary units.

PFC (Huppert et al., 2006), which creates the positive peak of HbO. After a few seconds due to the metabolic consumption of oxygen the oxyhemoglobin concentration decreases leading to a negative HbO amplitude. During the rest state which comes after the task block, the oxyhemoglobin concentration starts to increase and HbO returns to baseline. Toward the end of rest window there is an apparent task anticipating rise in HbO. The range of changes of HbO is obviously higher than those of HbR during the task periods. From 0- to 2-back the positive peak of HbO increases and then decreases from 2- to 3-back. HbO and HbR have the opposite sign and are hence negatively short time correlated in the rest state. However, this appears to change during task in ways depending on the value of n. The range of HbO changes increases with n although they appear to slightly decrease as n changes from 2 to 3. In (**Figures 8f,g**) we show the grand block average of all tasks v rest state for one specific fNIRS channel (channel 10, which is located at Fpz, near the center of the forehead). In this figure the curves corresponding to rest and task with all values of n have been shown in one plot

to make the comparison easier. And we just show the first 25 s of the n-back task block (out of the full 44 s). The shaded areas for standard deviation are omitted for clarity. **Figure 8f** shows that the peak amplitude of HbO is positive for task performance and negative for rest. In addition, it decreases with increasing load for $n > 0$. The area under curve clearly discriminates between rest and task since it is negative during rest and positive for all other n-back tasks. By contrast, the peaks of the amplitude of grand block average of HbR (**Figure 8g**) that occur after the 10 s have a positive correlation with the level of mental load. These patterns are not observed in the case of the 0-back task since it is related to perception only and is less involved with WM system. We also examined the time course of selected features that were extracted from the signals.

Figure 9 shows the PSD extracted from EEG, HbO/R Amp. from fNIRS, and NVO/R features from EEG+fNIRS change in relation to the degree of WM load. We use the term session to denote a task block and the following rest block. For each cognitive state, we then have 10 sessions per subject. In the case of
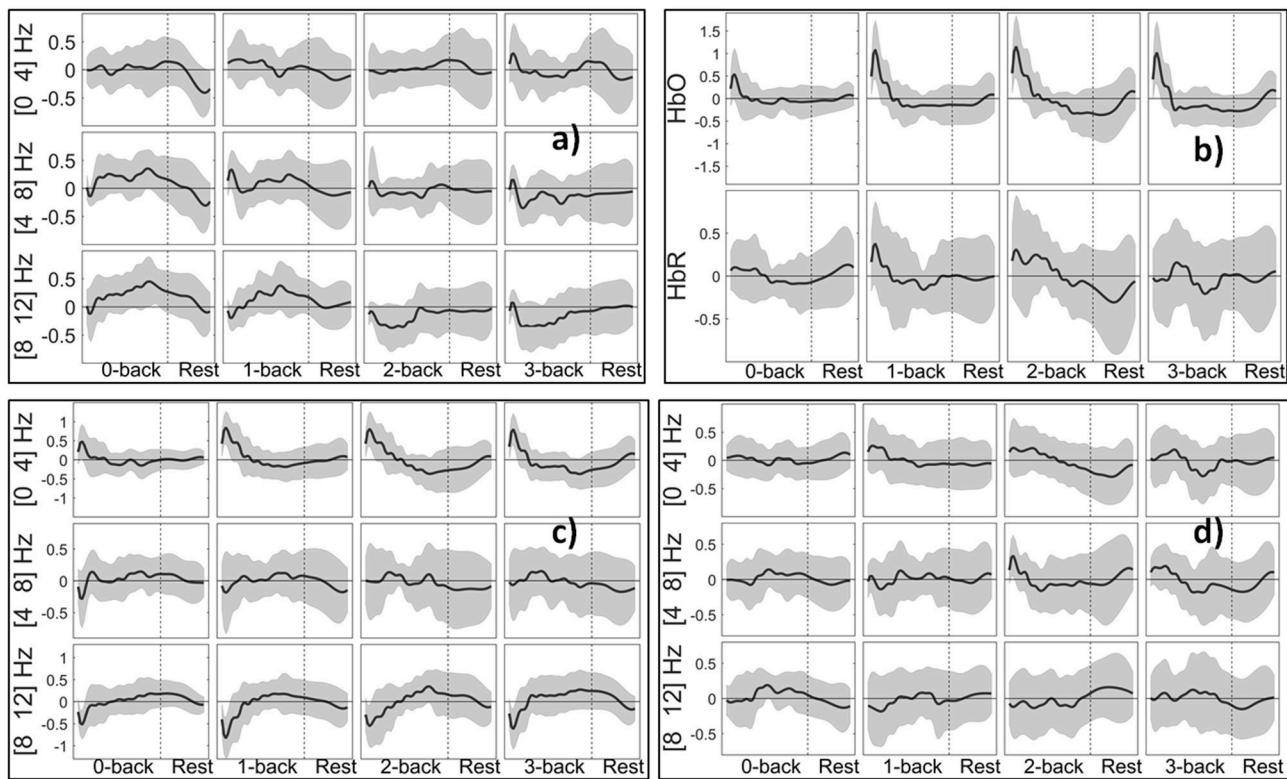
FIGURE 9 | Grand block average of normalized features from 5 s windows: **(a)** PSD (delta, theta, alpha bands) of channel O2. **(b)** HbO/R Amp. for channel 10. **(c)** NVO (delta, theta, alpha bands) for channel 10. **(d)** NVR features (delta, theta, alpha bands) for channel 10. Shaded areas indicate the standard deviation of inter-subject variability.

5 s windowing, for each feature, we have 21 values in each session (16 from task block and 5 from rest block). The curves in **Figure 9** were computed by first applying a simple triangular moving average filter covering three samples at each step, and then a cubic spline interpolation. The figure shows that the theta and alpha bands of EEG are positive during 0- and 1-back, although they become negative for 2- and 3-back tasks. The positive peak of HbO increases from 0- to 2-back and has a slightly lower peak for 3-back compared to 2-back. The figure also shows that the Hybrid features (such as NVO in the delta range) generally resemble the corresponding uni-modal features (such as HbO and PSD in the delta range) however they were dominated by neither, suggesting that the Hybrid feature contained additional information.

Table 1 shows the top 10 highest ranked features (based on R^2) for three subjects obtained during the 3-back v rest training set. The features are characterized by the description (e.g., PSD, PLV, HbO, as described in Section Methods) and the particular frequency band, where applicable. The frequency band is applicable only to the EEG and neurovascular features. The table also indicates the order of the feature according to the magnitude of its eigenvalue [ordered from the most energetic (1) to the least (19)]. Since PCA was not used in the case of PLV, the channel label is given instead of the PC order. For example, the highest ranked feature for subject one was the third most energetic PC from the EEG frequency band power in the theta range (4–8 Hz). For subject 3, the highest ranked feature was the

second most energetic PC from the neurovascular feature based on the correlation between HbR and the EEG frequency band power in the high beta range (28–32 Hz). The table illustrates that the types of features in the top ranked group may vary among subjects and that high discriminating ability of a feature does not imply high energy in the sense of the PCA.

Figure 10 shows the classification accuracies of various subsystems as well as the Hybrid system for the 3-back v rest using 5 s windows. The error bars represent the standard deviation of inter-subject variability. Within the EEG group (gray bars), the leftmost bar is the accuracy of a system based only the PSD features. On its immediate right is the accuracy of the subsystem based only on PLV features, and similarly for PAC and other feature types. The rightmost bar in the EEG group shows the accuracy of the full EEG system which includes all feature types based on EEG signals. Clearly the PSD is the primary contributor to the discriminating ability of the EEG, however, the accuracy appears to be slightly enhanced by including the other types of features. Among the fNIRS systems (red) the leftmost bar indicates that Hb amplitudes together with the HbO-HbR correlation is the primary contributor to the accuracy of detection. Unlike the EEG system, the other feature types such as slope and higher order statistics significantly enhance the accuracy of the fNIRS system. The overall accuracy of the fNIRS system is lower than the overall accuracy of the EEG system. The accuracy based only on the neurovascular features is indicated

by the leftmost bar in the Hybrid group (green). The middle bar in the Hybrid group represents the pooling of all features from the EEG and fNIRS systems. Finally the inclusion of the neurovascular features in the Hybrid system (rightmost green bar) appears to slightly enhance the accuracy.

Figure 10 compared the accuracies of various systems with each system containing the full set of features that belonged to it. The number of features in the full set was different for each system, for example the EEG, fNIRS, and Hybrid systems contained 360, 209, and 873 features, respectively. This may raise

FIGURE 10 | Accuracy of types of features in classifying rest v 3-back with 5 s feature windows. The error bars indicate the standard deviation of inter-subject variability. The union of neurovascular features is abbreviated as NV. Features are extracted from different systems: EEG (gray bars), fNIRS (red bars), and Hybrid (green bars).
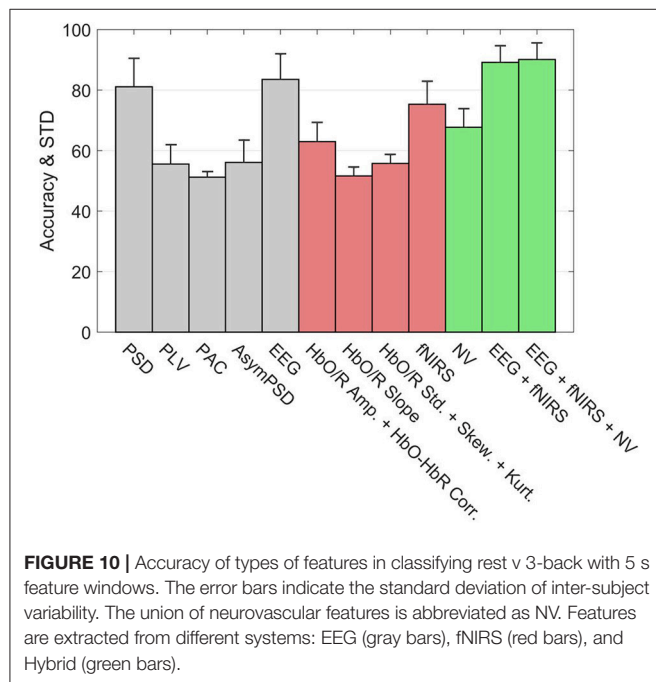
the question of to what extent these systems' accuracies were influenced by the number of features they contained, rather than by the information content per feature. In order to examine this topic, we computed the systems' accuracies after they had been truncated to contain the same number of features. This was done by selecting the top ranking group after the features had been sorted in order of descending values of R^2. The goal was to perform a comparison on an equal footing by truncating each system in the same way. The calculation was repeated by varying the number of features from two to the available maximum. **Figure 11a** indicates that the fNIRS system had the lowest accuracy over the entire range of the number of features. The EEG system performed better, while the Hybrid accuracy was consistently superior to either system, similar to the results shown in **Figure 10**. **Figure 11b** shows the cumulative sums of R^2 index v number of features for three systems which qualitatively agree with **Figure 11a**. The calculations are for the 3-back v rest using 5 s windows and they qualitatively agree with the results (not shown) of binary classifications of other pairs of classes and window sizes. The shaded areas indicate the standard deviation of inter-subject variability.

Figure 10 provided the results for only one type of binary classification (3-back v rest) and the variability over subjects as a standard deviation. However, it is highly instructive to examine the result for each subject as well as for every binary and multi-class problem that was described previously in our Methods. **Tables 2, 3** break down the accuracy of classification for each subject (S1, S2,..., S14), system type (EEG, fNIRS, Hybrid), and the type of classification problem. The mean as well as the minimum and maximum of the values for the subject population are provided as three separate columns on the left. The height of the rightmost bars within the EEG (gray), fNIRS (red), and Hybrid (green) groups in **Figure 10** correspond in **Table 2** to the accuracy percentages 83.5, 75.3, and 90.1 shown under the column "Mean" and the row "3-back v rest." In the columns for

FIGURE 11 | (a) Accuracy and **(b)** cumulative sum of R^2 for EEG (black), fNIRS (red), and Hybrid (green) systems as a function of system size. Mean and standard deviation over subjects are indicated by the solid curves and shaded areas, respectively. The classification task was rest v 3-back and feature window size was 5 s.

TABLE 2 | Binary classification accuracy for all subjects included in the study (S1 to S14) for 10-fold cross validation.

		Min	Max	Mean	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
1back v rest	EEG	65.9	92.5	78.9	81.3 ± 14	89.1 ± 13	73.1 ± 17	65.9 ± 13	84.4 ± 13	92.5 ± 4	90.6 ± 6	67.2 ± 15	74.1 ± 17	72.5 ± 17	91.3 ± 9	75.0 ± 14	80.0 ± 19	68.1 ± 13
	fNIRS	67.5	89.4	75.3	71.9 ± 13	70.0 ± 14	68.8 ± 11	69.7 ± 13	69.7 ± 11	82.2 ± 14	86.9 ± 5	80.6 ± 8	67.5 ± 14	69.1 ± 15	89.4 ± 10	83.1 ± 7	70.6 ± 14	75.0 ± 12
	Hybrid	76.5	97.2	87.2	89.1 ± 12	79.7 ± 18	76.6 ± 15	86.9 ± 12	76.6 ± 17	95.6 ± 4	90.6 ± 14	86.9 ± 9	86.9 ± 10	77.8 ± 16	97.2 ± 4	91.9 ± 8	93.8 ± 4	90.9 ± 8
2back v rest	EEG	69.4	95.3	82.9	90.6 ± 9	87.5 ± 9	86.3 ± 9	75.3 ± 10	87.5 ± 8	92.5 ± 11	83.4 ± 9	69.4 ± 15	72.2 ± 12	75.6 ± 16	95.3 ± 4	79.7 ± 12	92.2 ± 6	73.1 ± 15
	fNIRS	70.0	84.7	77.6	79.1 ± 15	70.0 ± 15	72.2 ± 11	71.6 ± 15	80.3 ± 13	84.7 ± 10	79.4 ± 7	78.1 ± 9	73.8 ± 16	78.8 ± 10	84.7 ± 11	74.1 ± 14	82.8 ± 5	77.2 ± 15
	Hybrid	71.2	97.2	89.3	91.6 ± 8	71.3 ± 17	92.5 ± 5	87.2 ± 12	92.2 ± 5	97.2 ± 3	87.2 ± 9	82.8 ± 12	84.7 ± 12	92.5 ± 5	93.4 ± 8	90.9 ± 4	92.2 ± 9	94.4 ± 5
3back v rest	EEG	67.2	96.6	83.5	90.9 ± 6	78.8 ± 17	82.2 ± 18	82.8 ± 12	82.2 ± 10	96.6 ± 6	87.2 ± 14	67.2 ± 12	88.8 ± 14	74.1 ± 15	94.4 ± 10	80.3 ± 24	89.7 ± 9	74.4 ± 19
	fNIRS	63.7	88.4	75.3	67.5 ± 16	68.4 ± 9	65.6 ± 13	63.7 ± 14	83.8 ± 5	82.5 ± 11	77.2 ± 11	78.1 ± 11	75.0 ± 13	82.2 ± 11	88.4 ± 12	78.8 ± 11	72.5 ± 11	70.6 ± 14
	Hybrid	82.2	96.6	90.1	95.6 ± 2	84.7 ± 17	89.4 ± 14	83.8 ± 17	92.8 ± 5	95.6 ± 5	82.5 ± 15	82.2 ± 14	93.4 ± 7	84.7 ± 15	96.6 ± 4	92.2 ± 4	94.7 ± 4	93.4 ± 6
1back v 0back	EEG	80.0	91.9	86.7	89.1 ± 5	91.9 ± 4	80.0 ± 8	87.5 ± 5	89.7 ± 4	84.7 ± 5	87.2 ± 6	83.4 ± 6	86.6 ± 6	81.6 ± 8	85.3 ± 5	90.3 ± 7	88.1 ± 5	88.8 ± 5
	fNIRS	67.8	82.8	74.0	68.8 ± 10	74.1 ± 12	73.4 ± 9	69.7 ± 10	70.9 ± 7	76.6 ± 7	74.1 ± 11	82.8 ± 6	70.0 ± 9	73.4 ± 11	77.8 ± 8	79.7 ± 5	67.8 ± 12	77.2 ± 8
	Hybrid	88.1	93.4	91.4	90.9 ± 5	91.6 ± 6	89.1 ± 5	91.3 ± 5	93.1 ± 4	92.2 ± 4	92.5 ± 4	93.4 ± 4	88.1 ± 7	88.1 ± 6	91.9 ± 5	93.1 ± 6	93.1 ± 6	91.9 ± 3
2back v 0back	EEG	85.0	94.7	91.2	89.4 ± 5	89.7 ± 5	94.7 ± 3	87.8 ± 5	93.1 ± 5	94.7 ± 5	92.2 ± 5	89.1 ± 5	91.9 ± 4	85.0 ± 5	93.8 ± 5	92.2 ± 4	93.8 ± 5	89.1 ± 3
	fNIRS	66.9	83.4	74.8	66.9 ± 8	72.8 ± 7	71.3 ± 5	69.7 ± 5	78.1 ± 11	77.2 ± 6	73.8 ± 8	75.3 ± 9	79.4 ± 6	72.8 ± 7	79.7 ± 5	83.4 ± 7	70.3 ± 10	77.2 ± 8
	Hybrid	91.6	96.6	93.8	91.9 ± 6	91.6 ± 4	93.8 ± 4	91.9 ± 6	94.7 ± 3	95.3 ± 5	92.8 ± 6	94.4 ± 4	93.8 ± 4	92.5 ± 4	96.6 ± 3	95.6 ± 4	94.7 ± 4	93.4 ± 5
3back v 0back	EEG	86.2	96.2	92.0	91.9 ± 6	86.3 ± 5	90.9 ± 4	90.3 ± 4	92.5 ± 3	95.3 ± 5	94.1 ± 6	88.4 ± 8	93.4 ± 4	86.3 ± 7	96.3 ± 6	94.4 ± 4	94.4 ± 4	93.1 ± 4
	fNIRS	66.6	77.5	71.6	66.9 ± 4	67.2 ± 8	77.5 ± 5	67.5 ± 8	74.1 ± 6	74.7 ± 7	71.9 ± 10	70.3 ± 6	74.7 ± 8	75.9 ± 6	73.8 ± 9	72.5 ± 9	68.8 ± 8	66.6 ± 9
	Hybrid	90.6	96.6	93.8	93.4 ± 5	90.6 ± 5	96.3 ± 2	92.2 ± 4	94.7 ± 3	95.9 ± 3	92.8 ± 5	91.3 ± 5	93.4 ± 5	91.9 ± 3	93.8 ± 5	95.0 ± 3	94.7 ± 4	96.6 ± 3

For each one of the subjects, ANOVA showed a significant difference (p < 0.0001) for comparison of EEG v Hybrid results and fNIRS v Hybrid results.
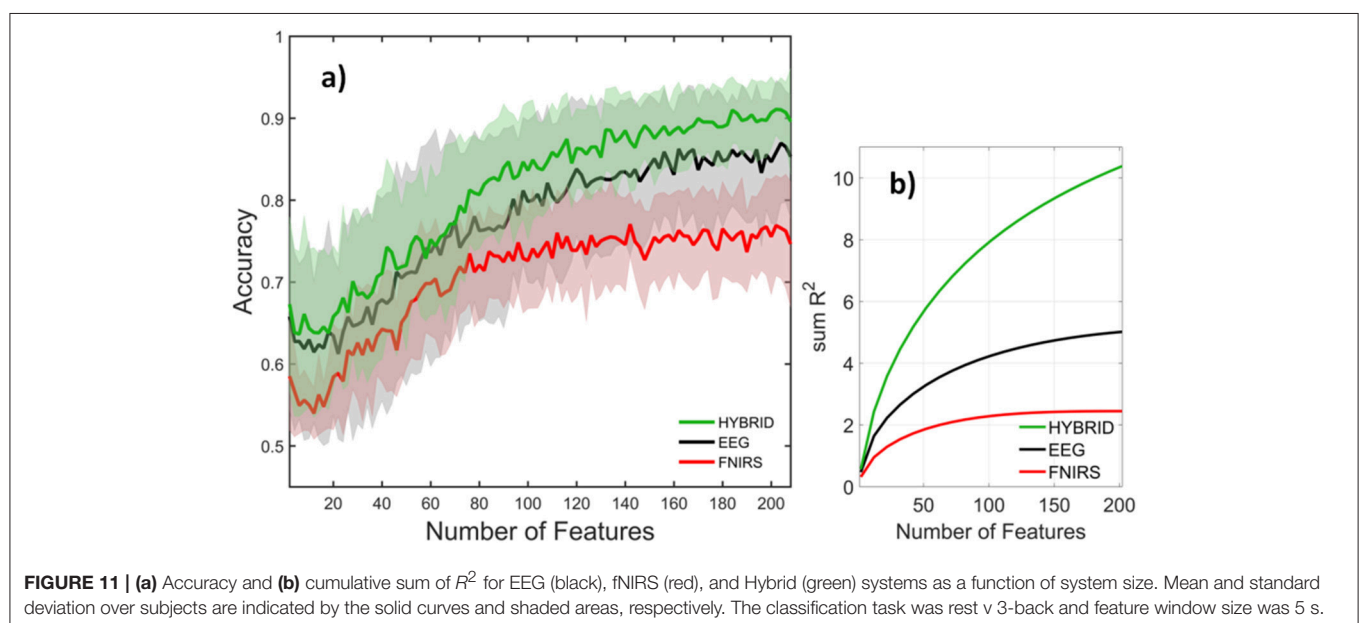
TABLE 3 | Multi-class classification accuracy for all subjects included in the study (S1 to S14) for 10-fold cross validation.

		Min	Max	Mean	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14
3back v 2back v 1back	EEG	72.9	85.4	80.8	76.5 ± 7	84.2 ± 5	74.2 ± 8	76.7 ± 7	82.3 ± 6	83.8 ± 6	85.4 ± 4	85.2 ± 4	83.5 ± 4	72.9 ± 6	83.5 ± 6	81.3 ± 7	82.5 ± 5	80.0 ± 8
	fNIRS	49.6	62.5	57.9	49.6 ± 8	61.3 ± 8	54.0 ± 10	56.3 ± 9	60.0 ± 6	62.5 ± 8	54.4 ± 8	58.8 ± 8	56.9 ± 6	55.0 ± 7	62.1 ± 8	60.2 ± 8	61.3 ± 10	59.0 ± 8
	Hybrid	82.5	90.2	87.3	82.5 ± 3	90.2 ± 4	85.4 ± 3	85.2 ± 5	90.0 ± 5	88.1 ± 3	88.7 ± 5	88.5 ± 4	85.4 ± 4	84.6 ± 5	90.2 ± 5	86.9 ± 5	87.3 ± 4	89.8 ± 5
3back v 2back v 1back v Rest	EEG	73.1	87.7	80.6	80.6 ± 5	85.6 ± 3	77.3 ± 6	76.7 ± 5	80.5 ± 5	87.7 ± 4	85.5 ± 5	74.5 ± 6	82.2 ± 4	73.1 ± 5	86.3 ± 4	77.5 ± 4	84.7 ± 4	76.3 ± 4
	fNIRS	55.8	69.5	62.3	58.4 ± 4	61.7 ± 3	57.5 ± 6	61.1 ± 6	64.7 ± 6	66.7 ± 5	55.8 ± 4	62.5 ± 4	61.3 ± 8	58.3 ± 4	69.5 ± 5	65.3 ± 7	63.7 ± 6	65.2 ± 4
	Hybrid	82.7	91.7	87.1	85.2 ± 5	88.4 ± 2	86.3 ± 3	85.0 ± 4	87.7 ± 4	89.1 ± 3	86.9 ± 4	82.7 ± 4	88.1 ± 2	84.1 ± 4	91.7 ± 3	86.6 ± 5	88.3 ± 4	89.7 ± 2
3back v 2back v 1back v 0back	EEG	70.3	83.9	79.2	77.5 ± 6	80.8 ± 4	72.3 ± 5	77.3 ± 6	79.7 ± 5	82.8 ± 2	83.4 ± 5	77.8 ± 3	80.2 ± 5	70.3 ± 5	82.5 ± 3	82.0 ± 6	83.9 ± 4	78.8 ± 4
	fNIRS	42.5	56.2	51.7	42.5 ± 5	51.2 ± 5	49.2 ± 7	48.4 ± 8	56.3 ± 6	53.3 ± 5	48.0 ± 6	55.8 ± 7	50.5 ± 10	49.8 ± 6	56.3 ± 7	55.5 ± 5	55.0 ± 7	52.3 ± 5
	Hybrid	81.4	88.9	84.9	81.7 ± 5	85.3 ± 4	82.8 ± 4	81.4 ± 4	87.5 ± 3	85.0 ± 5	83.9 ± 3	84.7 ± 2	84.8 ± 4	81.6 ± 7	88.9 ± 4	87.3 ± 4	86.1 ± 4	87.5 ± 3
3back v 2back v 1back v 0back v rest	EEG	69.6	85.0	78.0	79.0 ± 5	82.4 ± 5	73.0 ± 4	72.7 ± 3	77.4 ± 5	84.5 ± 2	82.3 ± 3	72.5 ± 4	79.9 ± 2	69.6 ± 5	85.0 ± 4	78.6 ± 3	83.8 ± 4	72.0 ± 4
	fNIRS	46.7	62.6	56.2	46.8 ± 7	55.8 ± 5	54.6 ± 7	53.0 ± 5	59.0 ± 5	62.6 ± 3	54.0 ± 3	57.4 ± 4	54.6 ± 6	52.5 ± 6	61.9 ± 3	59.9 ± 5	57.1 ± 5	58.4 ± 5
	Hybrid	81.1	92.1	85.4	85.5 ± 6	84.3 ± 3	83.0 ± 3	82.5 ± 3	86.0 ± 3	87.9 ± 3	85.1 ± 3	81.1 ± 2	84.2 ± 4	81.7 ± 5	92.1 ± 3	87.1 ± 3	86.5 ± 4	88.3 ± 4

For each one of the subjects, ANOVA showed a significant difference (p < 0.0001) for comparison of EEG v Hybrid results and fNIRS v Hybrid results.

TABLE 4 | Sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), and negative predictive value (NPV) are listed in percentage (%) for all classification cases (binary and multi-class) and all systems (EEG, fNIRS, Hybrid).

		Sens.			Spec.			PPV			NPV		
		EEG	fNIRS	Hybrid	EEG	fNIRS	Hybrid	EEG	fNIRS	Hybrid	EEG	fNIRS	Hybrid
Rest v 1back	Rest	94.5	94.5	96.7	63.3	56.7	79.6	72.1	68.6	82.6	92.0	91.1	96.1
	1back	63.3	56.7	79.6	94.5	94.5	96.7	92.0	91.1	96.1	72.1	68.6	82.6
Rest v 2back	Rest	94.0	95.2	95.8	71.8	56.7	82.5	76.9	68.7	84.6	92.3	92.2	95.1
	2back	71.8	56.7	82.5	94.0	95.2	95.8	92.3	92.2	95.1	76.9	68.7	84.6
Rest v 3back	Rest	94.4	94.2	96.6	72.7	56.7	83.3	77.5	68.5	85.2	92.8	90.8	96.1
	3back	72.7	56.7	83.3	94.4	94.2	96.6	92.8	90.8	96.1	77.5	68.5	85.2
0back v 1back	0back	86.7	69.2	90.9	86.7	72.9	92.4	86.7	71.9	92.3	86.7	70.3	91.0
	1back	86.7	72.9	92.4	86.7	69.2	90.9	86.7	70.3	91.0	86.7	71.9	92.3
0back v 2back	0back	90.8	67.5	92.5	91.5	76.7	94.3	91.4	74.3	94.2	90.9	70.3	92.6
	2back	91.5	76.7	94.3	90.8	67.5	92.5	90.9	70.3	92.6	91.4	74.3	94.2
0back v 3back	0back	90.7	68.8	93.0	93.3	71.2	94.0	93.1	70.5	93.9	90.9	69.6	93.1
	3back	93.3	71.2	94.0	90.7	68.8	93.0	90.9	69.6	93.1	93.1	70.5	93.9
1back v 2back v 3back	1back	76.0	51.1	82.5	94.4	79.3	95.5	87.1	55.3	90.1	88.7	76.4	91.6
	2back	83.8	60.4	89.6	87.7	78.3	90.9	77.3	58.1	83.1	91.6	79.8	94.6
	3back	82.7	57.8	87.1	89.2	77.1	93.2	79.3	55.7	86.5	91.2	78.5	93.5
0back v 1back v 2back v 3back	0back	76.4	45.0	81.4	94.6	85.4	96.0	82.6	50.7	87.1	92.3	82.3	93.9
	1back	79.9	48.1	86.7	91.7	82.6	93.9	76.2	48.0	82.5	93.2	82.7	95.5
	2back	81.0	53.4	87.0	92.3	81.5	94.2	77.9	49.1	83.4	93.6	84.0	95.6
	3back	79.7	50.7	83.8	93.7	82.8	95.5	80.8	49.5	86.2	93.3	83.4	94.6
Rest v 1back v 2back v 3back	Rest	81.4	79.6	86.4	95.4	91.5	97.3	85.5	75.8	91.4	93.9	93.1	95.6
	1back	79.3	51.6	86.6	93.9	85.1	95.6	81.3	53.5	86.8	93.2	84.0	95.5
	2back	80.8	56.0	87.0	92.3	85.8	94.7	77.8	56.9	84.5	93.5	85.4	95.6
	3back	80.8	54.7	86.4	92.5	84.9	94.6	78.3	54.7	84.2	93.5	84.9	95.4
Rest v 0back v 1back v 2back v 3back	Rest	78.6	77.9	84.5	95.3	93.4	97.5	80.8	74.8	89.3	94.7	94.4	96.2
	0back	79.4	47.3	85.4	94.2	87.5	95.9	77.5	48.7	84.0	94.8	86.9	96.3
	1back	78.0	44.8	84.4	93.7	86.9	95.3	75.7	46.0	81.8	94.5	86.3	96.1
	2back	75.9	49.9	84.3	94.4	87.7	96.1	77.3	50.4	84.6	94.0	87.5	96.1
	3back	78.3	48.8	85.2	94.8	86.6	96.1	79.1	47.8	84.5	94.6	87.1	96.3

individual subjects, **Table 2** shows the mean accuracy and the standard deviation from the trials in the 10-fold cross validation.

Table 2 suggests that the mean accuracy of classifying task against a baseline increases with n, as expected. The accuracy of detecting 0-back v n-back appears to be slightly greater than that of detecting rest v n-back ($n > 0$). For example, 87.2% for 1-back v rest and 91.4% for 1-back v 0-back. **Table 3** shows the results for multi-class classification. In this case, the accuracy tends to decline slightly as more classes are included in the classification problem. In all subjects and classification problems, the Hybrid system has the greatest accuracy without exception. We investigated whether the observed superiority of the Hybrid system was statistically significant. A two-way ANOVA was performed on every classification problem (a row of **Table 2** or

Table 3) by using as the two factors the type of system (EEG, fNIRS, Hybrid) and the subject. The analysis was repeated by taking the classification problems as, first, the binary types in **Table 2** and, second, the multi-class types in **Table 3**. In all cases, the differences of accuracy among the subjects were not significant and there were no interactions between system type and subject, while the differences in accuracy between the Hybrid and the uni-modal system was significant with a $p < 0.001$.

Table 4 lists the sensitivity (Sens.), specificity (Spec.), positive predictive value (PPV), and negative predictive value (NPV) for each individual class within a classification case. For example for the case of {Rest v 3back}, each one of rest and 3-back classes would have a Sens., Spec., PPV, and NPV, respectively. In addition, this table summarizes all these metrics for EEG,

fNIRS, and Hybrid systems in order to make it easier to compare between their capabilities.

The foregoing results corresponded to 5 s windows but qualitatively agreed with patterns we observed with other window sizes as well. We also assessed the effect of window length on classification accuracy for EEG, fNIRS, and Hybrid systems. **Figure 12** shows the results of this assessment. We examined four different lengths for the windows (5, 10, 20, and 25 s). Change of window length has the same effect on all three types of systems. By increasing the length from 5 to 20 the accuracy increases and declines thereafter.

DISCUSSION AND CONCLUSION

The functional activity of the human brain can be observed with various imaging techniques including fMRI, fNIRS, and EEG. Each of these modalities has its advantages and disadvantages. The advantage of using Hybrid EEG+fNIRS system can be divided into two main categories: First, each of these modalities is measuring the changes in a specific brain physiology. EEG results directly from the electrical activity of cortical and subcortical neurons with a sub-millisecond temporal resolution. On the other hand, fNIRS yields local measures of changes in HbO and HbR concentration and is, therefore, an indicator of metabolic/hemodynamic changes associated with neural activity. Second, the physics of measurement behind EEG and fNIRS are quite different. This property, for example, makes EEG signal prone to blink and muscle artifacts, while this is not the case for fNIRS. Hence using a multimodal recording system we are able to assess brain behavior from different physiological perspectives in addition to compensating for some weaknesses of one modality by the other one.

Our results suggest that EEG+fNIRS combined with a classifier are capable of robustly discriminating among various levels of MWL. In our study, the Hybrid system had an accuracy higher than either EEG or fNIRS alone for every subject. The pooling of EEG and fNIRS features and the inclusion of neurovascular features resulted in a synergistic enhancement, rather than in a diluting effect (which would have given a performance intermediate between the two modalities). In mission-critical contexts such as aviation or surgery, even small improvements in MWL detection can translate into significant gains in safety and efficiency. Our experiments were designed to use WM load (adjusted through the value of n in the n-back task) as a correlate of MWL in general. Furthermore, EEG and fNIRS can be integrated without excessive cost, effort, or intrusiveness for the user. The combination of all these considerations suggests that EEG+fNIRS should be preferred to only EEG or fNIRS, in developing passive BCIs and other applications which need to monitor users' MWL.

Our preliminary analysis of the experimental data was consistent with expectations. For example **Figure 7** indicated that the fraction of accurate responses declined more steeply as n increased. This can be explained by noting that temporal tagging is the cognitive process that imposes the greatest load in the n-back task, as compared to the other processes which are also involved such as encoding, storage, matching, and inhibition to dampen the oldest memory traces (Jonides et al., 1997).

Temporal tagging, unused in the 0- and 1-back, begins to affect MWL substantially as $n > 1$. Another interesting preliminary result was the observation that HbO showed an anticipatory increase near the end of the rest sessions (**Figure 8E**). This is consistent with related fMRI findings (Sakai and Passingham, 2003) and with the fact that the PFC is involved in planning future action. A negative correlation between HbO and HbR has been seen in the literature. Based on (**Figures 8A–E**), the shaded area, which is the standard deviation of normalized HbO and HbR variations within the block for all of the subjects and all of the fNIRS channels, is relatively high. This shows the high level of inter-subjects variability and that might be the reason that we are not seeing such anti-correlated pattern between HbO and HbR in **Figure 8**. Izzetoglu et al. (2004) showed that the reason behind the drop of the peak of HbO in **Figure 8d** is that when a participant reaches his maximum performance capacity or in another word starts to overloads cognitively, he loses his concentration on the task and as a result performance as well as the oxygenated hemoglobin changes decline.

Figure 5 did not show any differences between rest and task states that were obvious to visual inspection of the preprocessed EEG or fNIRS signals. Subject and block averaging of various features shown in **Figure 9** did, however, indicate that such systematic variations existed. Lower values of the EEG alpha band power in the 2- and 3-back task, and higher values of HbO in the beginning of the task period in the n-back (increasing with n) relative to those in the rest state were examples of such visible variations. To take advantage of such variations we employed discrimination through linear SVM. In the case of non-linear SVM, the kernel can help with the non-linearly separable data and map it into a new feature space in which the dataset are separable with a linear SVM. In non-linear SVM in order to improve the accuracy choosing the optimum kernel parameters, is necessary. This can reduce classifier's generalization potential for new subjects when we don't want to adjust the kernel's parameters. It will increase the probability of overfitting occurrence, since increasing the complexity level of a classifier gives it the flexibility to match exactly to the train set. In addition, the option of choosing a non-linear SVM depends on the exact application to make a trade-off between a slightly higher accuracy rather than calculation speed. The same trade-off we face to choose the windows size. The results shown in **Figure 10**, **Tables 2, 3** were highly promising for accurately discriminating among the rest and task states. As **Tables 2, 3** show, the subject averaged accuracy of the Hybrid system in binary discrimination was lowest (87.2%) for 1-back v rest and highest (96.6%) for 3-back v 0-back. The corresponding lowest and highest results for uni-modal systems were fNIRS (71.6%) and EEG (92.0%), both for 3-back v rest. We calculated the overall average of accuracy one time for all of the binary cases and one time for all of the muli-class cases. We did this calculation for EEG, fNIRS, and Hybrid systems separately. The results show that EEG, fNIRS, and Hybrid system, in the case of binary classification, have 85.9, 74.8, and 90.9% overall accuracy, respectively. EEG, fNIRS, and
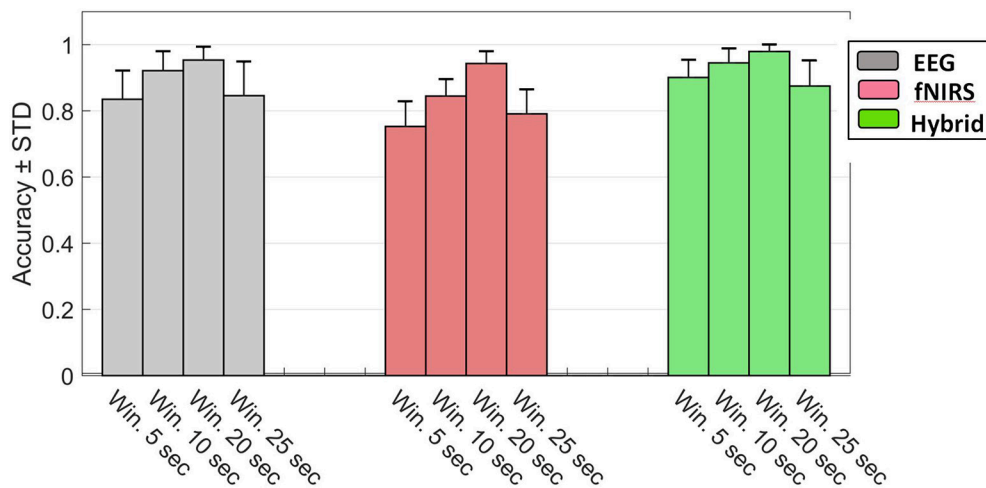
FIGURE 12 | Accuracy of the rest v 3-back classification as a function of window size for EEG (gray), fNIRS (red), and Hybrid (green) systems. Error bars indicate the standard deviation of inter-subject variability.

Hybrid system, in the case of multi-class classification, have 79.6, 57.0, 86.2% overall accuracy, respectively. These numbers convey that the accuracy of each one of EEG, fNIRS, and Hybrid systems are higher for the binary cases. The multi-class accuracies were generally lower; however, note that the chance level accuracy for multi-class classification is less than binary classification (33% for 3-back v 2-back v 1-back, 25% for 3-back v 2-back. v 1-back v rest and also 3-back v 2-back. v 1-back v 0-back, and 20% for 3-back v 2-back. v 1-back v 0-back v rest).

Table 4 reveals that, for all of the four extracted metrics from the confusion matrix (sensitivity, specificity, PPV, NPV), always Hybrid system has a higher value than EEG system and EEG system has a higher value than fNIRS system. In the cases of binary classification, for those that we are detecting between task and rest, the sensitivity of detection of rest state is significantly higher than the sensitivity of detection of task state. As its complement, the specificity of detecting task state is higher than detecting the rest state. In the cases of binary classification, for those that we are detecting between task and task, the sensitivity of detecting the task with a higher difficulty level is more than those with less level of difficulty, although the difference is not very significant. The PPV and NPV are usually more important than sensitivity and specificity. Patients and doctors want to know whether this particular patient is ill rather than whether the test can recognize ill people (Beleites et al., 2013). Here, our result (**Table 4**) shows that Hybrid system has at the same time a very promising PPV and NPV for all of the classification cases. Except for the case of {1-back v rest}, the minimum of sensitivity, specificity, PPV, NPV for the Hybrid system are 81.4, 82.5, 81.8, and 84.6%, respectively.

Selecting an optimal subset from the full set of features is crucial for achieving high accuracy and avoiding over-fitting. In some applications, e.g., those involving on-board real-time analysis, it may also be important to keep the system size small and avoid computational delays. **Figure 11b** shows the

cumulative sum of R^2 v number of features for three systems which qualitatively agrees with **Figure 11a**, suggesting that R^2 ranking is an effective method of feature selection. We have not used an explicit artifact rejection step in our analysis. However, it is well known that PCA can segregate non-cerebral artifacts (typically of higher amplitude than contributions of cortical origin) into distinct PCs. Our feature selection based on R^2 then assigns a lower rank to such PCs and they are excluded from a truncated system.

One of the main considerations in developing an online system is computational speed. It is instructive to review the computational loads of particular feature types in conjunction with how effectively they discriminate among rest and task states. For example, **Figure 10** shows that PAC is the least discriminating EEG feature. This may be important in designing a compact and efficient detector, as PAC is also the most computationally time-consuming feature. By contrast, the most effective EEG feature (PSD) was also the fastest to compute. In our study, the central processing unit (CPU) time required for computing PSD, PLV, PAC, and Asym_PSD were, respectively, 0.1, 14.3, 44.4, and 0.2 s. The CPU times required for other features were as follows: HbO/R Amp. and HbO-HbR Corr. (0.1 s), HbO/R slope (14.3 s), Std., Skew. and Kurt. collectively (3.3 s), and NVO and NVR (3.4 s).

Our results suggest that Hybrid outperforms the uni-modal systems for each subject (**Tables 2, 3**), every classification problem (**Tables 2, 3**), every number of features (**Figure 11**), and every window size (**Figure 12**). This could have been due to the neurovascular features that the uni-modal systems do not contain. NV obviously had a higher classification performance rather than any of fNIRS based feature subgroups. However, **Figure 10** indicates that such features contribute little if any to the accuracy (two rightmost bars) after the other EEG and fNIRS features have been pooled. The likely explanation instead is related to inter-subject variability. We have found that the

top ranked (in terms of R^2) features tend to differ among different subjects. Although, the EEG frequency band power (especially in the alpha range) tended to play an important role for most subjects, for other subjects other feature types, fNIRS- or Hybrid based, dominated the top ranks. An example of this is provided in **Table 1** where the third subject's most discriminating PCs were neurovascular. EEG and fNIRS use different physical processes for detection and the underlying physiology which they detect are different. Hence deficiencies such as artifacts, weak sensor coupling, or subject variability leading to a weak signal would selective affect only one modality rather than both. The Hybrid advantage may be associated primarily with the complementary nature of the individual modalities.

Figure 12 indicates that accuracy can be increased by using larger windows. But this presents a tradeoff between accuracy and rapid detection. The windows with the highest accuracy were 20 s long and may be impractical for some applications if the online response to rapid changes in MWL is desirable. As window size increases, although the amount of information per window likely increases, the number of windows available for training the classifier decreases. Fewer training data are expected to cause the classifiers to underperform (Grimes et al., 2008). The decline in accuracy in **Figure 12** for windows >20 s may be due to the excessively small number of data available for training.

A handful of studies on concurrent EEG and fNIRS and WM tasks have been previously published. Hirshfield et al. (2009) combined an 8 channel fNIRS covering the forehead with 32 channel whole-head EEG with $N = 4$ subjects as they performed a counting and mental arithmetic task with adjustable difficulty. They used separate classifiers for the fNIRS and EEG (k-nearest neighbor and Naive Bayes classification, respectively) and obtained a maximum accuracies of 64% (with fNIRS) and 82% (EEG). They did not attempt to use the multi-modal data concurrently. The generally higher accuracy of the EEG is consistent with our results. Their overall lower accuracies relative to our results may be due to the relatively short 2 s feature extraction window. Liu Y. et al. (2013) used a 16-optode fNIRS system covering the forehead and 28 EEG sensors at the standard 10–20 sites, with $N = 16$ subjects performing a n-back task. They found significant correlations between WM load and some EEG frequency band powers as well as HbO and HbR, however did not attempt classification. Their study focused on discovering neural correlates of the effects of practice time on performance. Coffey et al. (2012) recorded three fNIRS channels over the left forehead together with 8 EEG electrodes placed mainly in the frontal and central areas, from $N = 10$ subjects in a n-back task. They extracted EEG frequency band power and fNIRS Hb amplitude features from 5 s windows and employed them in linear discriminant analysis classifiers. They report maximum accuracies of 89.6% (EEG), 79.7% (fNIRS), and 91.0% (Hybrid). Their results differed from ours in that in some subjects all their systems had very low accuracies and their Hybrid accuracies were not always higher than those of both uni-modal systems. However, the fact that EEG generally had the higher uni-modal accuracy and that Hybrid could attain

the highest observed accuracy were consistent with our findings. The differences from our results could be attributed to the relatively lower number of sensors and fewer types of features they employed.

Acquiring the very low-frequency (VLF) oscillations (<0.5 Hz) in the EEG signal requires highly specialized amplifiers (DC-coupled, high input impedance, high DC stability, and a wide dynamic range; Demanuele et al., 2007). In addition, VLF oscillations are known to be linked with specific pathologies such as epileptic seizures or attention deficit hyperactivity disorder (Steriade et al., 1993; Vanhatalo et al., 2004; Demanuele et al., 2007) that are not within the range of interest in this study. On the other hand, some studies (Gevins et al., 1997; Berka et al., 2007; So et al., 2017) demonstrated EEG within gamma range as a biomarker for discrimination between different cognitive states. We defined the band-pass filter cutoff frequency (0.5–80 Hz) based on these criteria. Although in the feature extraction section, we did not consider gamma frequency range features and have considered this as the future work.

The present study had several limitations which we have not directly addressed due to constraints of available time or effort. Firstly, the group of subjects included only one female. This may have been due to the demographics of the subjects, who happened to be interested in volunteering for our study. In addition to the neural correlates of MWL, we have recorded the subjects' performance characteristics. However, it may prove insightful to collect data on the MWL by using additional techniques such as self-reporting, which was not done in this study. In some studies for assessment of MWL participants filled out the NASA Task Load Index (NASA TLX) questionnaire (Hart and Staveland, 1988) to provide a subjective evaluation of the mental demand induced by different levels of task difficulty. In this study, we implicitly used the assumption that an increase in the level of task difficulty will result in a higher MWL. This can be also considered in future studies. In addition, it is possible that during the course of an experiment the subjects' performance and MWL change through training effects. Studying the performance and neural correlates of MWL for subsets of our data could reveal differences in the beginning and at the end of the study. This would also require an additional investigation of statistical validity, and was not attempted. The statistical significance of the results of our study was demonstrated through a two-way ANOVA that showed significant differences in the accuracy of the Hybrid v uni-modal systems. However, we have not investigated whether a smaller group of subjects would still yield a significant result. We have investigated the capabilities of various subsets of the types of features that were available. It would also be illuminating to investigate the classification accuracy of subsets of the full array of our sensors. Such information can help design more compact headsets and is the subject of an ongoing study. The headset we used is lightweight and no discomfort was reported by any of the subjects. However, wearing it may nevertheless affect performance, and this could be revealed in a parallel set of experiments which we have not done. The primary goal of our study was to apply machine learning techniques in discriminating levels of MWL. We used multiple statistical techniques to ensure

that the statistical significance of the values of accuracy that we obtained for such discrimination. Our observations regarding the range of changes of Hb are therefore only qualitative and observational, serving to ensure that our results are consistent with expectations.

In this study, we have taken steps toward investigating the EEG+fNIRS feature extraction and analysis methods by using a popular WM task. We anticipate and hope that converging efforts in Hybrid hardware integration (Safaie et al., 2013) and data analysis (Biessmann et al., 2011; Keles et al., 2016), potentially based on detailed knowledge of underlying physiology (Bari et al., 2012; Mandrick et al., 2016a), will lead to more effective passive BMIs and other applications in neuroergonomics.

AUTHOR CONTRIBUTIONS

HA and AO designed the study. HA collected and analyzed the data. HA, AO, and MG wrote the manuscript.

FUNDING

REFERENCES

Aghajani, H., and Omurtag, A. (2016). "Assessment of mental workload by EEG + fNIRS," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Orlando, FL), 3773–3776.

Aghajani, H., Zahedi, E., Jalili, M., Keikhosravi, A., and Vahdat, B. V. (2013). Diagnosis of early Alzheimer's disease based on EEG source localization and a standardized realistic head model. *IEEE J. Biomed. Health Inform.* 17, 1039–1045. doi: 10.1109/JBHI.2013.2253326

Auria, L., and Moro, R. A. (2008). *Support Vector Machines (SVM) as a Technique for Solvency Analysis*. Rochester, NY: Social Science Research Network.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Bari, V., Calcagnile, P., Molteni, E., Re, R., Contini, D., Spinelli, L., et al. (2012). From neurovascular coupling to neurovascular cascade: a study on neural, autonomic and vascular transients in attention. *Physiol. Meas.* 33, 1379–1397. doi: 10.1088/0967-3334/33/8/1379

Beleites, C., Salzer, R., and Sergo, V. (2013). Validation of soft classification models using partial class memberships: an extended concept of sensitivity & co. applied to grading of astrocytoma tissues. *Chemom. Intell. Lab. Syst.* 122, 12–22. doi: 10.1016/j.chemolab.2012.12.003

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78(5 Suppl.), B231–B244.

Biessmann, F., Plis, S., Meinecke, F. C., Eichele, T., and Müller, K.-R. (2011). Analysis of multimodal neuroimaging data. *IEEE Rev. Biomed. Eng.* 4, 26–58. doi: 10.1109/RBME.2011.2170675

Boff, K. R., Kaufman, L., and Thomas, J. P. (1994). *Handbook of Perception and Human Performance, Vol. 2, Cognitive Processes and Performance.* New York, NY: John Wiley & Sons.

Brouwer, A.-M., Hogervorst, M. A., van Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the N-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Buccino, A. P., Keles, H. O., and Omurtag, A. (2016). Hybrid EEG-fNIRS asynchronous brain-computer interface for multiple motor tasks. *PLoS ONE* 11:e0146610. doi: 10.1371/journal.pone.0146610

Byrne, A. (2013). Mental workload as a key factor in clinical decision making. *Adv. Health Sci. Educ. Theory Pract.* 18, 537–545. doi: 10.1007/s10459-012-9360-5

Carswell, C. M., Clarke, D., and Seales, W. B. (2005). Assessing mental workload during laparoscopic surgery. *Surg. Innov.* 12, 80–90. doi: 10.1177/155335060501200112

Causse, M., and Matton, N. (2014). "Using near infrared spectroscopy to detect mental overload in flight simulator," in *Advances in Cognitive Engineering and Neuroergonomics,* eds M. K. Stanney, and S. K. Hale. Available online at: http://oatao.univ-toulouse.fr/16199/

Coffey, E. B. J., Brouwer, A.-M., and van Erp, J. B. F. (2012). Measuring workload using a combination of electroencephalography and near infrared spectroscopy. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 56, 1822–1826. doi: 10.1177/1071181312561367

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., et al. (1997). Temporal dynamics of brain activation during a working memory task. *Nature* 386, 604–608. doi: 10.1038/386604a0

Dehais, F., Causse, M., Vachon, F., Régis, N., Menant, E., and Tremblay, S. (2014). Failure to detect critical auditory alerts in the cockpit: evidence for inattentional deafness. *Hum. Factors* 56, 631–644. doi: 10.1177/0018720813510735

Demanuele, C., James, C. J., and Sonuga-Barke, E. J. (2007). Distinguishing low frequency oscillations within the 1/f spectral behaviour of electromagnetic brain signals. *Behav. Brain Funct.* 3, 62. doi: 10.1186/1744-9081-3-62

Dietterich, T. G., and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *J Artif. Int. Res.* 2, 263–286.

Dixon, B. J., Daly, M. J., Chan, H., Vescan, A. D., Witterick, I. J., and Irish, J. C. (2013). Surgeons blinded by enhanced navigation: the effect of augmented reality on attention. *Surg Endosc.* 27, 454–61. doi: 10.1007/s00464-012-2457-3

Dornhege, G., Millán, J. R., Hinterberger, T., McFarland, D., and Müller, K. (2007). "Improving human performance in a real operating environment through real-time mental workload detection," in *Toward Brain-Computer Interfacing* (MIT Press), 409–422. Available online at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6281209

Durantin, G., Gagnon, J.-F., Tremblay, S., and Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behav. Brain Res.* 259, 16–23. doi: 10.1016/j.bbr.2013.10.042

Dvorak, D., and Fenton, A. A. (2014). Towards a proper estimation of phase-amplitude coupling in neural oscillations. *J. Neurosci. Methods* 225, 42–56. doi: 10.1016/j.jneumeth.2014.01.002

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., et al. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage* 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Fishburn, F. A., Norr, M. E., Medvedev, A. V., and Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Front. Hum. Neurosci.* 8:76. doi: 10.3389/fnhum.2014.00076

Fitzgibbon, S. P., Lewis, T. W., Powers, D. M. W., Whitham, E. W., Willoughby, J. O., and Pope, K. J. (2013). Surface laplacian of central scalp electrical signals is insensitive to muscle contamination. *IEEE Trans. Biomed. Eng.* 60, 4–9. doi: 10.1109/TBME.2012.2195662

Fürnkranz, J. (2002). Round Robin classification. *J. Mach. Learn. Res.* 2, 721–747. doi: 10.1162/153244302320884605

Gagnon, J.-F., Durantin, G., Vachon, F., Causse, M., Tremblay, S., and Dehais, F. (2012). "Anticipating human error before it happens: towards a psychophysiological model for online prediction of mental workload," in *Human Factors: A View from an Integrative Perspective.* 57–66. Available online at: https://goo.gl/Ma6HbT

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., and Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PLoS ONE* 10:e0121279. doi: 10.1371/journal.pone.0121279

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Hum. Factors* 40, 79–91. doi: 10.1518/001872098779480578

Gevins, A., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-Resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cereb. Cortex* 7, 374–385. doi: 10.1093/cercor/7.4.374

Goncharova, I. I., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2003). EMG contamination of EEG: spectral and topographical characteristics. *Clin. Neurophysiol.* 114, 1580–1593. doi: 10.1016/S1388-2457(03)00093-2

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., and Rao, R. P. N. (2008). "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)* (New York, NY: ACM), 835–844.

Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., and Esfahani, E. T. (2015). Understanding cognitive performance during robot-assisted surgery. *Urology* 86, 751–757. doi: 10.1016/j.urology.2015.07.028

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in *Advances in Psychology, Vol. 52, Human Mental Workload*, eds P. A. Hancock and N. Meshkati, (Elsevier), 139–183. Available online at: http://www.sciencedirect.com/science/article/pii/S0166411508623869

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935

Hirshfield, L. M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E. T., Jacob, R. J. K., et al. (2009). "Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload," in *Foundations of Augmented Cognition Neuroergonomics and Operational Neuroscience,* eds D. D. Schmorrow, I. V. Estabrooke, M. Grootjen (Berlin; Heidelberg: Springer), 239–247. Available online at: http://link.springer.com/chapter/10.1007/978-3-642-02812-0_28

Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* 8:322. doi: 10.3389/fnins.2014.00322

Horrey, W. J., and Wickens, C. D. (2003). *Multiple Resource Modeling of Task Interference in Vehicle Control, Hazard Awareness and in-Vehicle Task Performance.* Available online at: https://trid.trb.org/view.aspx?id=698780

Hoshi, Y., Tsou, B. H., Billock, V. A., Tanosaki, M., Iguchi, Y., Shimada, M., et al. (2003). Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks. *Neuroimage* 20, 1493–1504. doi: 10.1016/S1053-8119(03)00412-9

Huppert, T. J., Hoge, R. D., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2006). A temporal comparison of BOLD, ASL, and NIRs hemodynamic responses to motor stimuli in adult humans. *Neuroimage* 29, 368–382. doi: 10.1016/j.neuroimage.2005.08.065

Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., and Pourrezaei, K. (2003). "fNIR spectroscopy as a measure of cognitive task load," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Vol.4 *(IEEE Cat No03CH9)* (Cancun), 3431–3434.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., and Chance, B. (2004). Functional optical brain imaging using near-infrared during cognitive tasks. *Int. J. Hum. Comp. Inter.* 17, 211–227. doi: 10.1207/s15327590ijhc1702_6

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., et al. (1997). Verbal working memory load affects regional brain activation as measured by PET. *J. Cogn. Neurosci.* 9, 462–475. doi: 10.1162/jocn.1997.9.4.462

Kaber, D. B., Onal, E., and Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator 'situation awareness, and subjective workload. *Hum. Factors Ergon. Manuf. Serv. Ind.* 10, 409–430. doi: 10.1002/hfm.20168

Keles, H. O., Barbour, R. L., Aghajani, H., and Omurtag, A. (2014b). "Multimodality mapping approach for evolving functional brain connectivity patterns: a fNIRs-EEG study," in *Biomedical Optics (Optical Society of America)* BT5B.2. OSA Technical Digest. Available online at: http://www.opticsinfobase.org/abstract.cfm?URI=BIOMED-2014-BT5B.2

Keles, H. O., Barbour, R. L., Aghajani, H., and Omurtag, A. (2014a). *Investigation of the Neurovascular Coupling from Simultaneous fNIRS-EEG System Using the Triplet Holder.* Montreal, QC. Available online at: https://goo.gl/UVmtwT

Keles, H. O., Barbour, R. L., and Omurtag, A. (2016). Hemodynamic correlates of spontaneous neural activity measured by human whole-head resting state EEG+fNIRs. *Neuroimage* 138, 76–87. doi: 10.1016/j.neuroimage.2016.05.058

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Liu, T., Xiao, T., and Shi, J.-N. (2013). Neural correlates of conflict control on facial expressions with a flanker paradigm. *PLoS ONE* 8:e69683. doi: 10.1371/journal.pone.0069683

Liu, Y., Ayaz, H., Curtin, A., Onaral, B., and Shewokis, P. A. (2013). "Towards a hybrid P300-based BCI using simultaneous fNIR and EEG," in *Foundations of Augmented Cognition* (Berlin; Heidelberg: Springer), 335–344. Available online at: http://link.springer.com/chapter/10.1007/978-3-642-39454-6_35

Longo, L. (2016). "Mental workload in medicine: foundations, applications, open problems, challenges and future perspectives," in *IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)* (Dublin), 106–111.

Mandrick, K., Chua, Z., Causse, M., Perrey, S., and Dehais, F. (2016a). Why a comprehensive understanding of mental workload through the measurement of neurovascular coupling is a key issue for neuroergonomics? *Front. Hum. Neurosci.* 10:250. doi: 10.3389/fnhum.2016.00250

Mandrick, K., Peysakhovich, V., Rémy, F., Lepron, E., and Causse, M. (2016b). Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biol. Psychol.* 121(Pt. A), 62–73. doi: 10.1016/j.biopsycho.2016.10.002

McBride, D. K., and Schmorrow, D. (2005). *Quantifying Human Information Processing.* Lanham, MD: Lexington Books.

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Front. Hum. Neurosci.* 7:138. doi: 10.3389/fnhum.2013.00138

Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics* 12, 229–244. doi: 10.1007/s12021-013-9204-3

Naseer, N., and Hong, K.-S. (2013). Classification of functional near-infrared spectroscopy signals corresponding to the right- and left-wrist motor imagery for development of a brain-computer interface. *Neurosci. Lett.* 553, 84–89. doi: 10.1016/j.neulet.2013.08.021

Naseer, N., and Hong, K.-S. (2015). fNIRs-based brain-computer interfaces: a review. *Front. Hum. Neurosci.* 9:3. doi: 10.3389/fnhum.2015.00003

Naseer, N., Hong, M. J., and Hong, K.-S. (2014). Online binary decision decoding using functional near-infrared spectroscopy for the development of brain-computer interface. *Exp. Brain Res.* 232, 555–564. doi: 10.1007/s00221-013-3764-1

Naseer, N., Noori, F. M., Qureshi, N. K., and Hong, K.-S. (2016a). Determining optimal feature-combination for LDA classification of functional near-infrared spectroscopy signals in brain-computer interface application. *Front. Hum. Neurosci.* 10:237. doi: 10.3389/fnhum.2016.00237

Naseer, N., Qureshi, N. K., Noori, F. M., and Hong, K.-S. (2016b). Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface. *Comput. Intell. Neurosci.* 2016:e5480760. doi: 10.1155/2016/5480760

Navon, D., and Gopher, D. (1979). On the economy of the human-processing system. *Psychol Rev.* 86, 214–255. doi: 10.1037/0033-295X.86.3.214

Omurtag, A., Aghajani, H., and Keles, H. O. (2017). *Classifying the Brain's Functional Status in Verbal Fluency Task: EEG+fNIRS.* Available online at: https://goo.gl/A4ttiQ

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59. doi: 10.1002/hbm.20131

Palmer, E. D., and Kobus, D. A. (2007). "The future of augmented cognition systems in education and training," in *Foundations of Augmented Cognition*, eds D. D. Schmorrow, L. M. Reeves (Berlin; Heidelberg:

Springer), 373–379. Lecture Notes in Computer Science. Available online at: http://link.springer.com/chapter/10.1007/978-3-540-73216-7_42

Parasuraman, R., and Wilson, G. F. (2008). Putting the brain to work: neuroergonomics past, present, and future. *Hum. Factors* 50, 468–474. doi: 10.1518/001872008X288349

Power, S. D., Kushki, A., and Chau, T. (2012). Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by NIRS. *PLoS ONE* 7:e37791. doi: 10.1371/journal.pone.0037791

Prabhu, A., Smith, W., Yurko, Y., Acker, C., and Stefanidis, D. (2010). Increased stress levels may explain the incomplete transfer of simulator-acquired skill to the operating room. *Surgery* 147, 640–645. doi: 10.1016/j.surg.2010.01.007

Putze, F., Hesslinger, S., Tse, C.-Y., Huang, Y., Herff, C., Guan, C., et al. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Front. Neurosci.* 8:373. doi: 10.3389/fnins.2014.00373

Radwan, B., Dvorak, D., and Fenton, A. A. (2016). Impaired cognitive discrimination and discoordination of coupled theta–gamma oscillations in Fmr1 knockout mice. *Neurobiol. Dis.* 88, 125–138. doi: 10.1016/j.nbd.2016.01.003

Safaie, J., Grebe, R., Moghaddam, H. A., and Wallois, F. (2013). Toward a fully integrated wireless wearable EEG-NIRs bimodal acquisition system. *J. Neural Eng.* 10:056001. doi: 10.1088/1741-2560/10/5/056001

Sahayadhas, A., Sundaraj, K., and Murugappan, M. (2012). Detecting driver drowsiness based on sensors: a review. *Sensors* 12, 16937–16953. doi: 10.3390/s121216937

Sakai, K., and Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nat. Neurosci.* 6, 75–81. doi: 10.1038/nn987

Sassaroli, A., and Fantini, S. (2004). Comment on the modified beer-lambert law for scattering media. *Phys. Med. Biol.* 49, N255–N257. doi: 10.1088/0031-9155/49/14/N07

Sato, H., Yahata, N., Funane, T., Takizawa, R., Katura, T., Atsumori, H., et al. (2013). A NIRS-fMRI investigation of prefrontal cortex activity during a working memory task. *Neuroimage* 83, 158–173. doi: 10.1016/j.neuroimage.2013.06.043

Smith, E. E., and Jonides, J. (1997). Working memory: a view from neuroimaging. *Cognit. Psychol.* 33, 5–42. doi: 10.1006/cogp.1997.0658

Smith, M. E., and Gevins, A. (2005a). "Assessing fitness-for-duty and predicting performance with cognitive neurophysiological measures," in *Proc. SPIE 5797, Biomonitoring for Physiological and Cognitive Performance during Military Operations* (Orlando, FL), 127–138. doi: 10.1117/12.602185

Smith, M. E., and Gevins, A. (2005b). "Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator", in *Proc. SPIE5797, Biomonitoring for Physiological and Cognitive Performance during Military Operations* (Orlando, FL), 116–126. doi: 10.1117/12.602181

So, W. K. Y., Wong, S. W. H., Mak, J. N., and Chan, R. H. M. (2017). An evaluation of mental workload with frontal EEG. *PLoS ONE* 12:e0174949. doi: 10.1371/journal.pone.0174949

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., and Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 267–278. doi: 10.1007/s11858-015-0754-8

Stanney, K. M., and Hale, K. S. (2012). *Advances in Cognitive Engineering and Neuroergonomics, 2nd Edn.* Boca, FL: CRC Press, Inc.

Stefanidis, D., Haluck, R., Pham, T., Dunne, J. B., Reinke, T., Markley, S., et al. (2007). Construct and face validity and task workload for laparoscopic camera navigation: virtual reality versus videotrainer systems at the sages learning center. *Surg. Endosc.* 21, 1158–1164. doi: 10.1007/s00464-006-9112-9

Steriade, M., Nuñez, A., and Amzica, F. (1993). A novel slow (< 1 Hz) oscillation of neocortical neurons *in vivo*: depolarizing and hyperpolarizing components. *J. Neurosci.* 13, 3252–3265.

Strangman, G., Culver, J. P., Thompson, J. H., and Boas, D. A. (2002). A quantitative comparison of simultaneous bold fmri and nirs recordings during functional brain activation. *Neuroimage* 17, 719–731. doi: 10.1006/nimg.2002.1227

Strayer, D. L., and Drews, F. A. (2007). Cell-phone–induced driver distraction. *Curr. Dir. Psychol. Sci.* 16, 128–131. doi: 10.1111/j.1467-8721.2007.00489.x

Strayer, D. L., and Johnston, W. A. (2001). Driven to distraction: dual-task studies of simulated driving and conversing on a cellular telephone. *Psychol. Sci.* 12, 462–466. doi: 10.1111/1467-9280.00386

Teplan, M. (2002). Fundamental of EEG measurement. *Meas. Sci. Rev.* 2, 1–11.

Thomas, L. C., and Wickens, C. D. (2001). Visual displays and cognitive tunneling: frames of reference effects on spatial judgments and change detection. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 45, 336–340. doi: 10.1177/154193120104500415

Tomita, Y., Vialatte, F.-B., Dreyfus, G., Mitsukura, Y., Bakardjian, H., and Cichocki, A. (2014). Bimodal BCI using simultaneously NIRS and EEG. *IEEE Trans. Biomed. Eng.* 61, 1274–1284. doi: 10.1109/TBME.2014.2300492

Vanhatalo, S., Palva, J. M., Holmes, M. D., Miller, J. W., Voipio, J., and Kaila, K. (2004). Infraslow oscillations modulate excitability and interictal epileptic activity in the human cortex during sleep. *Proc. Natl. Acad. Sci. U.S.A.* 101, 5053–5057. doi: 10.1073/pnas.0305375101

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., and Pennartz, C. M. A. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage* 55, 1548–1565. doi: 10.1016/j.neuroimage.2011.01.055

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* 3, 159–177. doi: 10.1080/14639220210123806

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Factors* 50, 449–455. doi: 10.1518/001872008X288394

Wolpaw, J., and Wolpaw, E. W. (2012). *Brain-Computer Interfaces: Principles and Practice, 1st Edn.* Oxford ; New York, NY: Oxford University Press.

Xiao, Y., Wang, Z., Wang, M., and Lan, Y. (2005). [The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index]. *Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi.* 23, 178–181.

Xie, B., and Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work Stress* 14, 74–99. doi: 10.1080/026783700417249

Young, M. S., and Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Hum. Factors* 44, 365–375. doi: 10.1518/0018720024497709

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., and Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. *Simul. Healthc.* 5, 267–271. doi: 10.1097/SIH.0b013e3181e3f329

Multisubject "Learning" for Mental Workload Classification Using Concurrent EEG, fNIRS, and Physiological Measures

Yichuan Liu [1,2], Hasan Ayaz [1,2,3,4*] and Patricia A. Shewokis [1,2,5]

[1] School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, United States, [2] Cognitive Neuroengineering and Quantitative Experimental Research Collaborative, Drexel University, Philadelphia, PA, United States, [3] Department of Family and Community Health, University of Pennsylvania, Philadelphia, PA, United States, [4] Division of General Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, United States, [5] Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University, Philadelphia, PA, United States

An accurate measure of mental workload level has diverse neuroergonomic applications ranging from brain computer interfacing to improving the efficiency of human operators. In this study, we integrated electroencephalogram (EEG), functional near-infrared spectroscopy (fNIRS), and physiological measures for the classification of three workload levels in an n-back working memory task. A significantly better than chance level classification was achieved by EEG-alone, fNIRS-alone, physiological alone, and EEG+fNIRS based approaches. The results confirmed our previous finding that integrating EEG and fNIRS significantly improved workload classification compared to using EEG-alone or fNIRS-alone. The inclusion of physiological measures, however, does not significantly improves EEG-based or fNIRS-based workload classification. A major limitation of currently available mental workload assessment approaches is the requirement to record lengthy calibration data from the target subject to train workload classifiers. We show that by learning from the data of other subjects, workload classification accuracy can be improved especially when the amount of data from the target subject is small.

Keywords: fNIRS, EEG, heart rate variability, respiration rate, n-back, mental workload, multimodal fusion, brain computer interface

INTRODUCTION

Mental workload refers to the cognitive and psychological effort required to complete given tasks. Continuous evaluation of mental workload enables real-time adjustment in the task load assigned to human operators so that their workload can be kept at a moderate level for improving human performance (Parasuraman et al., 1992; Parasuraman, 2003). Studies have thus far mainly decoded human workload levels from brain activity electroencephalogram (EEG) measures (Gevins et al., 1998; Brouwer et al., 2012). Cerebral hemodynamics have recently gained attention for applications in brain-computer interfaces (Naseer and Hong, 2015) and the decoding of mental workload level with the emergence of the portable measurement technique known as functional near-infrared spectroscopy (fNIRS) (Sassaroli et al., 2008; Ayaz et al., 2012; Herff et al., 2014). Previous studies have adopted a combination of EEG and non-brain measures such as heart rate variability, respiration rate, and eye movement (Hankins and Wilson, 1998; Wilson and Russell, 2003; Fairclough, 2009) for mental workload assessment. Moreover, results from our previous study

suggests that when combining EEG and fNIRS workload classification accuracies, they outperform the EEG-alone and fNIRS-alone results in mental workload level classification (Liu et al., 2017).

Before mental workload can be decoded from brain and body signals, it is typical that a time-consuming calibration process is required to derive a decoder for each individual operator. This is primarily due to the challenge that psychophysiological signals vary considerably between different people and over time. In the traditional calibration approach, lengthy psychophysiological signals (i.e., calibration data) need to be recorded from an operator so that a decoder can learn both the signal patterns specific to this operator and the variations of these patterns over time.

This problem is not unique to mental workload decoding. The lengthy calibration process is also required to decode other types of mental activities such as motor imagery (Blankertz et al., 2006). To address this problem for motor imagery decoding, Lotte and Guan proposed an alternative calibration approach (Lotte and Guan, 2010). In this approach, a decoder is derived using calibration data from both the target subject and some other subjects. Lotte and Guan argued that despite the large inter-subject variations, similar signal patterns can be found across some individuals so that less calibration data from the target subject is required to derive a decoder. This approach has been further investigated by other researchers, with positive results (Devlaminck et al., 2011; Samek et al., 2013). An alternative approach to learning from other subjects, is to identify which models incorporate the inter-subject variations from a large database (Fazli et al., 2009).

For mental workload decoding, only one preliminary study to date has explored the reduction of calibration time using a simulated aviation task (Wang et al., 2012). Authors have shown that calibrating decoders using data from both the target subject and a pool of other subjects did not degrade the decoding accuracies compared to using data only from the target subject. However, no benefit of including data from the other subjects has been shown.

In this study, the integration of EEG, fNIRS, and physiological signals was investigated for the classification of three workload levels induced by the n-back working memory task. The objective was two-fold: first, to compare the classification performance using the different modalities and their combinations; and, second, to investigate learning in a workload decoder using data from other subjects as an approach to improve workload classification performance when the sample size of the target subject is small.

MATERIALS AND METHODS

Participants

A total of 25 volunteers were recruited for participation in this study. Two of the participants were unable to finish the protocol. Another two participants were rejected from the analyses due to excessive movement. Consequently, a total of 21 valid subjects [all right-handed, 12 female, ages 25.9 ± 4.9 (mean $\pm$ SD)] were included in the analysis. The Edinburgh Handedness Inventory (Oldfield, 1971) showed that participants were right handed and

the average Laterality Quotient (L.Q.) and Decile is 78.7 ± 22.2 and 6.2 ± 3.4, respectively. Participants self-reported that they had their vision corrected to 20/20, did not have any history of neurological or psychiatric disorders and that they did not take any medication known to affect brain activity. Prior to the experiment, participants gave written informed consent for their participation in the study. The protocol was approved by the Institutional Review Board of Drexel University.

Recording

EEG, fNIRS, Heart rate, R-R interval, breath rate, and breath depth were simultaneously recorded during data collection. **Figure 1** shows an overview of the recording setup.

EEG were recorded using a Neuroscan Nuamp amplifier by Compumedics Neuroscan (http://compumedicsneuroscan.com/) from 26 locations according to the International 10–10 system (See **Figure 2**). Three additional electrodes, one placed above Nasion, the other two placed below the left/right outer canthus were used for electrooculography (EOG) artifact correction according to Schlögl et al. (2007). All 29 channels (26 EEG + 3 EOG) were band-pass filtered 0.1–100 Hz, digitally sampled at 500 Hz and referenced to a linked mastoid.

Prefrontal fNIRS were recorded from the forehead at a 2 Hz sampling rate using a 16-optode continuous wave fNIRS system developed at Drexel University (Ayaz et al., 2012, 2013) and manufactured by fNIR Devices LLC (http://fnirdevices.com/). The sensor included 4 light sources (LED) that can emit 730 and 850 nm wavelength light and 10 photon detectors (See **Figure 3**). The distance between light sources and detectors was 2.5 cm which allowed for a ∼1.2 cm penetration depth. To ensure repeatable sensor placement, the center of the sensor was aligned to the midline and the bottom of the sensor touched the participant's eye brow.

Systemic NIR were recorded from the right cheek at a 4 Hz sampling rate using a 2-optode continuous wave wireless fNIRS system developed at Drexel University (Ayaz et al., 2013) and manufactured by fNIR Devices LLC. The systemic NIR was not used in the current study.

Heart rate, R-R interval, breath rate, and breath depth were recorded using a Zephyr Bioharness chest band (https://www.zephyranywhere.com/).
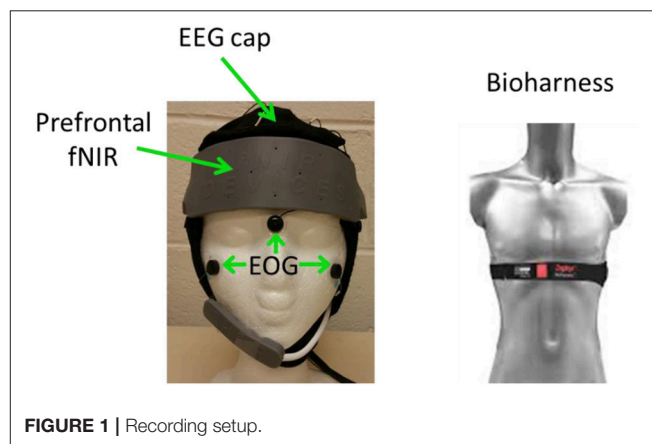
FIGURE 1 | Recording setup.
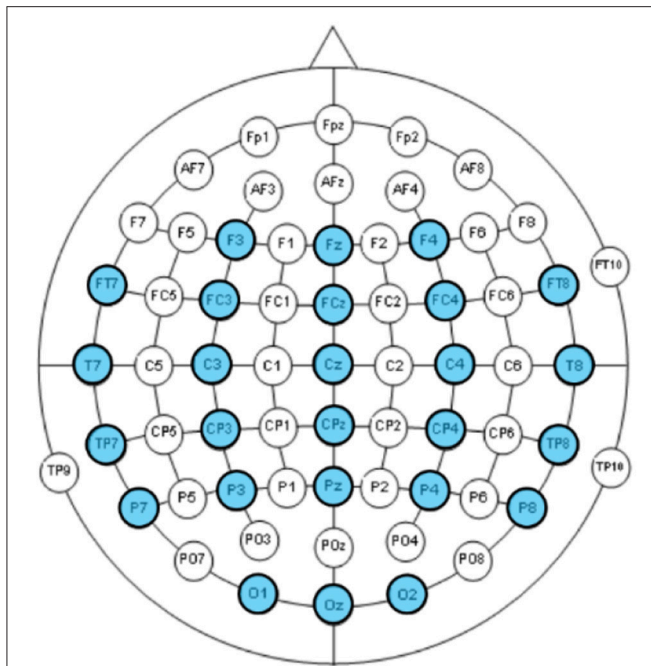
FIGURE 2 | EEG channels according to the International 10–10 system. The 26 recorded channels were highlighted.
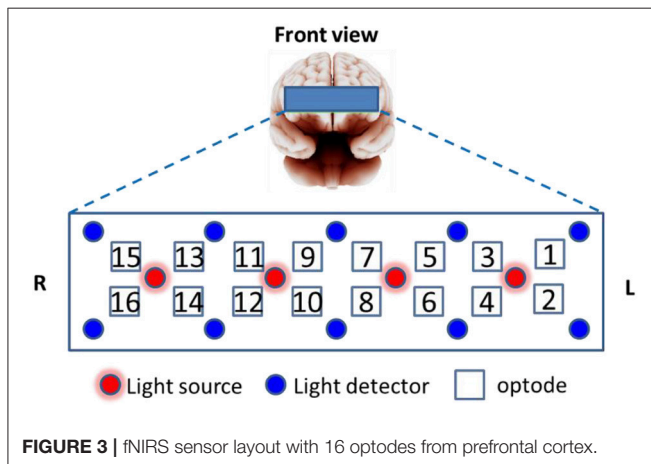
FIGURE 3 | fNIRS sensor layout with 16 optodes from prefrontal cortex.

Experiment

Subjects sat comfortably in front of an LED screen. Sequences of capitalized letter stimuli (∼1.7° visual angle) were shown on the center of the screen. The BCI2000 software was employed for stimulus delivery and for the recording of EEG and behavioral data (Schalk et al., 2004). Each letter was displayed for a duration of 480 ms and the inter-stimulus interval (ISI) was 2,520 ms. Subjects were instructed to click a keypad button with their right index finger in response to a "match stimulus" and click another keypad button with their right middle finger in response to a "non-match stimulus" as fast as possible. There were three workload conditions. In the 0-back condition, letter "X" was the match. In the 2-back condition, a letter was the match if it was shown two screens back. In the 3-back condition, a letter was the match if it was shown three screens back.

The letter stimuli were grouped into n-back blocks. Each block included 6 s of instruction, 45 s of task performance, and 15 s of fixation. The instruction period informed the subject which task (0-, 2-, or 3-back) to perform. During the task period, 15 letters were shown to the participants on the screen in a pseudo random order. Four of the letters were targets. No letters appeared more than twice in succession within a block. In the fixation period, subjects were instructed to focus their eye gaze on a white plus sign located at the center of the screen allowing fNIRS signals to return to the baseline. **Figure 4** shows the time line of a typical n-back block.

There were four recording sessions. Each session included 12 n-back blocks, 4 from each condition. Hence, there were 48 n-back blocks for the entire experiment, 16 from each condition. To reduce the correlation between adjacent samples and to balance time induced experimental factors such as fatigue across the three workload conditions, the 48 n-back blocks were grouped into 16 repetitions. Each repetition included one block from each workload conditions. The order of the blocks was further randomly shuffled so that no workload condition was repeated twice in succession within a session. Before the start of the first session, subjects practiced one block from each condition for familiarization with the procedure and an ∼5 min long EOG calibration session was performed during which subjects were instructed to rotate, blink and move (up/down, left/right) their eyes. A 5 min break was given to the subjects between the recording sessions. The entire recording time was about 1 h. **Figure 5** shows the outline of the experiment.

EEG Signal Processing

In this work, we extracted for each EEG channel the band powers of 1–3, 4–7, 8–12, 13–19, and 20–30 Hz bandwidths. This was performed at a single stimulus level, forming a feature vector f_{EEG} of *6 bands × 26 channels = 156* length for each of *the 48 blocks × 15 stimuli = 720* sample epochs for each subject.

Raw EEG and EOG signals are band-pass filtered 1–35 Hz. A regression-based approach was adopted to reduce EOG contamination by using the calibration data recorded before the n-back sessions started (Schlögl et al., 2007). Epochs were extracted 0–2.8 s and baseline corrected −0.2 to 0 s with respect to stimulus onset. The power spectral density of each epoch was then estimated using the Multitaper method (Thomson, 1982) with 8 Discrete Prolate Spheroidal Sequences (DPSS) window of 3 s long for subsequent analysis.

fNIRS Signal Processing

The average oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) amplitude change between (25 s, 45 s) and (−5 s, 5 s) with respect to the block start was used as a feature. The features were extracted from 14 forehead optodes, forming a feature vector f_{NIR} of *14 × 2(oxy/deoxy-Hb) = 28* length for each of the *48 sample blocks*. Optode 1 and 15 were rejected from analysis because they are over the hairline for most of the subjects. The average activation amplitude with respect to a baseline was adopted as the feature for characterizing the mental activities in many studies (Ayaz et al., 2007, 2012; Merzagora et al., 2009; Herff et al., 2012; Liu et al., 2013). This feature extraction strategy has been shown to be more reliable when

FIGURE 4 | Timeline of an n-back block.

FIGURE 5 | Experiment outline.
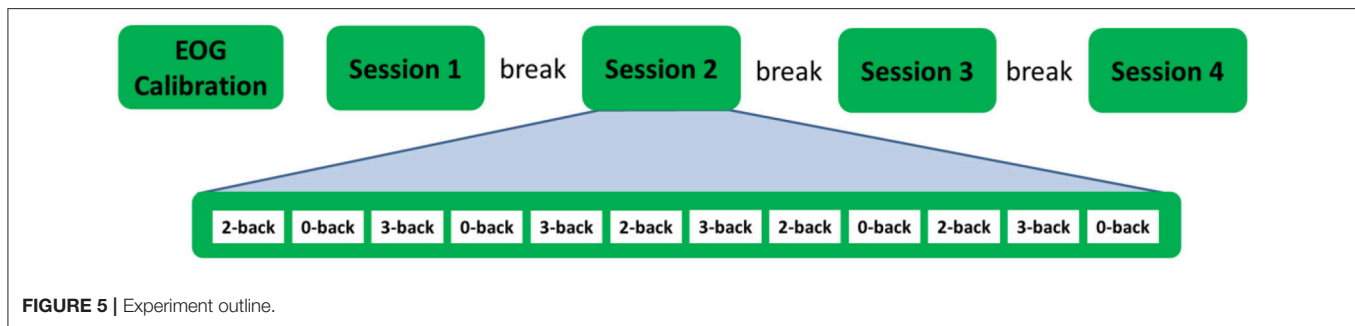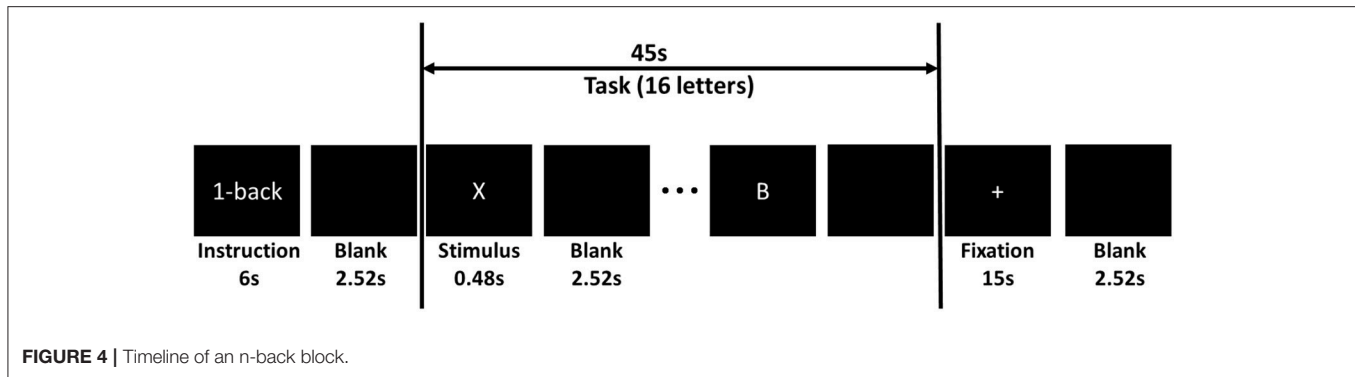
compared to other possible feature choices in our preliminary analysis.

Raw light intensities recorded from prefrontal fNIRS were first visually inspected to reject those optodes with inadequate contact or those positioned over the hairline. Raw light intensities were converted into concentration changes in oxy-hemoglobin (oxy-Hb) and deoxy-hemoglobin (deoxy-Hb) using the modified Beer-Lambert law (Cope and Delpy, 1988). Oxy-Hb/deoxy-Hb signals are band-pass filtered at 0.005–0.1 Hz for reducing artifacts from physiological signals (e.g., heartbeat and respiration) before subsequent analysis.

Heart Rate Variability (HRV) Processing

Heart rate variability (HRV) was estimated according to Clifford (2002) and Gritti et al. (2013). The R-to-R interval recorded by the Bioharness was first interpolated to form a 4 Hz time series. Epochs were extracted for each n-back block with (0 s, 45 s) time windows with respect to the onset of the first stimulus and the power spectral density (PSD) were estimated using a single DPSS window of 45 s long (Thomson, 1982) for evaluating the variability of the R-to-R interval. The average PSD in the bandwidths 0.02–0.06 Hz (mainly originated from body temperature regulation), 0.07–0.14 Hz (related to regulation of blood pressure), and 0.15–0.5 Hz (momentary respiratory influences on heart rate) were extracted as suggested by Scerbo et al. (2001).

In addition to HRV, the average of heart rate, breath rate, and breath depth for each n-back block recorded by Bioharness were extracted as features.

Multimodality Workload Classification

We considered the three-class classification problem of 3- vs. 2- vs. 0-back. A multiclass linear discriminant analysis (LDA) was adopted for classification. To prevent a covariance matrix from becoming singular due to small sample size, an automatic shrinkage using the Ledoit-Wolf lemma (Schafer and Strimmer, 2005) was adopted. The following eight different classifications were considered dependent on the adopted modalities (See **Figure 6**):

1) **EEG-alone**. A LDA was trained to classify EEG features at the single stimulus level (3 s time window with respect to a single stimulus). At the block level (45 s time window, included 15 stimuli), the LDA predicted probabilities for each of the 15 stimuli were Naïve-Bayes combined (Kuncheva et al., 2001) to produce $P\left(L|f_{EEG}\right)$ where $L \in \{0back, 2back, 3back\}$, which determined the predicted workload levels. More specifically, in Naïve-Bayes fusion, the product of the predicted probabilities from the 15 stimuli was calculated and normalized as follows:

$$P\left(L|f_{EEG}\right) = \frac{\prod_{i=1}^{15} P\left(L|f_{EEG}^i\right)}{\prod_{i=1}^{15} P\left(0back|f_{EEG}^i\right) + \prod_{i=1}^{15} P\left(2back|f_{EEG}^i\right) + \prod_{i=1}^{15} P\left(3back|f_{EEG}^i\right)} \qquad (1)$$
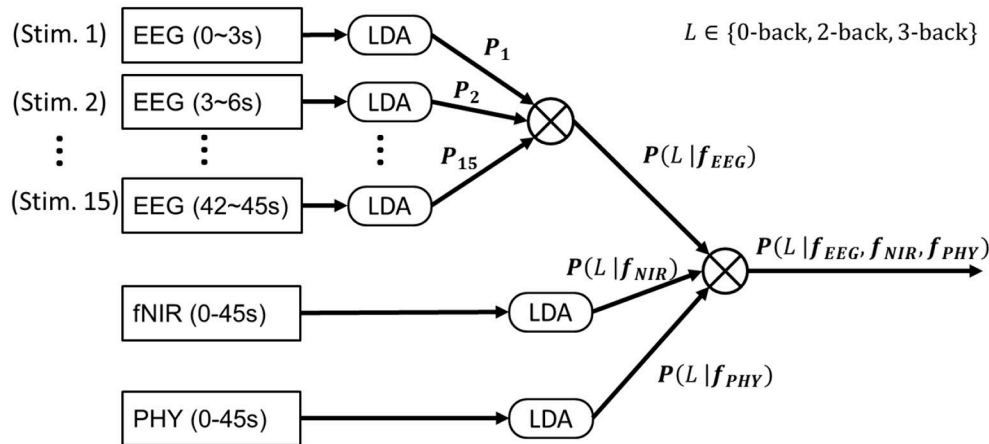
FIGURE 6 | EEG+fNIRS+PHY workload classification. A Linear Discriminant Analysis (LDA) was trained to classify EEG band power features at the single stimulus level (4 s epoch). The output probabilities from the 15 stimuli (of a block) were Naïve-Bayes combined to produce $P(L|f_{EEG})$. A second LDA was trained to classify fNIRS features extracted from each block (45 s epoch) to produce $P(L|f_{NIR})$. A third LDA was trained to classify PHY features extracted from each block to produce $P(L|f_{PHY})$. $P(L|f_{EEG})$, $P(L|f_{NIR})$, and $P(L|f_{PHY})$ were Naïve-Bayes combined for EEG+fNIRS+PHY classification. All of the above procedures were conducted on calibration data. The LDA classifiers were then applied on testing data to evaluate the classification performance.

2) **fNIRS-alone**. A LDA was trained to classify fNIRS features at the block level. The LDA probability output was termed $P(L|f_{NIR})$.

3) **PHY-alone**. A LDA was trained to classify PHY features (HRV, heart rate, respiration rate, and respiration depth) at the block level. The LDA probability output was termed $P(L|f_{PHY})$.

4) **EEG+fNIRS**. $P(L|f_{EEG})$ and $P(L|f_{NIR})$ were Naïve-Bayes combined for a final decision. More specifically, the product of the predicted probabilities from the two modalities was calculated and the output class c was determined as follows:

$$c = arg\ max_L[P(L|f_{EEG}) \cdot P(L|f_{NIR})] \qquad (2)$$

5) **EEG+PHY**. $P(L|f_{EEG})$ and $P(L|f_{PHY})$ were Naïve-Bayes combined for a final decision.

6) **fNIRS+PHY**. $P(L|f_{NIR})$ and $P(L|f_{PHY})$ were Naïve-Bayes combined for a final decision.

7) **EEG+fNIRS+PHY**. $P(L|f_{EEG})$, $P(L|f_{NIR})$ and $P(L|f_{PHY})$ were Naïve-Bayes combined.

Learning from Other Participants

We consider the following calibration approaches:

- **Traditional calibration**, which derives a classifier only with data from a specific subject. In terms of a LDA classifier, the mean and covariance matrix of the feature vector μ_i and Σ were estimated from a feature matrix extracted from the data of a target subject for estimating the posterior probability of a class given a feature vector. To obtain good classification accuracy, μ_i and Σ need to be estimated from a large data set recorded during a lengthy calibration session.

- **Proposed calibration**, which derives a classifier with data from both a specific subject and a pool of other subjects. As Lotte and Guan proposed, μ_i and Σ can be estimated from the target subject and a pool of other subjects to reduce the

calibration time of the target subject (Lotte and Guan, 2010). For each subject, the features were first z-score transformed to reduce the between-subject variations. For the target subject, only the training data was used for estimating the mean and variance of each feature. The mean and covariance matrix of the feature vector of each subject was then estimated. Finally, the mean and covariance matrices from all subjects were combined according to Equation (3) and Equation (4).

$$\mu_i = (1 - \lambda)\mu_i^t + \lambda \frac{1}{|S_t(\Omega)|} \sum_{j \in S_t(\Omega)} \mu_i^j \qquad (3)$$

$$\Sigma = (1 - \lambda)\Sigma^t + \lambda \frac{1}{|S_t(\Omega)|} \sum_{j \in S_t(\Omega)} \Sigma^j \qquad (4)$$

where μ_i^t and Σ^t are the mean and covariance estimated from the target subject, $S_t(\Omega)$ is a set of subjects that does not include the target subject (leave-one-subject-out) and $\lambda \in [0, 1]$ was a parameter representing the weight of other subjects. In this study, λ was empirically chosen to be 0.5.

When the sample size from the target subject is small, we expect that an improved classification performance can be achieved by incorporating the mean and covariance matrices estimated from other subjects.

Performance Evaluation

A repeated learning-testing method (Burman, 1989) was adopted for performance evaluation. The procedure was done as follows:

For subject $j = 1, \ldots, 21$:

1) For *iteration* $i = 1, \ldots, 100$:

   a. Data splitting:

The data of the target subject j were randomly split into a calibration set and a testing set three times with varying calibration sample size:

   i. 13 min calibration (12 samples), 39 min testing (36 samples).
   ii. 26 min calibration (24 samples), 26 min testing (24 samples).
   iii. 39 min calibration (36 samples), 13 min testing (12 samples).

  b. Classifier calibration:

   i. For traditional calibrations, the calibration sets were used to train the classifiers using LDA.
   ii. For the proposed calibrations, the calibration sets and data from all other subjects were used to train the classifiers.

  c. Classifier evaluation:

   i. The testing sets were used to evaluate the classification accuracy.

2) For each of the evaluated approaches, testing accuracies from the 100 iterations are averaged for a stable performance evaluation.

Multiple Comparisons

To correct for multiple testing, we adopted false discovery rate (FDR) control with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Without specification, we rejected null hypotheses for FDR $q < 0.05$.

RESULTS

Behavior Performance

To verify the successful manipulation of workload level with the adopted protocol, we evaluated the following three behavior measures:

1) **d-prime**, which was the key-press true positive rate minus false positive rate:

$$d\text{-}prime = \frac{N(stim = match \text{ and } responded = match)}{N(stim = match)}$$
$$- \frac{N(stim = non - match \text{ and } responded = match)}{N(responded = match)}$$

$$(5)$$

where $N(event)$ is the number of cases of an event, $stim = match/unmatch$ represent the true stimulus type, and $responded = match/unmatch$ represent subject's response.

2) **Accuracy**, which was the key-press accuracy.
3) **Response delay**: The time elapsed between stimulus onset and key-press.

For all three behavioral measures, one-way repeated measures ANOVAs revealed a significant effect of workload and *post-hoc* tests revealed significant differences (FDR $q < 0.05$) between all three workload levels, suggesting the successfully manipulation

of workload level (**Figure 7**). The generalized eta-squared (η^2) as reported by the ezANOVA library of R was used (Bakeman, 2005).

Effect of Workload on EEG Band Powers

Figure 8 depicts the topographic map of EEG band powers. A repeated measures ANOVA was applied to assess the effect of workload on the six mid-line channels Fz, FCz, Cz, CPz, Pz, and Oz and the results are shown in **Table 1**. For delta activity, a significant effect of workload was found at Cz and CPz (FDR $q < 0.05$) where the delta band power decreased with increased workload. Workload had a significant effect on theta band at channel Fz and Cz (FDR $q < 0.05$). At Fz, theta band increased with increased workload whereas at Cz, theta band power decreased with increased workload. Workload had a significant effect on alpha band power at all of the six midline channels Fz, FCz, Cz, CPz, Pz, and Oz (FDR $q < 0.05$). At all six channels, alpha band power decreased with increased workload. Workload has a significant effect on low beta band power at the six midline channels Fz, FCz, Cz, CPz, Pz, and Oz (FDR $q < 0.05$). At all six channels, low beta band power decreased with increased workload. Workload had a significant effect on high beta band power at Fz, FCz, Cz, CPz, and Pz (FDR $q < 0.05$). At these five channels, high beta band power decreased with increased workload. The significant effects that has been found in the low and high beta band may be confounded by motor responses as the 13–30 Hz range is typically associated with motor responses (Pfurtscheller et al., 1996, 2006). To investigate the effect of motor responses, a 2 (key press type: middle/index finger) × 3 workload level (0-/2-/3-back) ANOVA with repeated measures on both factors was conducted using the amount of key-press responses as the dependent. No significant effect of workload level [$F_{(2, 42)} = 0.83$, $p = 0.44$, $\eta^2 < 0.01$] or the interaction between key-press type and workload level [$F_{(2, 42)} = 2.01$, $p = 0.15$, $\eta^2 = 0.03$] was found. Mean and standard deviations of the number of key-presses within each block across the 21 participants for each of the three workload conditions can be found in Supplementary Table 1.

Effect of Workload on fNIRS Measures

Figure 9 shows the results from oxy-Hb. A common average reference approach was applied to remove the average oxy-Hb across all optodes and from each individual optode for reducing the effect of systemic physiological artifacts. Repeated measures ANOVA revealed a significant effect of workload on optode 5, 7, 8, and 14. *Post-hoc* tests revealed a significant 3-back > 0-back and 2-back > 0-back at optode 14. A optode 7, there was a significant effect of 3-back < 0-back and 3-back < 2-back. At optode 8, there was a significant effect of 3-back < 0-back and 2-back < 0-back. No significant *post-hoc* test results were detected at optode 5.

Effect of Workload on Physiological Measures

The effect of workload on physiological measurements are shown in **Figure 10**. For each subject, the average heart rate at 0-back, 2-back, and 3-back blocks were calculated respectively and the
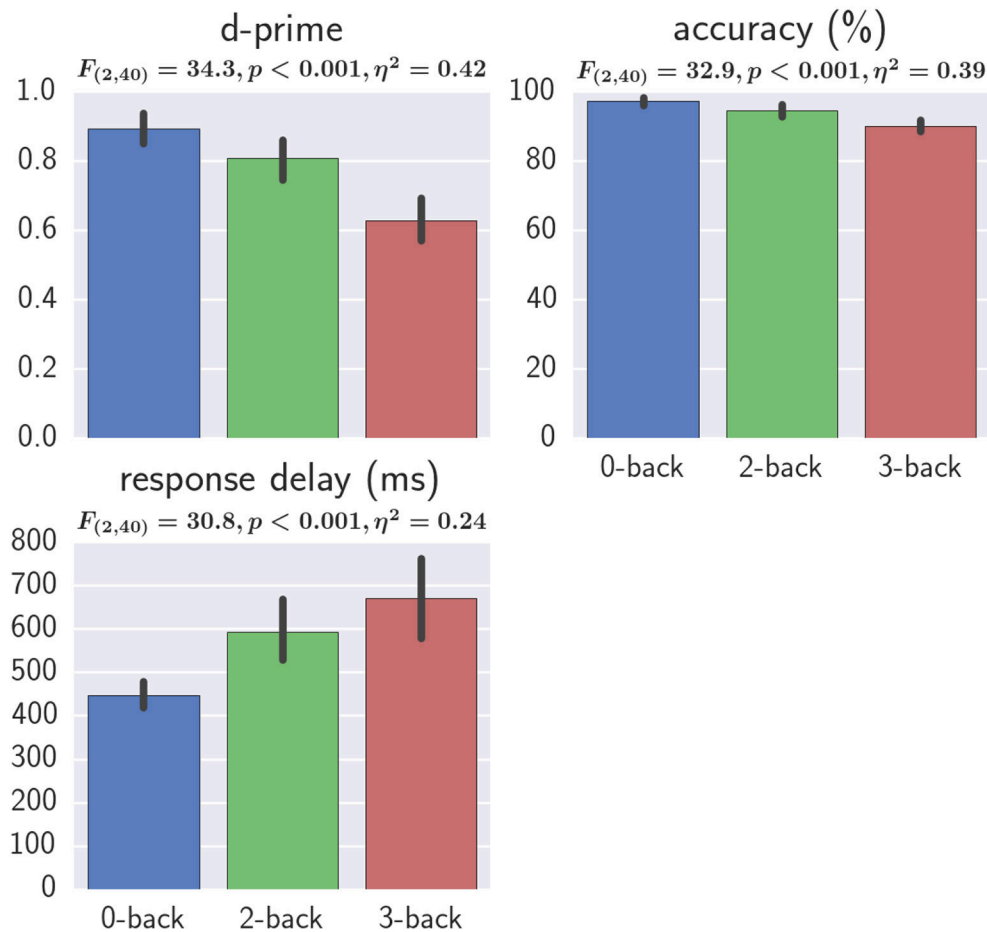
FIGURE 7 | Effect of workload on behavioral results. One-way repeated measures ANOVA results and the η^2 effect sizes with workload as the independent variable are shown. Error bars represent the bootstrapped 95% confidence interval.

three heart rate values were z-score standardized before analysis. The same preprocessing procedure was applied to the other physiological measurements: breath rate, breath amplitude and HRV. A repeated measures ANOVA revealed a significant effect of workload on breath amplitude, breath rate, heart rate, HRV 0.07–0.14 Hz, and HRV 0.15–0.5 Hz (FDR $q < 0.05$). *Post-hoc* tests revealed significant differences between 3-back and 0-back and also between 2-back and 0-back for breath amplitude, breath rate, heart rate HRV 0.07–0.14 Hz, and HRV 0.15–0.5 Hz (FDR $q < 0.05$).

Workload Classification

Workload classification results are shown in **Table 2**, **Figures 11**, **12** and Supplementary Tables 2, 3. For all investigated approaches and with the different calibration sample sizes, classification accuracy was significantly better than chance level (33.3%) as revealed by one-tailed Wilcoxon signed rank tests. **Figure 11** compares the accuracy using traditional and proposed calibration approaches. The results of the repeated measures ANOVAs indicate that the proposed calibration approach significantly outperforms the traditional

calibration approach for EEG-based classification, fNIRS-based classification, physiological based classification, and EEG-fNIRS multimodal classification ($p < 0.05$). The effect size of the results are shown in **Table 3**. *Post-hoc* analysis was conducted using a Wilcoxon Signed Rank test with FDR correction and the results are shown in Supplementary Table 4. For the calibration sample size of 13 min, the proposed calibration approach significantly outperformed the traditional calibration approach for EEG-alone, fNIRS-alone, PHY-alone, and EEG+fNIRS (FDR $q < 0.05$). For the calibration sample size of 26 min, the proposed calibration approach significantly outperformed the traditional calibration approach for EEG-alone, fNIRS-alone, and EEG+fNIRS (FDR $q < 0.05$). While for the calibration sample size of 39 min, no significant difference in classification accuracy can be found between the proposed and traditional calibration approach for all of the four modalities.

Figure 12 compares the classification accuracy using EEG-alone, fNIRS-alone with those using both EEG and fNIRS. A repeated measures ANOVA revealed that EEG-fNIRS significantly outperforms EEG-alone or fNIRS-alone for both traditional calibration approach and the proposed calibration
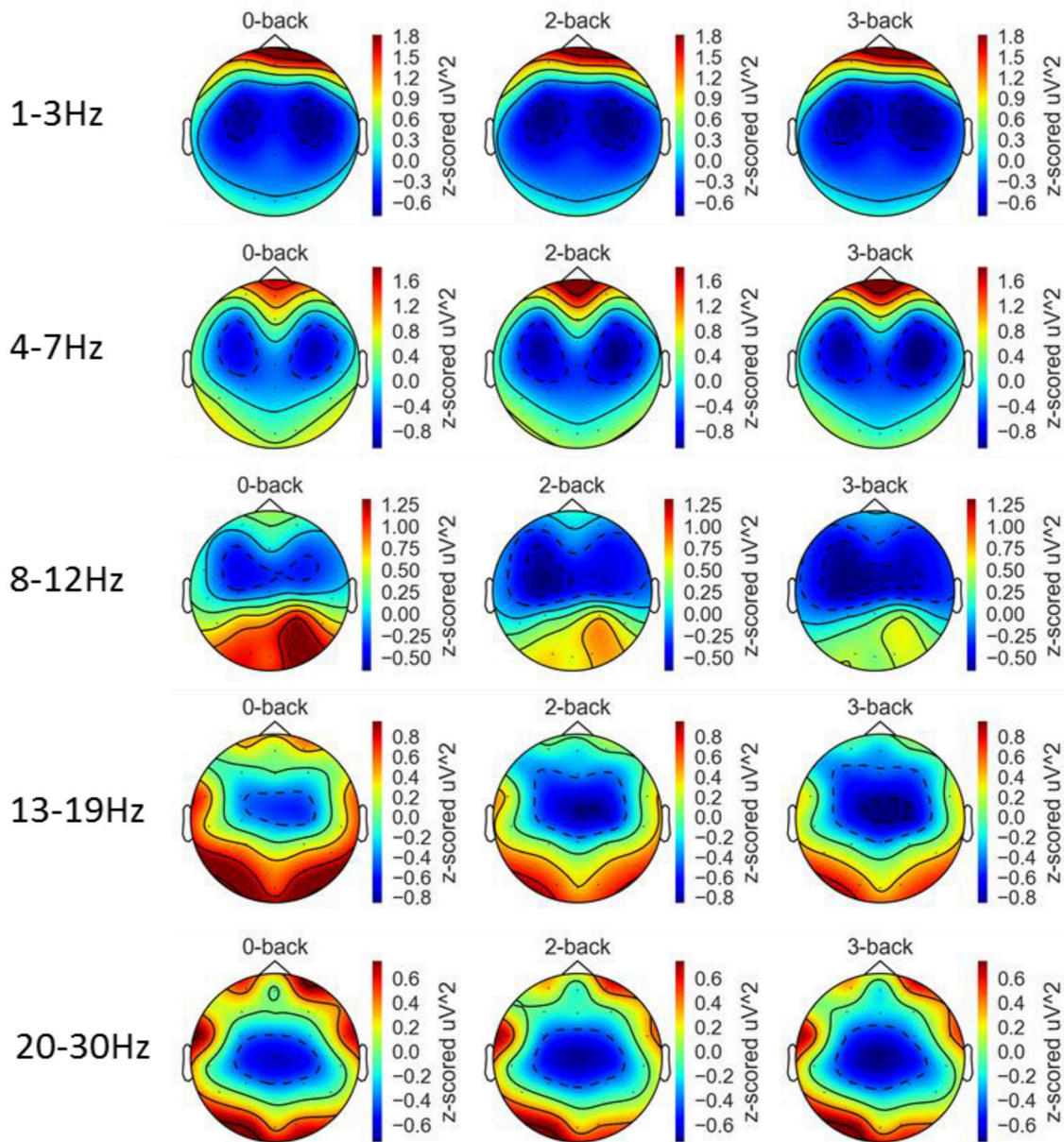
FIGURE 8 | Topographic maps of EEG band powers at the five frequency bands and three workload conditions averaged over 21 participants.

approach ($p < 0.001$). For the traditional calibration approach, an effect size dz of 0.81 [$t_{(20)} = 3.70$], 0.84 [$t_{(20)} = 3.83$], and 0.85 [$t_{(20)} = 3.88$] has been achieved for a calibration sample size of 13, 26, and 39 min, respectively, when comparing the EEG-alone and EEG+fNIRS approach. For the proposed calibration approach, an effect size dz of 0.89 [$t_{(20)} = 4.07$], 1.18 [$t_{(20)} = 5.43$], and 0.94 [$t_{(20)} = 4.33$] has been achieved for a calibration sample size of 13, 26, and 39 min, respectively, when comparing the EEG-alone and EEG+fNIRS approach. *Post-hoc* analysis was performed using a Wilcoxon Signed Rank test with FDR correction comparing EEG-alone and EEG+fNIRS with the results reported in Supplementary Table 5. For all three

calibration sample sizes and for both traditional and proposed calibration approaches, EEG+fNIRS significantly outperformed EEG-alone (FDR $q < 0.05$).

The effect of including a physiological-based classifier and combining them with EEG-alone, fNIRS-alone, and EEG-fNIRS classifier was studied and no significant improvement in classification was found.

DISCUSSION

In this study, the integration of EEG, fNIRS, and physiological measures investigated the classification of three workload levels.

To our knowledge, this is the first study that investigated the integration of fNIRS, EEG, and physiological signals for mental workload assessment. The n-back working memory task was adopted to induce three workload levels and the behavioral results suggested successful manipulation of the workload levels.

We first showed that in our data the EEG delta, alpha, low beta, and high beta activities decreased with increased workload levels

TABLE 1 | Effect of workload on EEG band powers.

	Delta $F_{(2, 40)}$, η^2	Theta $F_{(2, 40)}$, η^2	Alpha $F_{(2, 40)}$, η^2	Low beta $F_{(2, 40)}$, η^2	High beta $F_{(2, 40)}$, η^2
Fz	0.3, 0.01	**3.5, 0.15**	**21.2, 0.51**	**67.3, 0.77**	**17.8, 0.47**
FCz	2.9, 0.13	2.1, 0.10	**23.5, 0.54**	**66.4, 0.77**	**22.6, 0.53**
Cz	**7.3, 0.27**	**5.9, 0.23**	**14.5, 0.42**	**98.9, 0.83**	**31.6, 0.61**
CPz	**3.9, 0.16**	1.7, 0.08	**13.9, 0.41**	**85.8, 0.81**	**31.9, 0.61**
Pz	1.0, 0.05	1.5, 0.07	**13.4, 0.40**	**73.0, 0.78**	**28.6, 0.59**
Oz	0.1, <0.01	2.7, 0.12	**14.0, 0.41**	**10.7, 0.35**	1.0, 0.05

F-ratio statistics and the η^2 effect sizes are reported. Bold highlighted results are significant after correcting for multiple comparisons using FDR (Benjamini and Hochberg, 1995) (FDR q < 0.05).
Interpretation of the effect size η^2 are 0.02 = small, 0.13 = medium, and 0.26 or greater is large (Bakeman, 2005).

whereas theta activity increased with an increased workload level at the frontal site Fz. The suppression of alpha power in the posterior areas and increased theta power in the midline frontal areas under workload matches with the results reported in the literature (Gevins et al., 1997). It has been reported that beta activity decreased as workload increased at the midline central site Cz (Gevins et al., 1998). A previous study also suggested that the delta band decreased with increased workload level and the delta band carried information needed to characterize mental workload levels (Zarjam et al., 2011). Our results match those reported in the literature. A concern is that the workload effect on beta activities found in our study maybe caused by motor responses. The effect of workload and key-press type (middle/index finger) was assessed based on the number of key-press within each block and no significant effect of workload and the interaction between workload and key-press type was found. It is possible that motor activities other than key-presses could be affected by workload levels (e.g., subject may be more restless in the low workload condition) which need to be investigated in future studies.

For the fNIRS data, three prefrontal sites were found to be sensitive to workload changes with the 3-back task showing the highest level of activations. Previous fNIRS-based mental workload studies suggested that fNIRS was sensitive to workload changes (Ayaz et al., 2012; Fishburn et al., 2014; Herff et al., 2014).

FIGURE 9 | Effect of workload on fNIRS oxy-Hb. **(A)** Average fNIRS oxy-Hb at each workload level; **(B)** ANOVA F- ratio statistics and η^2 effect size of workload effect; **(C)** Average oxy-Hb for each optode. Highlighted optodes showed a significant workload effect (optode 5, 7, 8, and 14, FDR q < 0.05). Error bars illustrated the bootstrapped 95% confidence interval. Interpretation of the effect size η^2 are 0.02 = small, 0.13 = medium, and 0.26 or greater is large (Bakeman, 2005).

FIGURE 10 | Effect of workload on physiological measurements. One-way repeated measures ANOVA results and the η^2 effect sizes with workload as the independent variable are shown. All measures except for HRV (0.02–0.06 Hz) showed a significant effect of workload (FDR $q < 0.05$). Error bars represented the bootstrapped 95% confidence interval. BRAmplitude, breath amplitude. Interpretation of the effect size η^2 are 0.02 = small, 0.13 = medium, and 0.26 or greater is large (Bakeman, 2005).

TABLE 2 | Classification results using the traditional and proposed calibration approach.

Modality	Calibration sample size (min)					
	13		**26**		**39**	
	Proposed	**Traditional**	**Proposed**	**Traditional**	**Proposed**	**Traditional**
EEG-alone	56.0 ± 10.8	52.0 ± 10.6	61.1 ± 11.1	59.3 ± 11.5	63.4 ± 12.1	62.3 ± 12.5
fNIRS-alone	48.0 ± 10.1	46.1 ± 9.9	51.6 ± 12.0	49.8 ± 12.0	53.9 ± 12.8	51.9 ± 12.8
PHY-alone	41.0 ± 8.0	39.4 ± 8.1	41.8 ± 9.3	40.8 ± 9.5	42.2 ± 10.4	41.8 ± 10.0
EEG+fNIRS	56.9 ± 10.9	53.3 ± 11.0	62.6 ± 11.1	60.8 ± 11.8	65.0 ± 12.3	64.0 ± 12.7
EEG+PHY	55.9 ± 10.8	52.0 ± 10.4	61.1 ± 11.1	59.3 ± 11.4	63.4 ± 12.3	62.2 ± 12.5
fNIRS+PHY	48.7 ± 10.0	46.7 ± 9.7	52.7 ± 11.9	50.6 ± 12.1	55.0 ± 12.6	53.5 ± 13.1
EEG+fNIRS+PHY	56.8 ± 10.9	53.3 ± 10.9	62.5 ± 11.2	60.7 ± 11.7	65.1 ± 12.3	64.1 ± 12.6

Mean ± standard deviation of the classification accuracies are shown.

Again, our findings are consistent with the reported results in the literature.

For the physiological data, breath amplitude, breath rate, heart rate, HRV mid band (0.07–0.14 Hz), and HRV high band (0.15–0.5 Hz) were found to be sensitive to workload changes. The suppression of HRV spectral power in the 0.07–0.14 Hz range and 0.15–0.5 Hz range under workload was reported by the literature (Veltman and Gaillard, 1996; Nickel and Nachreiner, 2003). They suggested increased blood pressure and increased heart rate under high workload. Also reported was that breath rate increased with increased workload (Wilson and Eggemeier, 1991). Our results reflected these phenomena.

For workload classification, a significantly better than chance level classification was achieved by all investigated modalities: EEG-alone, fNIRS-alone, physiological alone, and EEG+fNIRS hybrid classification. For improving the classification accuracy when the calibration sample size is small, we proposed to calibrate classifiers using data from both the target subject and a pool of other subjects. Our results indicate that the proposed calibration approach significantly outperformed the traditional calibration approach which only used data from the target subject to calibrate classifiers regardless of the modality adopted. To our knowledge, this was the first study which demonstrated that learning from the data of multiple subjects outperforms learning
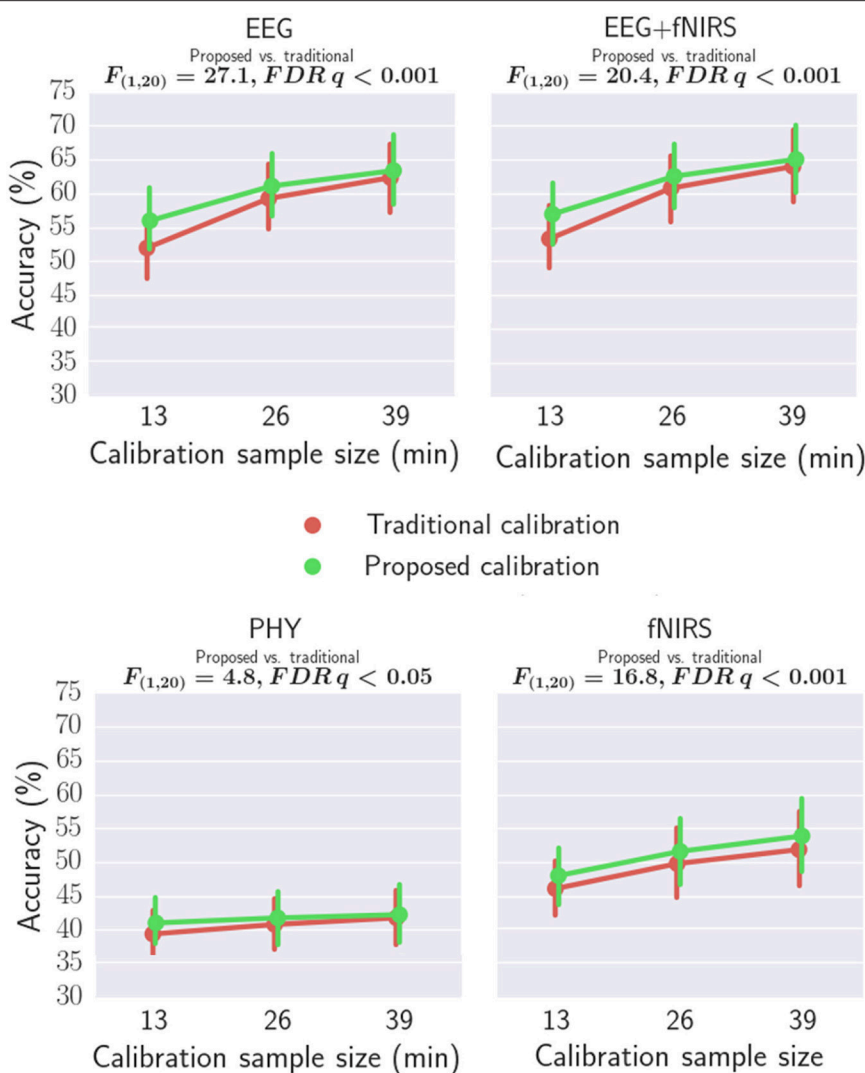
FIGURE 11 | Comparing the classification accuracies of traditional and proposed calibration approaches using EEG, fNIRS, PHY, and EEG+fNIRS.

TABLE 3 | Effect size and t-statistics of the improvement of the proposed calibration approach over the traditional approach.

Modality	Calibration sample size (min)	$t_{(20)}$	Cohen's dz	Effect size
EEG	13	8.66	1.89	Large
EEG	26	4.20	0.92	Large
EEG	39	1.76	0.38	Small
fNIRS	13	5.84	1.27	Large
fNIRS	26	3.63	0.79	Medium
fNIRS	39	2.75	0.60	Medium
PHY	13	3.96	0.87	Large
PHY	26	1.93	0.42	Small
PHY	39	0.77	0.17	Small
EEG+fNIRS	13	6.88	1.50	Large
EEG+fNIRS	26	3.55	0.78	Medium
EEG+fNIRS	39	1.77	0.39	Small

from a single subject for mental workload decoding accuracy. In the literature, various multisubject learning approaches have been proposed for the classification of different types of tasks. For example, Lotte et al. investigated multisubject learning for the classification of motor imagery tasks (Lotte and Guan, 2010) using EEG. To account for the inter-subject variability, they adopted a data-driven approach to select for each target subject a relevant subset of other subjects whose data can be used to improve the classification of the target subject. Reichert et al. investigated the classification of the phenomenal content of perception using fMRI (Reichert et al., 2014). To achieve cross-subject generalization, the weights of the classifiers trained from individual subjects were combined according to the individual classifier performance. Samek et al. found that the changes between training and testing data is similar across subjects and transferring this non-stationary information between subjects can help improve classification (Samek et al.,
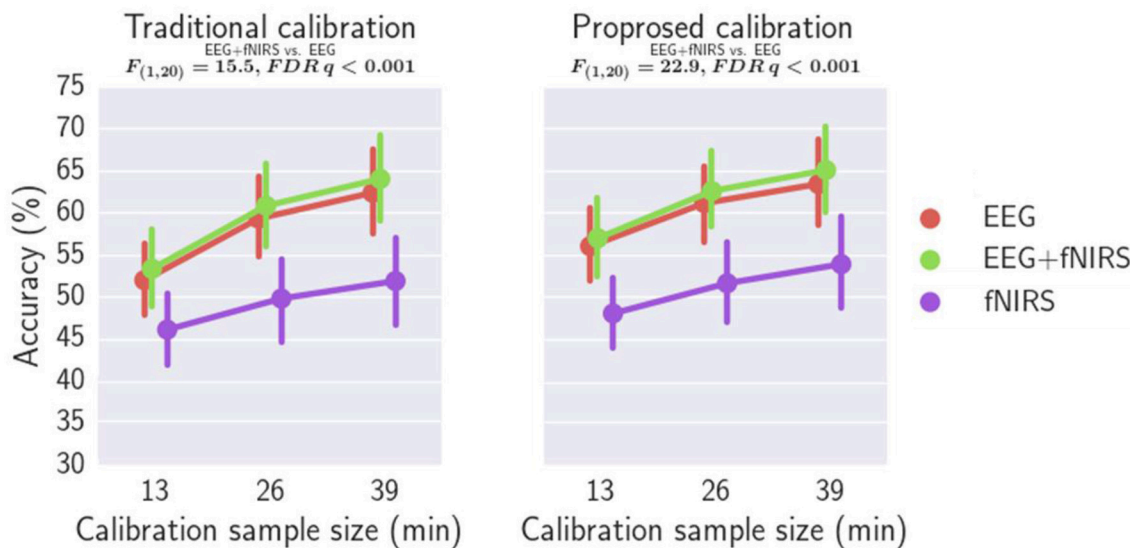
FIGURE 12 | Comparing the classification accuracies of EEG+fNIRS multimodal model and EEG-alone model.

2013). Our approach differs from these approaches in that the features for each subject were standardized before training to minimize the inter-subject variability. A further improvement to our approach may be achieved by performing subject subset selection as adopted by Lotte et al. or by weighting the mean and covariance matrices of each subject by their classification performance before applying Equation (3) and Equation (4). Finally, in this study, the hyperparameter λ in Equation (3) and Equation (4) is empirically chosen to be 0.5. The effect of λ on classification accuracies is provided in the Supplementary Figure 1. Estimating λ based on the individual classifier performance and the number of available samples from the target subject may further improve the performance of proposed approach.

Our results also suggest that EEG+fNIRS hybrid classification significantly outperformed EEG-alone or fNIRS-alone workload classification. These findings are consistent with our recent study (Liu et al., 2017) and indicate that there is complementary information about workload in EEG and fNIRS. However, the improvement of EEG+fNIRS over EEG-alone is only about 1–2% in classification accuracy. One possible reason behind this is the relatively low fNIRS-alone performance. It can be seen from **Table 2** that fNIRS-alone classification accuracy is about 10% lower than EEG-alone classification. A recent fNIRS-based workload estimation study reports that using only the forehead optodes resulted in a much-reduced workload estimation accuracy compared to using optodes from the whole head (Unni et al., 2017). We speculated that by using whole head coverage, the fNIRS-alone and EEG+fNIRS performance can be much improved. Finally, integrating physiological measures with EEG or fNIRS does not significantly improve workload classification. A reason for the lack of improvement in classification may be due the reduced reliability of the physiological based workload classification in comparison to the brain signal based approaches.

Another possibility may be that the physiological measurements do not provide additional information about workload to the brain signal measurements.

In conclusion, the current study presented various approaches for mental workload classification and demonstrated that with the integration of EEG and fNIRS and learning classifiers using the data from other subjects, workload classification performance can be improved. The proposed approaches may have applications in neuroegonomics research and applications such as adaptive aiding systems that are designed to improve the efficiency and safety of human-machine systems during critical tasks.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Institutional Review Board of Drexel University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Institutional Review Board of Drexel University.

AUTHOR CONTRIBUTIONS

YL had the largest contribution for all aspects of the work including the design and programming of the experimental testing, data acquisition, data processing, analysis, and interpretation as well as drafting and editing the manuscript. HA contributed to all aspects of the work with particular emphasis on the fNIRS and EEG application, signal processing, data analysis and interpretation, drafting and editing of the manuscript. PAS contributed to all aspects of the work with particular emphasis with the experimental design,

data analysis and interpretation, drafting and editing the manuscript as well as data acquisition. All authors agreed on the content and presentation of the submitted version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnhum.2017.00389/full#supplementary-material

REFERENCES

Ayaz, H., Izzetoglu, M., Bunce, S., Heiman-Patterson, T., and Onaral, B. (2007). "Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy," in *Proceedings of the 3rd International IEEE EMBS Conference on Neural Engineering* (Kohala Coast, HI: IEEE), 342–343.

Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P. A., McKendrick, R., and Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Front. Hum. Neurosci.* 7:871. doi: 10.3389/fnhum.2013.00871

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behav. Res. Methods* 37, 379–384. doi: 10.3758/BF03192707

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B.* 57, 289–300.

Blankertz, B., Dornhege, G., Krauledat, M., Muller, K., Kunzmann, V., Losch, F., et al. (2006). The Berlin brain-computer interface: EEG-based communication without subject training. *Neural Syst. Rehabil. Eng. IEEE Trans.* 14, 147–152. doi: 10.1109/TNSRE.2006.875557

Brouwer, A. M., Hogervorst, M. A., van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* 76, 503–514. doi: 10.1093/biomet/76.3.503

Clifford, G. D. (2002). *Signal Processing Methods for Heart Rate Variability.* Department of Engineering Science; University of Oxford.

Cope, M., and Delpy, D. T. (1988). System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Med. Biol. Eng. Comput.* 26, 289–294. doi: 10.1007/BF02447083

Devlaminck, D., Wyns, B., Grosse-Wentrup, M., Otte, G., and Santens, P. (2011). Multisubject learning for common spatial patterns in motor-imagery BCI. *Comput. Intell. Neurosci.* 2011:9. doi: 10.1155/2011/217987

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interact. Comput.* 21, 133–145. doi: 10.1016/j.intcom.2008.10.011

Fazli, S., Popescu, F., Danóczy, M., Blankertz, B., Müller, K. R., and Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural Netw.s* 22, 1305–1312. doi: 10.1016/j.neunet.2009.06.003

Fishburn, F., Norr, M., Medvedev, A., and Vaidya, C. (2014). Sensitivity of fNIRS to cognitive state and load. *Front. Hum. Neurosci.* 8:76. doi: 10.3389/fnhum.2014.00076

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Hum. Factors* 40, 79–91. doi: 10.1518/001872098779480578

Gevins, A., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cereb. Cortex* 7, 374–385. doi: 10.1093/cercor/7.4.374

Gritti, I., Defendi, S., Mauri, C., Banfi, G., Duca, P., and Roi, G. S. (2013). Heart rate variability, standard of measurement, physiological interpretation and clinical use in mountain marathon runners during sleep and after acclimatization at 3480 m. *J. Behav. Brain Sci.* 3:23. doi: 10.4236/jbbs.2013.31004

Hankins, T. C., and Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat. Space Environ. Med.* 69, 360–367.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935

Herff, C., Putze, F., Heger, D., Guan, C., and Schultz, T. (2012). "Speaking mode recognition from functional near infrared spectroscopy," in *Proceeding IEEE Engineering Medical Biology Society* (San Diego, CA), 1715–1718.

Kuncheva, L. I., Bezdek, J. C., and Duin, R. P. W. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognit.* 34, 299–314. doi: 10.1016/S0031-3203(99)00223-X

Liu, Y., Ayaz, H., Curtin, A., Onaral, B., and Shewokis, P. (2013). "Towards a hybrid P300-based BCI Using simultaneous fNIR and EEG," in *Foundations of Augmented Cognition*, eds D. Schmorrow and C. Fidopiastis (Berlin; Heidelberg: Springer), 335–344.

Liu, Y., Ayaz, H., and Shewokis, P. A. (2017). Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy. *Brain-Comput. Interfaces*, 1–11. doi: 10.1080/2326263X.2017.1304020

Lotte, F., and Guan, C. (2010). "Learning from other subjects helps reducing brain-computer interface calibration time," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference* (Dallas, TX), 614–617.

Merzagora, A. C., Izzetoglu, M., Polikar, R., Weisser, V., Onaral, B., and Schultheis, M. (2009). "Functional Near-Infrared Spectroscopy and Electroencephalography: a Multimodal Imaging Approach," in *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience,* eds D. Schmorrow, I. Estabrooke, and M. Grootjen (Berlin; Heidelberg: Springer), 417–426.

Naseer, N., and Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: a review. *Front. Hum. Neurosci.* 9:3. doi: 10.3389/fnhum.2015.00003

Nickel, P., and Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Hum. Factors* 45, 575–590. doi: 10.1518/hfes.45.4.575.27094

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Parasuraman, R. (2003). Neuroergonomics: research and practice. *Theor. Issues Ergon. Sci.* 4, 5–20. doi: 10.1080/14639220210199753

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., and Barnes, M. (1992). *Theory and Design of Adaptive Automation in Aviation Systems.* Available online at: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA254595

Pfurtscheller, G., Brunner, C., Schlögl, A., and Lopes da Silva, F. H. (2006). Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *Neuroimage* 31, 153–159. doi: 10.1016/j.neuroimage.2005.12.003

Pfurtscheller, G., Stancák, A. Jr., and Neuper, C. (1996). Post-movement beta synchronization. a correlate of an idling motor area? *Electroencephalogr. Clin. Neurophysiol.* 98, 281–293. doi: 10.1016/0013-4694(95)00258-8

Reichert, C., Fendrich, R., Bernarding, J., Tempelmann, C., Hinrichs, H., and Rieger, J. W. (2014). Online tracking of the contents of conscious perception using real-time fMRI. *Front. Neurosci.* 8:116. doi: 10.3389/fnins.2014.00116

Samek, W., Meinecke, F. C., and Muller, K. R. (2013). Transferring subspaces between subjects in brain–computer interfacing. *Biomed. Eng. IEEE Trans.* 60, 2289–2298. doi: 10.1109/TBME.2013.2253608

Sassaroli, A., Zheng, F., Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J., et al. (2008). Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *J. Innov. Opt. Health Sci.* 1, 227–237. doi: 10.1142/S1793545808000224

Scerbo, M. W., Freeman, F. G., Mikulka, P. J., Parasuraman, R., Di Nocero, F., and Prinzel, III., L. J. (2001). *The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology.* National Aeronautics and Space Administration; Langley Research Center.

Schafer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4, 1175–1189. doi: 10.2202/1544-6115.1175

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans Biomed Eng* 51, 1034–1043. doi: 10.1109/TBME.2004.827072

Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., and Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clin. Neurophysiol.* 118, 98–104. doi: 10.1016/j.clinph.2006.09.003

Thomson, D. J. (1982). "Spectrum estimation and harmonic analysis," in *Proceedings of the IEEE, Vol. 70*, 1055–1096.

Unni, A., Ihme, K., Jipp, M., and Rieger, J. W. (2017). Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: a realistic driving simulator study. *Front. Hum. Neurosci.* 11:167. doi: 10.3389/fnhum.2017.00167

Veltman, J. A., and Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biol. Psychol.* 42, 323–342. doi: 10.1016/0301-0511(95)05165-1

Wang, Z., Hope, R. M., Wang, Z., Ji, Q., and Gray, W. D. (2012). Cross-subject workload classification with a hierarchical Bayes model. *Neuroimage* 59, 64–69. doi: 10.1016/j.neuroimage.2011.07.094

Wilson, G. F., and Eggemeier, F. T. (1991). "Psychophysiological assessment of workload in multi-task environments." in *Multiple-Task Performance,* ed D. L. Damos (London: Taylor and Francis), 329–360.

Wilson, G. F., and Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Hum. Factors* 45, 635–643. doi: 10.1518/hfes.45.4.635.27088

Zarjam, P., Epps, J., and Chen, F. (2011). "Characterizing working memory load using EEG delta activity", in *Proceedings of the 19th European Signal Processing Conference EUSIPCO* (Barcelona), 1554–1558.

Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment

Carina Walter[1], Wolfgang Rosenstiel[1], Martin Bogdan[2], Peter Gerjets[3] and Martin Spüler[1]*

[1] *Department of Computer Engineering, Eberhard-Karls University Tübingen, Tübingen, Germany,* [2] *Department of Computer Engineering, University of Leipzig, Leipzig, Germany,* [3] *Knowledge Media Research Center, Tübingen, Germany*

In this paper, we demonstrate a closed-loop EEG-based learning environment, that adapts instructional learning material online, to improve learning success in students during arithmetic learning. The amount of cognitive workload during learning is crucial for successful learning and should be held in the optimal range for each learner. Based on EEG data from 10 subjects, we created a prediction model that estimates the learner's workload to obtain an unobtrusive workload measure. Furthermore, we developed an interactive learning environment that uses the prediction model to estimate the learner's workload online based on the EEG data and adapt the difficulty of the learning material to keep the learner's workload in an optimal range. The EEG-based learning environment was used by 13 subjects to learn arithmetic addition in the octal number system, leading to a significant learning effect. The results suggest that it is feasible to use EEG as an unobtrusive measure of cognitive workload to adapt the learning content. Further it demonstrates that a promptly workload prediction is possible using a generalized prediction model without the need for a user-specific calibration.

1. INTRODUCTION

Currently, there is an ongoing debate in mathematics education research on how to optimally support learners' during arithmetic learning (e.g., Askew, 2015; Calder, 2015). Obviously, learning outcome is most promising if the training program and learning content is tailored to the learner's specific needs (e.g., Gerjets and Hesse, 2004; Richards et al., 2007; for numerical interventions see Dowker, 2004; Karagiannakis and Cooreman, 2014). To optimally support the learner's efforts, learning content should neither be too easy, nor too difficult. Therefore, it is crucial for successful learning to keep the cognitive workload in the individual optimal range for each learner (Sweller et al., 1998; Gerjets et al., 2009). This can be achieved by adapting the difficulty of the learning content to the individual competencies of the learner.

Computer-supported learning (Kirschner and Gerjets, 2006) seems specifically suited for implementing adaptivity, because it is easy to implement algorithms that change the difficulty of the presented material based on the learner's behavioral response. This allows for an easy personalization of the learning environment to the user's individual needs, which is assumed to be necessary for efficient learning. So far, adaptive computer-supported learning environments rely on the user's interaction behavior for adaptation, e.g., error-adaptive systems, which change the task

difficulty based on the number of erroneous responses (Corbett, 2001; Graesser and McNamara, 2010; Käser et al., 2013). However, such behavioral measures are rather indirect and not very specific with respect to the cognitive processes required for performing the task at hand. For instance, more errors in a row may not only be caused by the difficulty of the task itself but also by task-unspecific processes (e.g., lapses of attention, fatigue, or disengagement).

It was recently proposed to measure cognitive processes directly to enhance human-computer interaction (Zander and Kothe, 2011). This approach, called passive brain-computer interface, could also be applied to improve arithmetic learning environments. Measuring neural correlates of specific cognitive processes allows for a more direct and implicit monitoring of the learner's cognitive state and should thereby allow for a better adaptation of the training content to improve the learning success of the user (Gerjets et al., 2014).

One cognitive state that is important within this context, is the working memory load, or in short workload. Throughout this paper, the term workload will be used as the amount of mental resources that are used to execute a specific task (based on Gevins and Smith, 2006). Working memory describes the small amount of information that can be stored and manipulated in mind simultaneously for the execution of a current cognitive task (Cowan, 2014). As the capacity for storing information at a time is limited, workload describes the extent to which this capacity is used or to which extent the working memory is filled. According to the cognitive load theory (Sweller et al., 1998), this is a bottleneck in learning, as the learning process is hindered if the amount of information the learner has to process exceeds the capacity of the working memory storage. Consequently, if the working memory load can be measured, this allows to adapt the presented learning content in a way that the storage capacities are never exceeded and the workload is always in an optimal range.

The amount of cognitive workload can be measured by Electroencephalography (EEG), which has been shown by multiple studies (Gevins et al., 1997; Murata, 2005; Berka et al., 2007; Wang et al., 2012). Basically, the amount of workload is reflected in two components of the EEG: the power spectrum and event-related potentials. Regarding the effect of workload on event-related potentials, two types of stimuli can be distinguished, task-independent as well as task-dependent stimuli. Using the n-back tasks for task-dependent stimuli, an increase in workload leads to a diminished P300 amplitude (Scharinger et al., 2017). Also for more complex tasks which include task-dependent as well as task-independent stimuli, like a piloting task, it was shown that the P300 amplitude is lower in high workload conditions (Causse et al., 2015). Brouwer et al. (2012) stated this diminishing effect in the P300 amplitude for task-independent stimuli. Besides the P300, also other components of the event-related potential were shown to be sensitive to workload, like the N100 (Ullsperger et al., 2001) or the N200 and mismatch negativity (Kramer et al., 1995), using task-independent stimuli. In Roy et al. (2016a) a significant decrease of the P200 component was stated when workload increased. While ignoring infrequent task-independent auditory probes, they were able to assess mental workload efficiently.

Further, the oscillatory activity in EEG is also affected by workload. Pesonen et al. (2007) have shown that there are workload related changes in theta-, alpha-, and beta-band. Brouwer et al. (2012) also found an increase in frontal theta power and a decrease in occipital alpha power in an n-back task. Specific to arithmetic tasks, it was shown that the cognitive demand results in an increasing power of the theta band and a decreasing power in the alpha band (Harmony et al., 1999).

Regarding an EEG-based prediction of workload, Kohlmorgen et al. (2007) have presented a real-time system in which the workload induced by mental calculation task while driving could be predicted. Brouwer et al. (2012) classified high- against low workload in an n-back task and achieved classification accuracies above 80% using either spectral features, ERPs or a combination of both. Roy et al. (2016b) also compared the classification accuracy of power spectrum and ERP features during a Sternberg memory task and achieved a low performance (60%) with spectral data, while achieving high performance (91%) with ERP data. In a previous study (Walter et al., 2014), we tried to predict the difficulty of arithmetic tasks based on theta, alpha and beta power and achieved an average correlation coefficient of up to 0.88. In a following study (Spüler et al., 2016), we improved the cross-subject prediction and trained a prediction model that works across subjects yielding an average correlation coefficient of 0.82.

While there is enough evidence showing that an EEG-based workload prediction is possible and can be implemented in a real-time system, it has not been used for a closed-loop adaptation in an online learning environment, to the best of our knowledge. In this paper, we show how we developed a learning environment for arithmetic exercises that adapts the task difficulty based on learner's cognitive workload as predicted from the EEG. To concentrate on the usability of the EEG-based learning environment, we wanted a system that works out-of-the-box without the need for a subject specific calibration phase. As we also wanted a stimulus-independent system, only the power spectrum was used for prediction. In the following, we present the EEG-based workload prediction and describe its application in the EEG-based learning environment. To show the feasibility of the EEG-based learning environment, the learning effect for 13 subjects testing this environment is compared to a control group using an error-based learning environment.

2. METHODS

For the development of an EEG-based workload environment, the work was split in two studies. In the first study, EEG data were collected while subjects solved arithmetic tasks of varying difficulty. Based on this data a prediction model was created that predicts the amount of workload based on the user's EEG. In the second study, which is the main part of this paper, the prediction model was used to estimate a learner's workload online and adapt the task difficulty accordingly. To evaluate the EEG-based learning environment, subjects used it to learn arithmetic addition in the octal number system (e.g., $3 + 5 = 10$) and the learning success was compared

to a control group, who learned the same task using a learning environment that adapts based on the number of correct responses.

2.1. Task Design

The participants solved addition tasks with diverse levels of difficulty, which where presented on a desktop computer with a 19 inch TFT display. In the first study, addition tasks in the decimal number system were presented, while the participants had to learn addition in the octal number system in the second study. The difficulty level of the presented addition exercises was defined by their Q-value (Thomas, 1963), which reflects the information content of an arithmetic task. This difficulty measurement takes into account both, problem size and the need for a carry over operation, which are the main parameters for problem difficulty in addition.

A more detailed description on calculating the Q-value can be found in Thomas (1963) or in Spüler et al. (2016), where also some examples are shown. The addition problems presented in this work were ranging from $Q = 0.6$ (easy, single-digit, e.g., $1 + 1$) to $Q = 7.2$ (difficult, four-digit, e.g., $3721 + 1452$).

Each trial, in which one arithmetic problem should be solved, consisted of four phases, which are depicted in **Figure 1**. First, the calculation phase occurred, where the problem to be solved was shown for 5 s. Subsequently, subjects had a maximum of 3.5 s to type in their result. In the first study, with arithmetic's in the decimal number system, the subjects did not receive feedback. In the second study, the subjects had to learn arithmetic's in the octal number system and therefore got feedback by presenting the correct answer for 3.5 s. Each trial ended with an inter-trial interval (ISI) of 1.5 s, resulting in a total length of approximately 45 min. To avoid the prediction model of being based on perceptual-motor confounds, the time windows used for analyzing EEG data should not contain motor events. As typing in the answer leads to motor artifacts, the calculation phase was used for EEG analysis only.

2.2. EEG Recording

A set of 28 active electrodes (actiCap, BrainProducts GmbH), was used to record EEG signals. They were attached to the scalp, placed according to the extended international electrode $10-20$ placement system (FPz, AFz, F3, Fz, F4, F8, FT7, FC3, FCz, FC4,

FIGURE 1 | Schematic flow of the learning phase for the EEG-based, as well as the error-based adaptive learning environment. The gray line indicates the calculation phase (I_A), whereas the black dashed line represents the response interval. Subsequently a feedback phase occurs indicated by the black line, followed by an inter-stimulus interval (ISI) shown as white line.

FT8, T7, C3, Cz, C4, T8, CPz, P7, P3, Pz, P4, P8, PO7, POz, PO8, O1, Oz, and O2). Three additional electrodes were used to record an electrooculogram (EOG); two of them were placed horizontally at the outer canthus of the left and right eye to measure horizontal eye movements and one was placed in the middle of the forehead between the eyes to measure vertical eye movements. Ground and reference electrodes were placed on the left and right mastoids. EOG- and EEG-signals were amplified by two 16-channel biosignal amplifier systems (g.USBamp, g.tec) and sampled at a rate of 512 Hz and the impedance of each electrode was less than $5 k\Omega$. EEG data were band-pass filtered between 0.5 and 60 Hz with a Chebyshev filter of order 8 during the recording. Furthermore, a notch-filter (Chebyshev, order 4) was applied between 48 and 52 Hz to filter out power line noise. The signal processing pipeline is documented in detail in **Figure 2**.

2.3. Study 1: Workload during Decimal Arithmetic Tasks

2.3.1. Study Design and Participants

Ten students (4 male and 6 female; range: $17 - 32$ years, $M = 24.9$ years, $SD = 5.3$ years) participated voluntarily in this study and received monetary compensation for participation. The study was approved by the local ethics committee of the Medical Faculty at the University of Tübingen and written informed consent was obtained by the participants. All participants reported normal or corrected to normal vision and no mathematical problems. Participants were chosen randomly. Nevertheless, all were university students (with different fields of study) and can thus be considered as having a high educational background.

Each participant had to solve 240 addition exercises in the decimal number system while EEG was measured. The exercises were presented with an increasing difficulty due to learning effects. If a subject learns and gets better at performing the tasks, exercises with the same objective difficulty (Q) can lead to a different perceived difficulty at the start (before learning) and at the end of the session (after learning). If exercises are presented in increasing difficulty, the relationship that a task with higher perceived difficulty has also a higher objective difficulty (Q) still holds.

However, it should be noted that the increasing task difficulty over times leads to potential confounds, as there might also be changes in EEG over time that are not related to task difficulty. The only possibility to counter these confounds would be a randomization of the task difficulty, which in turn would corrupt the labels, as the relationship between Q and the perceived difficulty would change over time due to learning effects. Since we needed reliable labels for the training of the prediction model, we decided against a randomization.

2.3.2. Analysis of EEG Data

EEG data were corrected for eye-movements using an EOG-based regression method (Schlögl et al., 2007). Furthermore, Burg's maximum entropy method (Cover and Thomas, 2006) with a model order of 32 was used to estimate the power spectrum from 1 to 40 Hz in 1 Hz bins. After calculating the power

FIGURE 2 | Flow chart of the EEG signal processing pipeline to generate a classifier for workload prediction, used in both studies.

spectrum, the data was *z-score* normalized along the channels to correct for inter-subject variability in the subjects baseline EEG power. To analyze which electrodes and frequencies change with increasing task difficulty, the signed squared correlation coefficients (sign $(R) * R^2$) between the power at each frequency bin (for each electrode), as well as Q-values of the corresponding trials were calculated.

2.4. Study 2: EEG-Based Learning Environment for Octal Arithmetic Tasks
2.4.1. Study Design and Participants
The participants in the second study were divided into two groups: an experimental group using the EEG-based learning environment and a control group using an error-adaptive learning environment. In both groups, the subjects learned

arithmetic addition in the octal number system (e.g., $5 + 3 = 10$), which was a completely new task to all subjects.

To evaluate the learning success, each subject did a pre-test and a post-test, before and after using the learning environment for approximately 45 min (180 exercises). The tests consisted of 11 exercises, with varying difficulty. Although difficulty of the exercises were the same for pre- and post-test, the exercises itself were different.

The participants of both groups were university students of various discipline, reported to have normal or corrected to normal vision and participated voluntarily in the EEG experiment. 13 subjects (7 male and 6 female; range $21 - 35$ years, $M = 28.1$ years, $SD = 4.3$ years) participated in the experimental group using the EEG-based learning environment. The control group consisted of 11 subjects (7 male and 4 female; range $22 - 27$ years, $M = 23.4$ years, $SD = 1.4$ years), using an error-adaptive learning environment which is state of the art. The study was approved by the local ethics committee of the Medical Faculty at the University of Tübingen and written informed consent was obtained by the participants.

2.4.2. Cross-Subject Regression for Online Workload Prediction

Based on the EEG data from study 1, we created a prediction model that was used to predict the cognitive workload in a timely manner and further was able to adjust the learning environment accordingly. In terms of usability, we wanted the EEG-based learning environment to be useable out-of-the-box without the need for a subject-specific calibration phase, which is why a cross-subject regression method as presented by Walter et al. (2014) was applied.

Therefore, EEG data obtained in study 1 were used for training a linear ridge regression model with a regularization parameter of $\lambda = 10^3$. For training the ridge regression, we used the MATLAB function *ridge* and the regularization parameter was found to be optimal in the previous study, where it was determined by cross-validation. The number of electrodes used for online adaptation was reduced to 16 inner electrodes (FPz, AFz, F3, Fz, FC3, FCz, FC4, C3, Cz, C4, CPz, P3, Pz, P4, Oz, and POz), to be consistent with the electrode positions used in the previous cross-subject study (Walter et al., 2014), where the outer electrodes were not used as those are more prone to artifacts and contain less relevant information. Furthermore, only trials with a Q-value smaller than 6 were used to train the regression model, since trials with higher Q-value showed similar EEG patterns as very easy trials, which is most likely due to a disengagement of the subjects (Spüler et al., 2016). The power spectrum was calculated for the 5 s time frame of the calculation phase using the Burg's maximum entropy with a model order of 32. To correct for inter-subject variability in the subjects baseline EEG power, the data was *z-score* normalized along the channels. For the final prediction output, we calculated the moving average with a window length of 6 trials in 1 trial steps, which still guaranteed a response time smaller than 1 min of the system. This delay is feasible for the detection of workload since it is not recommendable to adapt an online learning environment too fast (i.e., single trial duration). The moving average also leads to a more robust

prediction (Walter et al., 2014), but makes the system react slower to sudden changes in workload, which is feasible since it is not recommendable to adapt the difficulty of an online learning environment too rapidly.

In contrast to study 1, we did not use an EOG-correction in this study. Although it is technically possible, it would not fit with our approach to build an out-of-the-box system, as training data for the EOG-correction would be needed for each subject. This would cost additional time before the user can use the system (and start learning) and therefore would decrease usability of the system.

The so trained regression model was then applied in the EEG-based learning environment, to predict the amount of cognitive workload for novice subjects online.

2.4.3. Online Adaptation of the EEG-Based Learning Environment

For the experimental group, the EEG data served as workload indicator. Therefore, we used the output of the previously trained regression model to predict the current workload state of each learner and differentiated three difficulty levels. If the predicted workload was less than $Q = 0.8$, the presented task difficulty was assumed to be too easy. Thus the following Q-value was increased by 0.2. Vice versa, the target Q of the subsequent task decreased by 0.2 when the predicted workload was greater than $Q = 3.5$. In this case, the presented task difficulty was assumed to be too difficult. If the predicted workload was between $Q = 0.8$ and $Q = 3.5$, the Q-value for the next presented task remained the same and the difficulty level was kept constant. These thresholds were defined based on the results in Spüler et al. (2016). Trials with $Q < 1$ were solved correctly in all cases, while none of the subjects were able to solve trials with a $Q > 6$. 50% of the trials with a $Q = 3.5$ were successfully solved on average. The learning session for the experimental group started with an exercise of difficulty level $Q = 2$.

2.4.4. Adaptation of the Error-Based Learning Environment

For the control group, an error-adaptive learning environment was used. The number of wrong answers served as performance and adaptation measure. When subjects solved five consecutive tasks correctly, the difficulty level (Q) increased by 1. Vice versa, the difficulty level decreased by 1 when participants made three errors in a row. Otherwise, the Q-value did not change and the difficulty level was held constant. The adaptation scheme was kept similar in the control group, as in common tutoring systems. The learning session for the control group started with an exercise of difficulty level $Q = 2$.

2.4.5. Evaluating Learning Success

To compare the learning success of the two groups, the learning effect after completing the learning phase serves as performance measure and is used as an indicator of how successful each subject was supported during learning. Hence, each subject had to perform a pre-test before the learning phase started. This was used to assess the prior knowledge of each user. After the learning session, each participant had to solve a post-test, and the

difference in score between the two tests served as indicator of the learning effect. Results of the pre- and post-tests were compared statistically using a two-sided Wilcoxon's ranksum test.

3. RESULTS

3.1. Study 1: Workload Related Effects in EEG Data

As results from study 1 were already published in Spüler et al. (2016), only the most important results relevant for the implementation of the online learning environment are shown here. For a more detailed analysis, we refer to the original publication.

Results from the analysis of the EEG data regarding the association between the Q-value and the power at each electrode and frequency bin, the R^2-values are shown in **Figure 3**. In the delta frequency band, there is a small difficulty-related effect over the central electrodes, while the effect in the theta, alpha, and beta frequency band is located over the parieto-occipital electrodes. This effect was strongest for the alpha band ($8 - 12\,Hz$). While the lower beta band ($13 - 24\,Hz$) still shows some effects related to task difficulty, they cannot be observed in the upper beta band ($25 - 40\,Hz$).

3.2. Study 2: Task Performance Results

The behavioral results, how well the subjects performed the octal arithmetic task, are shown in **Table 1** for both groups. For the experimental group, 45.5% of the 180 exercises were solved correctly on average. Averaged over all subjects, a maximum Q-value of 5.85 was reached by using the EEG-based learning

environment. Each subject achieved at least the difficulty level of $Q = 3.2$.

The control group answered 64% of all 180 assignments correctly on average. Since the error-rate was used for adapting the difficulty level of the presented learning material, the number of correctly solved trials was similar across subjects. On average, a maximum Q-value of 4.64 was reached (see **Table 1**). The best subjects reached a maximum Q-value of 6, whereas each participant achieved at least the difficulty level of $Q = 4$.

3.3. Study 2: Learning Effect Using Adaptive Learning Environments

To evaluate if the EEG-based learning environment works and how it compares to an error-adaptive learning environment, we analyzed the learning effect of each subject by pre- and post-tests. Furthermore, the learning effects between the experimental and the control group were compared.

Table 2 reports the learning effect of each individual subject, as well as the group averages for both groups. After using the EEG-based learning environment and thus learning how to calculate in an octal number system, a learning effect can be recognized for almost every subject of the experimental group, except for subject S02 and S05. On average, 5.08 assignments from 11 post-test tasks were solved correctly after completing the learning phase, 3.54 more assignments compared to the pre-test. On average, a significant learning effect can be verified between the pre- and post-test ($p = 0.0026$, two-sided Wilcoxon test).

The control group solved 1.55 tasks from the 11 pre-test assignments on average correctly. As the difference to the experimental group is not significant ($p > 0.05$, two-sided

FIGURE 3 | Topographic plots of the signed squared correlation coefficient averaged over all subjects between the frequencies in each band for each electrode and the task difficulty as indicated by the Q-value.
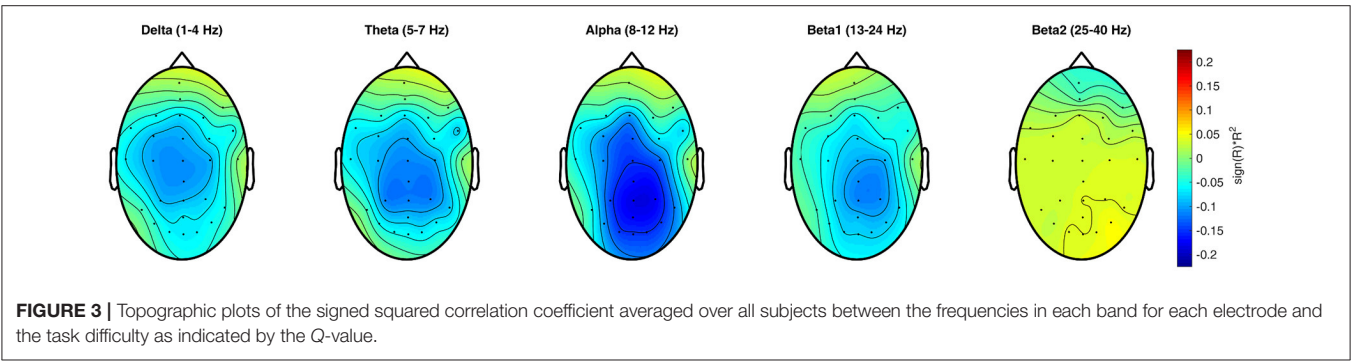
TABLE 1 | Task performance for all subjects.

EEG-based learning	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	mean
Correct	38.3	88.9	60.6	10.0	28.9	25.6	31.1	17.8	66.1	33.3	16.7	92.2	82.2	45.5
max. Q	6.6	3.2	6.6	6.6	6.6	6.6	6.6	6.6	5.2	6.6	6.6	3.6	4.6	5.85
Control group	**S14**	**S15**	**S16**	**S17**	**S18**	**S19**	**S20**	**S21**	**S22**	**S23**	**S24**			**mean**
Correct	64.4	62.2	57.2	66.7	63.3	61.1	63.3	67.8	65.0	66.1	66.7			64.0
max. Q	6	4	5	6	4	5	4	4	4	5	4			4.64

The relative amount of correctly solved trials in % and the maximum Q-value for each subject, as well as the group average are shown. Results of the experimental group using the EEG-based learning environment are displayed at the top, while results from the control group using an error-adaptive learning environment are displayed at the bottom.

TABLE 2 | Number of correctly solved trials in the pre-and post-test, as well as the difference, which indicates the learning success for the individual subjects and the group means.

EEG-based learning	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	mean
Pre-Test	3	0	6	0	2	1	0	1	0	5	1	1	0	1.54
Post-Test	7	0	9	8	2	3	3	2	8	7	7	2	8	5.08
Difference	4	0	3	8	0	2	3	1	8	2	6	1	8	3.54
Control group	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24			mean
Pre-Test	4	3	1	0	2	1	2	1	1	2	0			1.55
Post-Test	8	2	4	7	5	8	3	4	4	6	4			5
Difference	4	-1	3	7	3	7	1	3	3	4	4			3.45

Results of the experimental group using the EEG-based learning environment are displayed at the top, while results from the control group using an error-adaptive learning environment are displayed at the bottom.

Wilcoxon test), equal prior knowledge can be implied for both groups. The best subject of the control group performed 36.36 % of the pre-test tasks accurately, whereas the worst subjects gave no correct answers. For almost every subject, a learning effect for calculating in the octal number system is noticeable after the error-adaptive learning session, except for subject S15. On average, 3.45 more tasks were solved correctly in the post-test, compared to the pre-test, which also shows a significant learning effect between the pre- and post-test ($p = 0.0016$, two-sided Wilcoxon test).

Although the learning effect for the experimental group using the EEG-based learning environment is higher than for the control group, this difference is not significant ($p > 0.05$, two-sided Wilcoxon test).

4. DISCUSSION

In the presented work, we have shown that EEG can be used to measure a learner's workload online and adapt a learning environment accordingly. By using a cross-subject regression method, a subject-specific calibration phase can be omitted. Although the cross-subject prediction model was build using data recorded while subjects did arithmetics in the decimal number system, it could be successfully applied to predict the workload during learning of arithmetics in the octal number system.

In the following, the benefits and drawbacks, as well as further ideas for adaptive learning environments will be discussed.

4.1. Evaluating an EEG-Based Learning Environment

Commonly, if neural signals (like EEG) are used to estimate a user's mental state or a user's intention, the performance of the prediction method is assessed in terms of accuracy, correlation or other metrics (Spüler et al., 2015) that try to quantify how well the prediction model is working. While we have also evaluated the prediction performance of our model in a previous publication (Spüler et al., 2016), this kind of assessment is no more feasible in the here presented scenario with an online learning environment.

The reason for this is the lack of an objective measure for the user's workload. For the creation of the prediction model we used EEG data from a task that all subjects were able to

do fluently (addition in decimal system). As no learning effects are expected in this case, the difficulty of the task (measured by Q) was used as a subjective measure of expected workload. For the online learning environment, the relationship between task difficulty and workload does not hold anymore, as the learning environment induces learning effects. At the beginning, when the task is unknown to the user, even easy exercises will induce a high workload. After using the learning environment, the user may have mastered arithmetic tasks in the octal number system, and even exercises with moderate difficulty will result in a low workload. As the relationship between task difficulty and workload changes in the course of learning, the predicted task difficulty cannot be used for performance evaluation in an online learning environment.

As the task difficulty measured by the *Q*-value is the only objective measure we have, and the relationship to workload is invalidated in the online scenario, we have no means to objectively assess the prediction performance of our model. If the reader is interested in an assessment of the model performance, we refer to our previous publication (Spüler et al., 2016), where the model was evaluated on offline data.

Although we cannot assess the performance of the prediction model in the online scenario, we can evaluate the EEG-based learning environment with regards to its effect on the learning success. Learning success was defined as the difference in score between the pre- and post-tests, which were done by the subjects before and after using the learning environment and is a common measure for the evaluation of learning environments (Chi et al., 2011). As it is not important for the learner how accurate the workload-prediction works, but it is important how much the learning success can be improved, learning success is also the most user-centered metric.

Due to these facts, that other commonly used metrics are not applicable for the scenario of an online learning environment and that learning success is the most user-centered metric, we used the learning success as prime outcome measure for this study.

4.2. Proving the Concept of an EEG-based Learning Environment

As subjects using the EEG-based learning environment showed a significant learning effect, this work is a successful

proof-of-concept that an EEG-based learning environment works. So far, the use of EEG in a reading tutor has been investigated (Chang et al., 2013) or the cognitive and emotional state of the user is modeled to improve a tutoring system (Graesser et al., 2007), but to the best of our knowledge, this work presents for the first time a closed-loop system using an EEG-based workload adaptation in an arithmetic learning environment.

When comparing the experimental group using the EEG-based learning environment with the control group using the error-based learning environment, the learning success was higher for the experimental group, but the difference was not significant. As this study only compared an EEG-adaptive system to an error-adaptive system, it would be interesting for future studies to compare an EEG-based system against non-adaptive learning environments to see how it compares to those and if a significant performance difference can be achieved. Nevertheless, the results achieved with this study indicate that an EEG-based learning environment is an alternative to the state-of-the-art approach, but usability of the system is still an open issue.

Although, we aimed at a high usability for the presented system by using a cross-subject prediction model to omit a subject-specific calibration phase, the use of gel-based EEG still needs some time and effort to prepare, thereby making it impractical to use an EEG-based learning environment on a wide basis. With the recent development of dry EEG electrodes it was shown that dry electrodes provide lower signal-to-noise ratio than gel-based electrodes, but signal quality is still good enough for brain-computer interface control (Spüler, 2017). While usability of such a system could be improved by dry EEG electrodes, the increased cost can likely not be justified when using EEG-based learning environments on a broad population (e.g., in a class room). However, this technique could prove helpful for special cases in which the user suffers from learning disability or other problems. In the presented study, one subject of the experimental group suffered from test anxiety and reported to feel very comfortable using the EEG-based learning environment, as the system turned down the difficulty every time the subject was closed to feeling overwhelmed, thereby providing a good learning experience.

4.3. Improving the EEG-Based Learning Environment

As this work should merely serve as a proof-of-concept to show the feasibility of an EEG-based learning environment, it should be discussed how such a system could be potentially improved.

One possible improvement could be made by not only detecting workload, but also other cognitive properties like vigilance, attention or engagement. As we already mentioned in Spüler et al. (2016), arithmetic exercises with a very high difficulty show similar EEG patterns as exercises with very low difficulty, which is likely due to a disengagement effect (Chanel et al., 2008) where subjects do not even try to solve the task. Therefore, taking additional parameters into account when adapting the learning material, seems to be advisable. Besides workload, also vigilance and engagement are important factors for solving a

task correctly and learn efficiently. Decreasing vigilance is often specified as the decline in attention-requiring performance over an extended period of time. Furthermore, vigilance increases steeper in the context of difficult, compared to easy tasks. Tiwari et al. (2009) describe the interrelationship of vigilance and workload. An increasing workload is accompanied by a vigilance decrease. Engagement and workload increased as a function of task difficulty during learning and memory tasks (Berka et al., 2007). The results from previous studies (Oken et al., 2006; Berka et al., 2007; Tiwari et al., 2009) showed, the detection of various vigilance and engagement states is possible. Being able to detect these cognitive states based on the EEG could thereby improve the workload prediction and also add another layer to the learning environment where not only the learning material is presented, but also other cues could be integrated to motivate the learner when signs of disengagement or boredom are detected. Roy et al. (2013) have also shown that there is an interaction between fatigue and workload, which potentially decreases classification performance over time. As Roy et al. (2016b) have also shown that ERPs are a more robust indicator of workload than power spectral features, ERPs could be used as an additional feature to predict workload in future EEG-based learning environments. But, as ERP-based workload detection is stimulus-dependent, this approach is less flexible than using the power spectrum, which can be used for workload estimation independent of any stimuli.

5. CONCLUSION

In this paper, we presented an EEG-based learning environment, which unobtrusively detects the user's workload online and adapts the learning material accordingly to support each learner optimally. In a first study we collected EEG of subjects solving arithmetic exercises in the decimal number system. Based on this data a cross-subject prediction model was created that allows to predict the workload of other subjects without the need for a subject-specific calibration phase. This prediction model was used online to predict the workload of subjects learning arithmetic in the octal number system and the learning material was adapted promptly to keep the learner's workload in an optimal level. Utilizing the EEG-based learning environment showed a learning success similar to using a state-of-the-art system, which suggests the feasibility of an EEG-based learning environment.

AUTHOR CONTRIBUTIONS

CW and MS conceived and designed the studies. CW programmed the online learning environment. CW collected the data. CW and MS analyzed the data. CW and MS wrote the article, partly based on the doctoral dissertation by CW, and aided by MB, WR, and PG.

FUNDING

REFERENCES

Askew, M. (2015). Numeracy for the 21st century: a commentary. *ZDM* 47, 707–712. doi: 10.1007/s11858-015-0709-0

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78(Suppl. 1), B231–B244.

Brouwer, A.-M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Calder, N. (2015). Student wonderings: Scaffolding student understanding within student-centred inquiry learning. *ZDM* 47, 1121–1131. doi: 10.1007/s11858-015-0734-z

Causse, M., Fabre, E., Giraudet, L., Gonzalez, M., and Peysakhovich, V. (2015). Eeg/erp as a measure of mental workload in a simple piloting task. *Proc. Manufac.* 3, 5230–5236. doi: 10.1016/j.promfg.2015.07.594

Chanel, G., Rebetez, C., Bétrancourt, M., and Pun, T. (2008). "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games," in *Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era* (Tampere: ACM), 13–17.

Chang, K.-m., Nelson, J., Pant, U., and Mostow, J. (2013). Toward exploiting EEG input in a reading tutor. *Int. J. Artif. Intell. Educ.* 22, 19–38. doi: 10.3233/JAI-130033

Chi, M., VanLehn, K., Litman, D., and Jordan, P. (2011). An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *Int. J. Artif. Intell. Educ.* 21, 83–113. doi: 10.3233/JAI-2011-014

Corbett, A. (2001). "Cognitive computer tutors: solving the two-sigma problem," in *Proceedings of the 8th International Conference on User Modeling* (Sonthofen), 137–147.

Cover, T., and Thomas, J. (2006). *Elements of Information Theory*. Hoboken, NJ: Wiley-Interscience.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educ. Psychol. Rev.* 26, 197–223. doi: 10.1007/s10648-013-9246-y

Dowker, A. (2004). *What works for Children with Mathematical Difficulties?* Vol. 554. Nottingham: DfES Publications.

Gerjets, P., Scheiter, K., and Cierniak, G. (2009). The scientific value of cognitive load theory: a research agenda based on the structuralist view of theories. *Educ. Psychol. Rev.* 21, 43–54. doi: 10.1007/s10648-008-9096-1

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Front. Neurosci.* 8:385. doi: 10.3389/fnins.2014.00385

Gerjets, P. H., and Hesse, F. W. (2004). When are powerful learning environments effective? the role of learner activities and of students' conceptions of educational technology. *Int. J. Educ. Res.* 41, 445–465. doi: 10.1016/j.ijer.2005.08.011

Gevins, A., Smith, M., McEvoy, L., and Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cereb. Cortex* 7, 374–385. doi: 10.1093/cercor/7.4.374

Gevins, A., and Smith, M. E. (2006). "Electroencephalography (EEG) in neuroergonomics," in *Neuroergonomics: The Brain at Work*, eds R. Parasuraman, and M. Rizzo (New York, NY: Oxford University Press) 15–31.

Graesser, A., Chipman, P., King, B., McDaniel, B., and D'Mello, S. (2007). Emotions and learning with auto tutor. *Front. Artif. Intell. Appli.* 158:569.

Graesser, A., and McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educ. Psychol.* 45, 234–244. doi: 10.1080/00461520.2010.515933

Harmony, T., Fernández, T., Silva, J., Bosch, J., Valdés, P., Fernández-Bouzas, A., et al. (1999). Do specific EEG frequencies indicate different processes during mental calculation? *Neurosci. Lett.* 266, 25–28. doi: 10.1016/S0304-3940(99)00244-X

Karagiannakis, G. N., and Cooreman, A. (2014). "Focused MLD intervention based on the classification of MLD subtypes," in *The Routledge International Handbook of Dyscalculia and Mathematical Learning Difficulties*, ed S. Chinn (New York, NY: Routledge), 265.

Käser, T., Baschera, G.-M., Busetto, A. G., Klingler, S., Solenthaler, B., Buhmann, J. M., and Gross, M. (2013). Towards a Framework for Modelling Engagement Dynamics in Multiple Learning Domains. *Int. J. Artif. Intell. Educ.* 22, 59–83. doi: 10.3233/JAI-130026

Kirschner, P., and Gerjets, P. (2006). Instructional design for effective and enjoyable computer-supported learning. *Comput. Hum. Behav.* 22, 1–8. doi: 10.1016/j.chb.2005.01.004

Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Müller, K.-R., Curio, G., et al. (2007). Improving human performance in a real operating environment through real-time mental workload detection. *Toward Brain-Computer Interfacing*, eds G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller (Cambridge, MA: MIT press), 409–422.

Kramer, A. F., Trejo, L. J., and Humphrey, D. (1995). Assessment of mental workload with task-irrelevant auditory probes. *Biol. Psychol.* 40, 83–100. doi: 10.1016/0301-0511(95)05108-2

Murata, A. (2005). An attempt to evaluate mental workload using wavelet transform of EEG. *Hum. Fact.* 47, 498–508. doi: 10.1518/001872005774860096

Oken, B. S., Salinsky, M. C., and Elsas, S. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clin. Neurophysiol.* 117, 1885–1901. doi: 10.1016/j.clinph.2006.01.017

Pesonen, M., Hämäläinen, H., and Krause, C. M. (2007). Brain oscillatory 4–30 hz responses during a visual n-back memory task with varying memory load. *Brain Res.* 1138, 171–177. doi: 10.1016/j.brainres.2006.12.076

Richards, K. C., Enderlin, C. A., Beck, C., McSweeney, J. C., Jones, T. C., and Roberson, P. K. (2007). Tailored biobehavioral interventions: a literature review and synthesis. *Res. Theory Nurs. Pract.* 21, 271–285. doi: 10.1891/088971807782428029

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2013). "Mental fatigue and working memory load estimation: interaction and implications for EEG-based passive BCI," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (Osaka: IEEE), 6607–6610.

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2016a). Efficient workload classification based on ignored auditory probes: A proof of concept. *Front. Hum. Neurosci.* 10:519. doi: 10.3389/fnhum.2016.00519

Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016b). Efficient mental workload estimation using task-independent eeg features. *J. Neural Eng.* 13:026019. doi: 10.1088/1741-2560/13/2/026019

Scharinger, C., Soutschek, A., Schubert, T., and Gerjets, P. (2017). Comparison of the working memory load in n-back and working memory span tasks by means of eeg frequency band power and p300 amplitude. *Front. Hum. Neurosci.* 11:6. doi: 10.3389/fnhum.2017.00006

Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., and Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clin. Neurophysiol.* 118, 98–104. doi: 10.1016/j.clinph.2006.09.003

Spüler, M. (2017). A high-speed brain-computer interface (BCI) using dry eeg electrodes. *PLoS ONE* 12:e0172400. doi: 10.1371/journal.pone.0172400

Spüler, M., Sarasola-Sanz, A., Birbaumer, N., Rosenstiel, W., and Ramos-Murguialday, A. (2015). "Comparing metrics to evaluate performance of regression methods for decoding of neural signals," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (Milano: IEEE), 1083–1086.

Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., and Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 267–278. doi: 10.1007/s11858-015-0754-8

Sweller, J., Van Merriënboer, J. J. G., and Paas, F. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

Thomas, H. B. G. (1963). Communication theory and the constellation hypothesis of calculation. *Q. J. Exp. Psychol.* 15, 173–191. doi: 10.1080/17470216308416323

Tiwari, T., Singh, A. L., and Singh, I. L. (2009). Task demand and workload: effects on vigilance performance and stress. *J. Indian Acad. Appl. Psychol.* 35, 265–275.

Ullsperger, P., Freude, G., and Erdmann, U. (2001). Auditory probe sensitivity to mental workload changes–an event-related potential study. *Int. J. Psychophysiol.* 40, 201–209. doi: 10.1016/S0167-8760(00)00188-4

Walter, C., Wolter, P., Rosenstiel, W., Bogdan, M., and Spüler, M. (2014). "Towards Cross-Subject Workload Prediction," in *Proceedings of the 6th International Brain-Computer Interface Conference* (Graz).

Wang, Z., Hope, R. M., Wang, Z., Ji, Q., and Gray, W. D. (2012). Cross-subject workload classification with a hierarchical bayes model. *Neuroimage* 59, 64–69. doi: 10.1016/j.neuroimage.2011.07.094

Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Check for updates

Dynamic Threshold Selection for a Biocybernetic Loop in an Adaptive Video Game Context

Elise Labonte-Lemoyne*, François Courtemanche, Victoire Louis, Marc Fredette, Sylvain Sénécal and Pierre-Majorique Léger

Tech3Lab, HEC Montréal, Université de Montréal, Montreal, QC, Canada

Passive Brain-Computer interfaces (pBCIs) are a human-computer communication tool where the computer can detect from neurophysiological signals the current mental or emotional state of the user. The system can then adjust itself to guide the user toward a desired state. One challenge facing developers of pBCIs is that the system's parameters are generally set at the onset of the interaction and remain stable throughout, not adapting to potential changes over time such as fatigue. The goal of this paper is to investigate the improvement of pBCIs with settings adjusted according to the information provided by a second neurophysiological signal. With the use of a second signal, making the system a hybrid pBCI, those parameters can be continuously adjusted with dynamic thresholding to respond to variations such as fatigue or learning. In this experiment, we hypothesize that the adaptive system with dynamic thresholding will improve perceived game experience and objective game performance compared to two other conditions: an adaptive system with single primary signal biocybernetic loop and a control non-adaptive game. A within-subject experiment was conducted with 16 participants using three versions of the game Tetris. Each participant plays 15 min of Tetris under three experimental conditions. The control condition is the traditional game of Tetris with a progressive increase in speed. The second condition is a cognitive load only biocybernetic loop with the parameters presented in Ewing et al. (2016). The third condition is our proposed biocybernetic loop using dynamic threshold selection. Electroencephalography was used as the primary signal and automatic facial expression analysis as the secondary signal. Our results show that, contrary to our expectations, the adaptive systems did not improve the participants' experience as participants had more negative affect from the BCI conditions than in the control condition. We endeavored to develop a system that improved upon the authentic version of the Tetris game, however, our proposed adaptive system neither improved players' perceived experience, nor their objective performance. Nevertheless, this experience can inform developers of hybrid passive BCIs on a novel way to employ various neurophysiological features simultaneously.

Keywords: dynamic adaptation, passive brain computer interface, hybrid brain computer interface, BCI, pBCI, hBCI, EEG, tetris

INTRODUCTION

We have become quite adept at communicating with our computer systems and having them do what we wish. However, this type of communication lacks the subtleties that accompany human-human interactions through non-verbal signals. A promising new mode of communicating with technology has emerged that could help our computer systems understand more understated information about us, and that is passive brain-computer interfaces (pBCIs) (Zander and Kothe, 2011). This paper presents suggested improvements to the current state of pBCI development.

With pBCIs, the user's state is communicated to the computer without any conscious effort from the user and the system takes that into account to adapt itself to ultimately lead the user to an optimal state. Passive BCIs can bring about many different types of adaptations, such as modifying the level of difficulty of a game to ensure it is as stimulating as possible for players (Van De Laar et al., 2013), set off an alarm when truck drivers are too tired to be safe on the road (Hajinoroozi et al., 2016), or disengaging the autopilot to prevent airline pilots from getting so bored that they fail to notice important warning signs (Pope et al., 1995). Passive BCIs, as opposed to active BCIs, are mostly being developed for average users to improve their daily experiences rather than to replace other input devices such as using eye tracking instead of a mouse (Frisoli et al., 2012). As mentioned by Robert Jacob, "active BCIs have the potential to bring a great improvement to the lives of a small number of people, while passive BCIs can bring small improvements to the lives of a great number of people" (Jacob, 2017, p. 14).

While multiple research groups have been building functional pBCI systems (Prinzel et al., 2000; Scerbo et al., 2003; Berka et al., 2004; Lin et al., 2006; Ewing et al., 2016), many technical challenges remain to be addressed (Allanson and Fairclough, 2004; Fairclough, 2011; Brouwer et al., 2015), such as personalization. As biosignals vary greatly from one person to another, personalization is required for systems to function properly (Makeig et al., 2012). Developers will often use personalization tasks to individually tailor thresholds before the use of a system (Johnson et al., 2011). Their participants will thus have to be recorded twice, once to measure their responses to a task designed to induce a given state and a second time once the system has been personalized to actually use it. This requires time and resources and results in highly personalized systems that are, nonetheless, not responsive to changes over time, such as fatigue. This is the challenge we attempt to address within this paper.

A solution that has been proposed to many such BCI problems is the combination of multiple neurophysiological signals (hybrid BCI) to improve reliability, proficiency, bandwidth, convenience, and utility (see Banville and Falk, 2016 for a review; Nijholt et al., 2011). The type of system presented in this article can incorporate both neurophysiological and psychophysiological measures into the same system (Pfurtscheller et al., 2010; Zander et al., 2010; Banville and Falk, 2016). It is also commonly known as multimodal BCI. We look to the hybrid BCI to personalize our system in real-time and diminish the time and resources that are required before a user can begin using a system for the first time. While most hybrid BCIs use a second signal to improve the robustness of the system, to our knowledge, no other group has used a secondary neurophysiological signal to adjust the parameters of the system's adaptation to the first signal. Basically, the second signal would be used as a successfulness indicator of the adaptation, which is based on the first signal, leading it to be personalized. To our knowledge, this paper presents a first attempt to do so.

Our approach takes a slightly different view of pBCIs as most systems intend the subject to be unaware of the adaptations taking place. We propose instead that subjects do not need to be completely conscious of the changes, but that they are probably, on some level, aware that changes are taking place. By measuring the impact of these changes with a secondary, affective signal, we intend to measure the user's reaction to the adaptation in real time. This should inform the system whether the adaptation improved their experience (and/or performance) or not, thus telling it whether the adaptation parameters were correct or not. Accordingly, we propose a new approach for dynamic threshold selection that performs a continuous calibration. The biocybernetic loop adapts to one physiologically inferred metric, in this case, cognitive load, using electroencephalography (EEG), while the thresholds are continuously adjusted based on a second metric, in this case, emotional valence, inferred using automatic facial expression analysis.

The goal of this paper is to evaluate if dynamic thresholding with a secondary neurophysiological signal can continuously personalize the adaptive system. We hypothesize that the adaptive system with dynamic thresholding will improve perceived game experience and objective performance.

METHODS

Participants

This study was carried out in accordance with the Research Ethics Board of our institution and with written informed consent from all subjects. Participants were students at local universities. There were 16 in total (6 women). Average age was 27.75 (SD = 8.55). Participants had to be over 18 and have no known health issues. They received a \$30 gift card as a compensation.

Initial Baselining

We used a modified vanilla baseline task (Jennings et al., 1992) as a baseline calibration task. This was chosen because it is an emotionally neutral task with similar visual characteristics as the Tetris game, i.e., color and movement.

Experimental Design

As Ewing et al. (2016) underlined, game periods need to be longer than 5 min to get an accurate representation of the system's adaptation. Thus, participants played each condition for 15 min. The order of the three conditions was counterbalanced between participants. For the adaptive conditions, the system managed the game difficulty as in the Dynamic Difficulty Adjustment framework (Hunicke and Chapman, 2004).

Game

Tetris is a game that is often used in research as it is easily adaptable and because gameplay is generally understood by most. In this instance, Tetris was also advantageous as there is no emotional component to the gameplay. This prevented us from having to tease out the effect of gameplay from the effects of the adaptive system on the measured valence, the latter being more relevant to our study.

To easily control game parameters, we developed an in-house version of Tetris. The game window consisted of a rectangle of 10 by 20 squares. Players controlled the game with a Logitech F310 gamepad (Logitech, Lausanne). They could move the pieces laterally with the left and right arrows and rotate the pieces with the "A" and "X" buttons. No other manipulation was allowed.

For each condition, the initial speed was 400 ms per line. The slowest allowed speed was 1,500 ms per line and the fastest allowed speed was 100 ms per line. Speed variations between levels were of 100 ms at a time. These parameters were chosen after pilot testing was conducted. The only parameter that was atypical was the absence of the control that allows the user to manually make the blocks fall faster. The ability to force the blocks down which might prevent the conditions from being uniform between subjects, thus it was not implemented in our game.

Condition 1

In the control condition, there was no adaptation taking place. The speed increased by one step (speed levels are defined in the next section) every time four rows were successfully cleared. This condition was chosen to resemble a traditional game of Tetris as closely as possible.

Condition 2

The second condition was developed to model as closely as possible to Ewing et al. (2016) conservative condition. The system adapts to the user state as measured by the EEG. Any 100% or two-fold increase/decrease from the baseline value (e.g., for a starting baseline value of 4, a 100% increase would be a value of 8, a 100% decrease would be a value of 2) or more from the benchmarking task triggers a speed change (see **Figure 1a**). Ewing et al. (2016) employed the upper alpha (10.5–13 Hz) in the P4 location and theta bands (4–8 Hz) in the Fz location in the following manner: an increase of theta accompanied by an increase in alpha (both should be more than 100% baseline) corresponds to the user being in a state of boredom and the speed is increased; an increase of theta accompanied by a decrease in alpha (both 100% or more) corresponds to the user being in a state of overload and the speed is decreased. Similarly, our system adaptations of speed are triggered by 100% or more changes to the baseline level of the upper alpha band (10.5–13 Hz). This band was chosen from Ewing et al. (2016) following pilot testing.

Condition 3

The third condition is similar to the second condition (see **Figures 1a,b** for a side by side comparison) with the addition of adaptive thresholds. The thresholds adapt to the emotional valence of the subject, i.e., the extent to which an emotion is positive (Lang et al., 1995). Valence is measured with the use of facial expression analysis, which is explained in the next section. The emotional valence is measured before and after each speed adaptation over 2 s. If the after-before delta is positive, the subject is deemed satisfied with the adaptation, the threshold value is considered appropriate and is maintained. If the delta is negative, the subject is deemed dissatisfied with the adaptation and the threshold value is modified as follows. There are two threshold values, the lower threshold which, when crossed, leads to an increase in speed, and the upper threshold, which, when crossed leads to a decrease in speed. When the subject is deemed dissatisfied with the adaptation, only the threshold that was crossed is changed. The thresholds move away from the middle, i.e., the upper threshold increases, or the lower threshold decreases, leading to a less severe adaptation criteria (see **Figure 2**). For example, if the upper threshold value was 1, and the baseline alpha value was 4, when the current alpha value reaches past 8 for a given window (a 100% increase), it has crossed the threshold. Thus, the decision-rule determines that the game speed must slow down. At the moment where the speed is modified, the delta valence is measured. The valence value from before the speed change is subtracted from the valence value after the speed change to see if the change is received positively or negatively. For example, if the valence before the speed change was 2 and the valence after the speed change is 1, the delta valence is negative (-1). The threshold value from our previous example, a value of 1 for the upper threshold, should be modified. It will move away from the center; therefore, the upper threshold value will increase to 1.1. In the next window, for the speed to change, the alpha value will need to reach 8.4 (an increase of 110% of the baseline value of 4). In addition, when, for a given 10 second epoch, no adaptation takes place while delta valence is negative, both thresholds are considered too loose and are brought closer to the middle, i.e., the upper threshold decreases, and the lower threshold increases, leading to a more severe adaptation criterion.

Measurement Tools and Processing

EEG Recording and Processing

The EEG was recorded from 32 Ag-AgCl preamplified electrodes mounted on the actiCap and with a brainAmp amplifier (Brainvision, Morrisville). The recording reference was FCz and the acquisition rate was 500 Hz. All the EEG processing was performed in the NeuroRT software (Mensia, Renne). The following steps were performed in this order for the online preprocessing of the data: down-sampling to 256 Hz, bandpass filtering with an infinite impulse response filter at 1–50 Hz, notch filtering at 60 Hz, blink removal through blind source separation, re-referencing to the common average reference, and artifact detection by computing the Riemannian distance between the covariance matrix and the online mean. The data was then filtered to keep only the upper alpha band (10.5–13 Hz) and squared. Only data from P3 and P4 was kept in accordance with Ewing (Ewing et al., 2016). Finally, each 0.5 s epoch was divided by the average of the 1-min calibration period to normalize the data. This value, which ranged from 0.008 to 6.87, was then sent to the adaptation system to be used in the decision rule-based process.

FIGURE 1 | (a) Flowchart of the pBCI in Condition 2. The game speed adapts according to changes in the EEG from baseline. In this condition the threshold at which changes in the EEG register as sufficient to trigger a change in speed is fixed at 100%. **(b)** Flowchart of the pBCI with adaptive thresholds in Condition 3. The game speed adapts according to changes in the EEG from baseline. In this condition, the threshold at which changes in the EEG register as sufficient to trigger a change in speed is continually adjusted according to the valence changes surrounding the previous speed changes.

Facial Expression Analysis

Emotional valence was obtained through the use of facial expression analysis performed in real time with the FaceReader software (Noldus, Wageningen). Automatic facial expression analysis uses the Facial Action Coding System (Ekman et al., 1980) that has traditionally been employed by human experts who manually code video playback for facial expressions. Specific combinations of muscle contractions are associated to certain emotions. Computer image recognition is now able to detect these same muscle contraction combinations and code the user's emotions in real time (Bartlett et al., 1999). We used the Facereader valence ratio which is calculated as the intensity of positive emotion minus the intensity of the negative expression with the highest intensity. The software maps the facial muscles
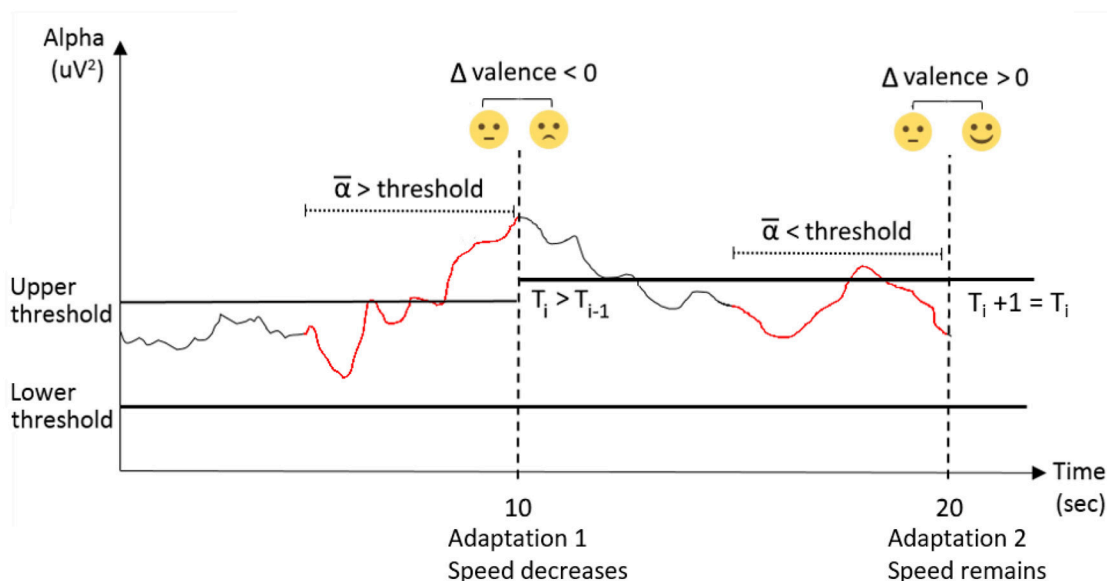
FIGURE 2 | Illustration of the dynamic thresholds adaptation in four steps (adaptation pacing = 10 s): (1) The average of alpha over the 5 s (in red) preceding adaptation 1 is higher than the upper threshold; (2) The system therefore choose to decrease the game speed at adaptation 1; (3) The difference in valence before (average of 2 s) and after (average of 2 s) adaptation 1 is negative; (4) The system qualifies adaptation 1 as a fail and raises the upper threshold in order to be less sensitive during adaptation 2. T_i refers to the threshold value at the offset of a given adaptation (i).

of the user and derives valence at a rate of 30 Hz. It streamed to the adaptive system that used the 2 s before and the 2 s after an adaptation to calculate the change in valence caused by the adaptation. A positive change tells the system that the adaptation was appropriate, meaning that the threshold parameters were correct, while a negative change tells the system that the adaptation was wrong and that the thresholds should be modified.

Game Performance

The behavior of each system (Conditions 1–3) was logged in order to compare them and understand what drove user experience. The game speed was controlled as the duration (ms) it took for a block to move down one line. This was converted *post-hoc* into levels. The smaller the level, the slower the game and there was 100 ms difference between levels (i.e., level 1 = 1,500 ms, level 15 = 100 ms). The number of "game deaths" per subject and the score of the subject at the end of each condition were also logged (**Table 1**). Participants scored 100 points each time they cleared a line, 300 points for two lines, 500 points for three lines, and 800 points for four lines.

Perceived Measures

In addition to game performance, we also measured the perception of the participant at the end of each experimental condition. We used the In-game and Post-game Game Experience Questionnaire (GEQ), a shortened version of the core questionnaire which was developed to assess multiple iterations of an experiment (IJsselsteijn et al., 2008). These questionnaires respectively have 14 and 17 items that lead to scores on the following dimensions: competence, flow, annoyance, challenge,

TABLE 1 | High and low speed comparisons for each experimental condition.

	Condition 1	Condition 2	Condition 3
Alpha	NA	Low speed = 0.65 High speed = 0.82 $p = 0.21$	Low speed = 0.69 High speed = 0.86 $p = 0.03$
Valence	NA	Low speed = 0.03 High speed = 0.14 $p = 0.21$	Low speed = 0.03 High speed = 0.13 $p = 0.11$
Average game level	Low speed = 12.37 High speed = 12.69 $p = 0.39$	Low speed = 3.94 High speed = 7.69 $p = 0.05$	Low speed = 4.41 High speed = 10.84 $p < 0.001$
Game deaths	Low speed = 3.6 High speed = 3.9 $p = 0.84$	Low speed = 0.3 High speed = 1.3 $p = 0.72$	Low speed = 0.14 High speed = 1.3 $p = 0.09$
Score	Low speed = 4,186 High speed = 5,367 $p = 0.35$	Low speed = 2,043 High speed = 2,989 $p = 0.26$	low speed=1957 high speed=956 $p = 0.008$

positive, and negative affect are the 6 dimensions of the In-game GEQ and positive experience, negative experience, tiredness, and returning to reality are the 3 dimensions of the pots-game questionnaire. Each item has a 5-point scale ranging from 0 (Don't get the feeling described in the statement at all) to 4 (Extremely get the feeling described in the statement).

Statistical Analysis

Diverse statistical tests have been performed between the three conditions (control, EEG, adaptive thresholds) and between clusters of participants afterwards. In order to compare conditions (GEQ, total scores, average game level, valence, alpha

power), we used Wilcoxon tests to compare the 3 conditions and Mann–Whitney tests to compare the clusters of participants. Both these tests are non-parametric as is appropriate considering the sample size. Also, a Kolmogorov–Smirnov test was used to test for normality and found only one variable to be normally distributed, the alpha power in conditions B and C ($D = 0{,}12$, $p = 0.15$; $D = 0.16$, $p = 0.15$) further justifying the use of non-parametric tests. We used Poisson regressions with random effects to account for repeated measures from the same participant to compare total number of deaths and total number of level changes between the 3 conditions.

RESULTS

As can be seen in **Figure 3**, when compared to the control condition (Condition 1), the number of game deaths was significantly lesser in the EEG only BCI (Condition 2), and in the adaptive thresholds condition (Condition 3). The score, when compared with the control condition, was lesser in both BCI conditions. The average game level (seen in **Figure 4**) in each condition was 12.53 for control (Condition 1), 6.18 for EEG only (Condition 2), and 8.92 for adaptive thresholds (Condition 3). All differences were significant at $p < 0.05$. These results show that both BCI conditions seem to have "over adapted" the game resulting on average in slower games (lower game level) and decreased player performance (score) compared to the control condition.

As for the GEQ, results and significant differences can be seen in **Figures 5**, **6**. Participants in the EEG condition (Condition 2) perceived the experience as less challenging (comparison 1 vs. 2, $p = 0.004$; comparison 2 vs. 3, $p = 0.059$), negatively charged (comparison 1 vs. 2 $p = 0.012$), and less conducive to flow (comparison 1 vs. 2, $p = 0.019$; comparison 2 vs. 3, $p = 0.086$) compared to the two other conditions. The fixed threshold condition (Condition 3) is perceived as less annoying (comparison 2 vs. 3 $p = 0.032$) and participants in this condition develop a higher perceived competence than in the two other conditions (comparison 1 vs. 2 $p = 0.008$; comparison 2 vs. 3 $p = 0.033$). Participants felt the experience in Condition 1 more negatively than in the two other conditions (comparison 1 vs. 2 $p = 0.031$; comparison 1 vs. 3 $p = 0.041$).

When explored further, our results appeared to show two different clusters of responses in Condition 3. Thus, we evaluated how many subjects spent 5 or more periods of 10 s at the fastest speed or at the slowest speed. Once these two groups were identified, we compared them with Wilcoxon tests to evaluate if they differed in their alpha power change from benchmark, valence, GEQ, number of game deaths, and game scores.

Results suggest that in Condition 3, 9 participants spent a large portion of their time at the slowest speed, while 10 participants spent most of their time at the fastest speed. Only three participants spent <25 s at either extremity. When compared, these two groups (slow speed group vs. high speed group, Condition 3) are statistically different for their average EEG alpha power change from baseline ($p = 0.01$), where the high-speed group has more alpha (high speed = 0.99; low speed = 0.67).

They are also statistically different in terms of valence values ($p = 0.02$; high speed = 0.01; low speed = 0.20).

We also evaluated if participants differed in Condition 2 to see if they were intrinsically different or if they were only different while using the dynamic threshold adaptive system. For those two groups (high and low-speed), their values did not differ in Condition 2 (alpha $p = 0.35$; valence $p = 0.23$; **Table 1**). Also, these groups differed on some dimensions of the GEQ. In Condition 2, they reported different annoyance perceptions (low speed = 2.78; high speed = 1.7; $p = 0.02$) and in Condition 3, challenge (low speed = 2.17; high speed = 3.65; $p = 0.03$) and competence perceptions differed (low speed = 3.22; high speed = 1.35; $p < 0.01$). Finally, the two groups significantly differed in game deaths in Condition 3 (high speed = 0; low speed = 19) and score in Conditions 1 and 2 (respectively, low speed = 4028.57, high speed = 5233.33; $p = 0.001$ and low speed = 2028.57, high speed = 2,700; $p = 0.001$).

DISCUSSION

Our results show that the adaptive systems did not improve the participants' experience while playing Tetris. Contrary to our expectation, participants had more negative affect from the fixed threshold and the dynamic thresholds conditions than in the traditional Tetris game (control condition). In the EEG only condition, similarly to Ewing et al. (2016), participants had fewer game deaths, a lower score, and a lower average speed; showing that the system slowed down the game, leading to the participants feeling more confident, but less challenged and ultimately feeling more negative affect.

If we examine the adaptive threshold condition by separating the participants according to how much time they spent at each extreme level, we can see that participants have significantly different alpha and valence. However, they did not differ in EEG only condition, showing that they are not fundamentally different in their neurophysiology, but that their neurophysiological responses to the use of an adaptive system were different. Obviously, the high-speed group found the challenge higher and their competence lower as they were generally unable to keep up with the game and died many more times than the low speed group.

One potential explanation of our counterintuitive results could be found in Csikszentmihalyi's flow theory (Nakamura, 2014). This specific flow theory postulates that an optimal positive valence experience emerges in an autotelic context when one's competence is in alignment with a given challenge. In self-motivating contexts, such as games, the theory predicts that when the challenge is above one's competency, a state of anxiety and frustration is likely to emerge. As the competency of the player evolves in addressing this challenge, the theory suggests that the individual is likely to move back toward a positively valenced equilibrium state. In our study, it is possible that the parameters to capture the change in the valence might not have been optimized to account for the time it takes for the player to come back to this equilibrium. This would suggest that future research needs to not only personalize the values that trigger

FIGURE 3 | Average number of game deaths per minute and average score per minute per experimental condition. Error bars represent standard deviation. *$p < 0.05$.

FIGURE 4 | Average game level for each experimental condition. Each line is the game level of a study participant over time in milliseconds. A higher game level means that the speed at which the blocs descend increases, thus diminishing the time available to position the bloc properly and increasing difficulty. In Condition 1, participants began at level 12 and the level increased progressively. In Condition 2, for most participants the level decreased over time. In Condition 3, a split appears, where some participants see their level increase irreversibly, while other participants have a result more similar to Condition 2. In a functioning system, there would be a lot of variations early on and after some time the levels would stabilize and adjust linearly. For Condition 3, we present only results for which the adaptive mechanism leads to thresholds that were mathematically attainable. This issue is discussed in further details in the next section.

FIGURE 5 | Average answers to the in-game portion of the Game Experience Questionnaire (GEQ) showing the results for each dimension measured. Error bars represent standard deviation. *$p < 0.05$.

FIGURE 6 | Average answers for the post-game portion of the Game Experience Questionnaire (GEQ) results showing the results for each dimension measured. Error bars represent standard deviation. *$p < 0.05$.

adaptation, but also needs to personalize the adaption window duration. This could be done through a more complex baseline task that would be able to capture this idiosyncratic latency in players.

With both these personalizations in place (threshold and adaptation time), we could hope to reach both subject independence and context independence. Subject independence and context independence are the big challenges currently facing passive BCIs in their quest for use in everyday life (BNCI Horizon 2020, 2015). The hope is to have a system that can be simply put on the user's head and work straight away in a variety of contexts without the need for lengthy calibration

tasks. Our goal was to develop a system that continuously adjusts its action triggering threshold rules which would have been a step in the direction of subject independence, or at the least, a system that is subject dependent and context independent. While our system needs more testing and calibration to evaluate its feasibility, we trust that continuous adaptation and the use of multimodal neurophysiological signals hold the key to those challenges. Other researchers should keep in mind that the use of valence-based tools may differ in the context of games with more emotionally charged objects and not extrapolate our results to those contexts. Future works will explore more comprehensive EEG data than alpha alone, and various window durations to see if the affective response to the speed adaptation appears after a delay, or if it may simply take some time for the user to get comfortable at a given level thus the speed changes should happen less frequently. In addition, as speed changes in quick succession are not expected when playing Tetris, perhaps novelty played a role in the user's dissatisfaction, thus a variety of adaptation windows could be tested.

While evaluating a variety of neurophysiological data sources and time windows could lead to a solution, other possible avenues of exploration have emerged. As mentioned in the literature, machine learning would greatly improve this type of system (Scherer et al., 2013; Brouwer et al., 2015; Ewing et al., 2016). On one side, recent advances in deep learning techniques could help enhance the accuracy of cognitive state inference models. Deep learning models are well suited to find complex hierarchical patterns in high dimensionality data such as neurophysiological data. Research using such approaches for psychophysiological inference (Martinez et al., 2013) or EEG signal decoding (Schirrmeister et al., 2017) show promising results. On the side of the interaction loop, we would recommend the use of reinforcement learning in the case of hybrid BCIs. Reinforcement learning uses a problem space to determine where

it is in relation to the optimal location. The system adapts within this problem space and gets rewarded as it gets closer to the goal. In the case of our proposed biocybernetic loop, the problem space would be defined as a grid of difficulty by neurophysiological state and the reward could be determined by a second neurophysiological signal. Consider the system presented in this paper. As a reinforcement learning system, the problem space in which the model would evolve and try to find an optimal path would be composed of the two dimensions of cognitive load and current speed of the game. Each action undertaken by the model would be rewarded by the local changes in the subject's valence following the adaptation. Doing so, the system should converge toward an optimal state of difficulty vs. cognitive load that is specific to each subject and that may change over time. In this approach, the adaptation thresholds would be replaced by the transitioning policy optimized over time that the system uses to decide which action to perform. We can imagine a similar system with a problem space composed of simulator challenge and vigilance level with a reward based on stress levels. The idea behind this is that an affective index could confirm or infirm if the action performed by the system improved the user experience in real time instead of having to validate with the user verbally after they test it for a while and adjusting the system manually afterwards. Our future work will explore this promising alternative.

We endeavored to develop a system that improved upon the authentic version of the Tetris game, however, our proposed adaptive system neither improved players' perceived experience, nor their objective performance. In practice, for condition 3, the way we modified the threshold values based on valence could eventually lead to values that were impossible to attain. For some participants, the thresholds crossed, leading to some lower thresholds being higher than upper thresholds and for some thresholds to be negative which is impossible to attain from the alpha ratio. Nevertheless, this experience can inform developers of hybrid passive BCIs on a novel way to merge various neurophysiological features. If the goal of a passive BCI system is to be widely adopted, it must be better than the current version of a similarly oriented system. Comparisons to other versions of a given BCI (Liu et al., 2009; Chanel et al., 2012) are very informative to better understand how parameters of the BCI influence the user-system relationship, however, this is only one part of the picture when it comes to the future adoption of passive BCIs in everyday lives. As mentioned by Nijholt et al. (2011), new types of human-computer interaction take years of research before they come into their own, thus all advances should be presented and discussed. Ultimately, the true test will be to compare any new system to the status quo in order to improve it.

AUTHOR CONTRIBUTIONS

EL-L and FC designed the study and developed the system. All authors contributed to the analysis and the writing.

REFERENCES

Allanson, J., and Fairclough, S. H. (2004). A research agenda for physiological computing. *Interact. Comput.* 16, 857–878. doi: 10.1016/j.intcom.2004.08.001

Banville, H., and Falk, T. H. (2016). Recent advances and open challenges in hybrid brain-computer interfacing: a technological review of non-invasive human research. *Brain Comput. Interfaces* 3, 9–46. doi: 10.1080/2326263X.2015.1134958

Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology* 36, 253–263. doi: 10.1017/S0048577299971664

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., et al. (2004). Real-Time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *Int. J. Hum. Comput. Interact.* 17, 151–170. doi: 10.1207/s15327590ijhc1702_3

BNCI Horizon 2020 (2015). Roadmap: the future in brain/neural-computer interaction. *Horizon 2020.* doi: 10.3217/978-3-85125-379-5

Brouwer, A. M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., and Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Front. Neurosci.* 9:136. doi: 10.3389/fnins.2015.00136

Chanel, G., Kivikangas, J. M., and Ravaja, N. (2012). Physiological compliance for social gaming analysis: cooperative versus competitive play. *Interact. Comput.* 24, 306–316. doi: 10.1016/j.intcom.2012.04.012

Ekman, P., Freisen, W., and Ancoli, S. (1980). Facial signs of emotional experience. *J. Pers. Soc. Psychol.* 39, 1125–1134. doi: 10.1037/h0077722

Ewing, K. C., Fairclough, S. H., and Gilleade, K. (2016). Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. *Front. Hum. Neurosci.* 10:223. doi: 10.3389/fnhum.2016.00223

Fairclough, S. (2011). "Physiological computing: interfacing with the human nervous system," in *Sensing Emotions,* eds J. Westerink, M. Krans, and M. Ouwerkerk (Dordrecht: Springer), 1–20.

Frisoli, A., Loconsole, C., Leonardis, D., Banno, F., Barsotti, M., Chisari, C., et al. (2012). A new Gaze-BCI-Driven control of an upper limb exoskeleton for rehabilitation in real-world tasks. *IEEE Trans. Syst. Man Cybern. Part C* 42, 1169–1179. doi: 10.1109/TSMCC.2012.2226444

Hajinoroozi, M., Mao, Z., Jung, T. P., Lin, C. T., and Huang, Y. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Process. Image Commun.* 47, 549–555. doi: 10.1016/j.image.2016.05.018

Hunicke, R., and Chapman, V. (2004). AI for dynamic difficulty adjustment in games. *Challenges Game Artif. Intell. AAAI* 91–96. doi: 10.1145/1178477.1178573

IJsselsteijn, W. A., de Kort, Y. A. W., and Poels, K. (2008). "The game experience questionnaire," in *Working Paper.*

Jacob, R. (2017). "Implicit user interfaces," in *Neuroadaptive Technology Conference* (Berlin).

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., and Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: vanilla or resting baseline. *Psychophysiology* 29, 742–750. doi: 10.1111/j.1469-8986.1992.tb02052.x

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., and Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biol. Psychol.* 87, 241–250. doi: 10.1016/j.biopsycho.2011.03.003

Lang, A., Dhillon, K., and Dong, Q. (1995). The effects of emotional arousal and valence on television viewers' cognitive capacity and memory. *J. Broadcast. Electron. Media* 39, 313–327.

Lin, C. T., Ko, L. W., Chung, I. F., Huang, T. Y., Chen, Y. C., Jung, T. P., et al. (2006). Adaptive EEG-based alertness estimation system by using ICA-based

fuzzy neural networks. *IEEE Trans. Circuits Syst. I Regul. Pap.* 53, 2469–2476. doi: 10.1109/TCSI.2006.884408

Liu, C., Agrawal, P., Sarkar, N., and Chen, S. (2009). Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *Int. J. Hum. Comput. Interact.* 25, 506–529. doi: 10.1080/10447310902963944

Makeig, S., Kothe, C., Mullen, T., Bigdely-Shamlo, N., Zhang, Z., and Kreutz-Delgado, K. (2012). Evolving signal processing for brain-computer interfaces. *Proc. IEEE* 100, 1567–1584. doi: 10.1109/JPROC.2012.2185009

Martinez, H. P., Bengio, Y., and Yannakakis, G. N. (2013). Learning deep physiological models of affect. *IEEE Comput. Intell. Mag.* 8, 20–33. doi: 10.1109/MCI.2013.2247823

Nakamura, J. (2014). "The concept of flow," in *Flow and the Foundations of Positive Psychology*, ed M. Csikszentmihalyi (Dordrecht: Springer), 239–263.

Nijholt, A., Allison, B. Z., and Jacob, R. J. K. (2011). "Brain-computer interaction: can multimodality help?," in *Proceedings 13th International Conference on Multimodal Interfaces (ICMI 2011)* (New York, NY), 35–39.

Pfurtscheller, G., Allison, B. Z., Brunner, C., Bauernfeind, G. G., Solis-Escalante, T., Scherer, R., et al. (2010). The hybrid BCI. *Front. Neurosci.* 4:30. doi: 10.3389/fnpro.2010.00003

Pope, A. T., Bogart, E. H., and Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biol. Psychol.* 40, 187–195. doi: 10.1016/0301-0511(95)05116-3

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., and Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive task allocation. *Int. J. Aviat. Psychol.* 10, 393–410. doi: 10.1207/S15327108IJAP1004_6

Scerbo, M. W., Freeman, F. G., and Mikulka, P. J. (2003). A brain-based system for adaptive automation. *Theor. Issues Ergon. Sci.* 4, 200–219. doi: 10.1080/1463922021000020891

Scherer, R., Faller, J., Balderas, D., Friedrich, E. V. C., Pröll, M., Allison, B., et al. (2013). Brain-computer interfacing: more than the sum of its parts. *Soft Comput.* 17, 317–331. doi: 10.1007/s00500-012-0895-4

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* 38, 5391–5420. doi: 10.1002/hbm.23730

Van De Laar, B., Gürkök, H., Plass-Oude Bos, D., Poel, M., and Nijholt, A. (2013). Experiencing BCI control in a popular computer game. *IEEE Trans. Comput. Intell. AI Games* 5, 176–184. doi: 10.1109/TCIAIG.2013.2253778

Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:25005. doi: 10.1088/1741-2560/8/2/025005

Zander, T. O., Kothe, C., Jatzev, S., and Gaertner, M. (2010). "Enhancing Human-Computer Interaction with input from active and passive Brain-Computer Interfaces," in *Brain-Computer Interfaces: Human-Computer Interaction Series*, eds D. S. Tan and A. Nijholt (London: Springer), 181–199.

Single-Trial EEG Analysis Predicts Memory Retrieval and Reveals Source-Dependent Differences

Eunho Noh[1], Kueida Liao[1], Matthew V. Mollison[2], Tim Curran[2] and Virginia R. de Sa[3]*

[1] Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA, United States,
[2] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, United States, [3] Department of Cognitive Science, University of California, San Diego, San Diego, CA, United States

We used pattern classifiers to extract features related to recognition memory retrieval from the temporal information in single-trial electroencephalography (EEG) data during attempted memory retrieval. Two-class classification was conducted on correctly remembered trials with accurate context (or source) judgments vs. correctly rejected trials. The average accuracy for datasets recorded in a single session was 61% while the average accuracy for datasets recorded in two separate sessions was 56%. To further understand the basis of the classifier's performance, two other pattern classifiers were trained on different pairs of behavioral conditions. The first of these was designed to use information related to remembering the item and the second to use information related to remembering the contextual information (or source) about the item. Mollison and Curran (2012) had earlier shown that subjects' familiarity judgments contributed to improved memory of spatial contextual information but not of extrinsic associated color information. These behavioral results were similarly reflected in the event-related potential (ERP) known as the FN400 (an early frontal effect relating to familiarity) which revealed differences between correct and incorrect context memories in the spatial but not color conditions. In our analyses we show that a classifier designed to distinguish between correct and incorrect context memories, more strongly involves early activity (400–500 ms) over the frontal channels for the location distinctions, than for the extrinsic color associations. In contrast, the classifier designed to classify memory for the item (without memory for the context), had more frontal channel involvement for the color associated experiments than for the spatial experiments. Taken together these results argue that location may be bound more tightly with the item than an extrinsic color association. The multivariate classification approach also showed that trial-by-trial variation in EEG corresponding to these ERP components were predictive of subjects' behavioral responses. Additionally, the multivariate classification approach enabled analysis of error conditions that did not have sufficient trials for standard ERP analyses. These results suggested that false alarms were primarily attributable to item memory (as opposed to memory of associated context), as commonly predicted, but with little previous corroborating EEG evidence.

Keywords: EEG, memory retrieval, old/new effect, multi-variate analysis, prediction

INTRODUCTION

Previous recognition memory studies have used electroencephalography (EEG) to identify neural substrates of recognition memory. The 'parietal old/new effect' is a positive-going event-related potential (ERP) typically observed in the parietal electrodes between 500 and 800 ms and typically left lateralized. It shows greater amplitude for the correctly recognized old (hits) compared to the new (correct rejections) test items. It has been found that this effect correlates with the amount of information retrieved from the study episode (Wilding and Rugg, 1996; Curran, 2000; Wilding, 2000; Rugg and Curran, 2007; Tsivilis et al., 2015); hence, it is understood as a neural correlate of recollection. The 'frontal old/new effect' (or the FN400) is a frontally distributed and negative-going ERP which peaks earlier around 400 ms. The FN400 is interpreted as a neural correlate of familiarity since it shows a more negative peak for less familiar items while it typically does not vary for different amounts of recollected context information (Curran, 2000; Rugg and Curran, 2007; Tsivilis et al., 2015).

Pattern classification methods have been recently applied to EEG data to reveal novel findings during encoding of episodic memory (Jafarpour et al., 2014; Noh et al., 2014a; Anderson et al., 2015; Ratcliff et al., 2016). In Noh et al. (2014a), the classifier was used as a discriminative dimensionality reduction method to project the high-dimensional EEG data onto a discriminative space. These projections revealed neural correlates of levels of encoding in the pre- and during-stimulus periods of the study phase. This multivariate analysis directly controls for the multiple comparison problem (MCP) by effectively reducing the number of test variables. A major advantage of this approach is that it is possible to compare the brain activity across conditions even when the trial count is low, provided that a sufficient number of classifier training trials are used to establish the initial hyperplane(s) (Noh and de Sa, 2014). Hence conditions that divide subtle behavioral differences can be readily compared. In ERP studies, these data are usually ignored or combined with other conditions to acquire reasonable ERPs for analysis. This may result in losing the ability to reveal the neural mechanisms underlying subtle behavioral differences.

Our study aims to create classifiers to discriminate between the correctly identified old/new trials during the recognition phase of episodic memory experiments on a single trial basis. We also utilize pattern classifiers as multivariate analysis tools to analyze the brain activity during retrieval of recognition memory using the time domain information of the EEG data. The EEG data were collected from three separate visual memory task experiments with extrinsic source information. Two types of source information were considered in these experiments. Spatial information (the location of the item) was of interest in Experiment 1 and extrinsic color information (the color of an external frame) was of interest in Experiment 2. In Experiment 3, both source types were considered. Data collected from these experiments were used to conduct multivariate analysis via pattern classifiers. The data used were previously collected by Mollison and Curran (2012). In the experiments, subjects were asked to remember items as well as the contextual source

information (side of the screen, or color of outlined box). In the test phase they were asked to indicate whether they believe they have seen the item before, and if so to give the associated source information as well as their confidence in that judgment by specifying whether they remember the source information, any other information, or whether the item is just familiar. Mollison and Curran (2012) found that even familiar judgments were associated with above chance source judgments and that the FN400 distinguished between the source-correct and source-incorrect responses only for the location-source information but not the box-color source information. In this work, we specifically train separate classifiers to extract information related to item memory (without correct source memory) and source memory (for correctly remembered items) to observe any source-dependent differences that the classifiers extract between the experiments with different source types.

The average projection values (or classifier scores) of the different source retrieval conditions and different subjective rating conditions are also compared to reveal the relationship between the different conditions and memory retrieval strength. Furthermore, data from the error conditions (incorrectly identified new trials, incorrectly rejected old trials) are projected onto the discriminative vector characterized by the different classifiers. The average projection values of these error trials are compared to those given by the other conditions and across the different projection directions.

MATERIALS AND METHODS

Electroencephalography for the current study was previously recorded in three separate visual memory task experiments (Mollison and Curran, 2012). All procedures were approved by the Institutional Review Board at the University of Colorado Boulder and were conducted in accordance with this approval. All participants gave written informed consent before the experiment.

Experiment 1
Participants
The subjects were right-handed University of Colorado undergraduate students (ages 18–28, mean = 21.4) who volunteered for paid participation ($15 per hour) or course credit (17 male, 13 female). All subjects were native English speakers and had normal or corrected-to-normal vision.

Experimental Paradigm and EEG Acquisition
The experiment was divided into four blocks consisting of a study and recognition phase. The stimuli were color images of physical objects, animals, and people. A total of 1297 images were selected from http://www.clipart.com, the stimuli set by Brady et al. (2008), and image search on the Internet. All images were resized to 240 pixels × 240 pixels and presented on a square white background. For each subject, a total of 416 images were randomly selected as the study items (104 items per block). The test lists consisted of 100 old items from the preceding study list with 50 foil items given in random order. The first and last two

stimuli in the study list were excluded from the test list to reduce primacy and recency effects.

During the study phase, the study items were presented on either the left or right side of the fixation cross. The subjects were instructed to memorize the side of the screen on which each study item was given. The spatial location of the item was considered as the source information in this experiment. A study item was shown for 1000 ms followed by an inter-stimulus interval with varying lengths (uniformly distributed within 625 ± 125 ms). A visual Gaussian noise image was shown at the locations of study item presentations whenever an item was not being presented to prevent after-image effects from the stimulus. The area containing the possible study image locations subtended a visual angle of $11.4°$ wide $\times$ $5.6°$ high.

In the recognition phase, a fixation cross appeared on the center of the screen for 750 ms. A test item was shown for 750 ms on top of the fixation cross followed by a 1500 ms long fixation cross. The visual angle of each test probe image was $4.3°$ wide $\times$ $4.3°$ high. Then the subjects were given two consecutive questions where the second question type depended on the subject's answer on the first one. An inter-stimulus interval of 625 ± 125 ms followed each response. In the first question, subjects were asked to make a source/new judgment where source was the location of the item in the study phase. The first question had three options: left, right (given as L and R, respectively) and a new judgment (given as N). If the subjects responded with L or R in the first question, they were asked to give a modified R-K judgment in the second question. The R-K judgment question had three options: remember side (given as RS), remember other (given as RO), and familiar (given as F). Subjects were instructed to respond with RS if they remembered the source information, RO if they remembered something other than the source information, and F if they could not remember any details of learning the item but it looked familiar. If the subjects responded with new in the first question, they were asked to give a confidence of that response: sure (given as S) or maybe (given as M) based on how confident they were about it being a new item. See **Figure 1A** for an illustration of the study and test tasks in the experiment. The keys for left responses were assigned to the left hand (z or x key), the keys for right responses were assigned to the right hand (. or / key), and the keys for new responses were assigned to one of the outermost keys (z or / key). For the confidence judgments, the keys were set up from left to right to follow memory strength in either descending or ascending order. The familiar (F) responses and remember (RS/RO) responses were always assigned to different hands. The key assignment was fixed for a given subject, but all possible key combinations were distributed to an equal number of subjects.

EEG was recorded with a 128-channel Geodesic Sensor Net[TM] [HydroCel GSN 200, v.2.1; Tucker (1993)] at 250 Hz sampling rate using an AC-coupled 128-channel, high-input impedance amplifier (300 MΩ, Net Amps[TM]; Electrical Geodesics Inc., Eugene, OR, United States) with a 0.1–100 Hz bandpass filter. Initial common reference was the vertex channel (Cz) and the individual electrodes were adjusted until impedance measurements were lower than 40 kΩ. **Figure 2** shows the locations of the electrodes.

Experiment 2
Participants
The subjects were right-handed University of Colorado undergraduate students (ages 18–27, mean = 21.2) who volunteered for paid participation ($15 per hour) or course credit (17 male, 13 female). All subjects were native English speakers and had normal or corrected- to-normal vision.

Experimental Paradigm and EEG Acquisition
The stimuli set used in Experiment 1 was used in Experiment 2. In the study phase, the study items were presented with a 48-pixel wide color frame with eight possible colors (purple, green, blue, pink, red, orange, yellow, and brown). The color of the frame was considered as the source information in this experiment. Two of the four study lists used six colors and the two other study lists used the two remaining colors. Half of the subjects received the two-color condition in the even blocks and the other half of the subjects received the two-color condition in the odd blocks. All colors were randomly and evenly distributed over the study items.

During the study phase, the subjects were instructed to memorize the frame color with each of the presented study items. A study item was shown for 1500 ms followed by an inter-stimulus interval with varying lengths (625 ± 125 ms). A visual Gaussian noise image was given at the location of study item presentation whenever an item was not being presented to prevent after-image effects from the stimulus.

In the recognition phase, a fixation cross appeared for 750 ms with a preview of the two colors the subject would be choosing from immediately following the test item presentation. The number of preview colors were set to two for both six- and two-color conditions. If the test item was old (i.e., given in the preceding study list), its corresponding frame color was given in the preview. After the color preview, a test item was shown for 750 ms followed by a 1500 ms long fixation cross. Then the subjects were given two consecutive questions where the second question type depended on the subjects' answer on the first one. In the first question, subjects were asked to make a source/new judgment where source was the frame color given with the item in the study phase. The first question had three options: two colors (given as solid color squares) and a new judgment (given as N). If the subjects responded with a color in the first question, they were asked to give a modified R-K judgment in the second question. The R-K judgment question had three options: remember color (given as RC), remember other (given as RO), and familiar (given as F). Subjects were instructed to respond with RC if they remembered the source information, RO if they remembered something other than the source information, and F if they could not remember any details of learning the item but it looked familiar. If the subjects responded with new in the first question, they were asked to give a confidence of that response: sure (given as S) or maybe (given as M) based on how confident they were about it being a new item. See **Figure 1B** for an illustration of the study and test tasks in the experiment.

EEG was recorded as for Experiment 1.
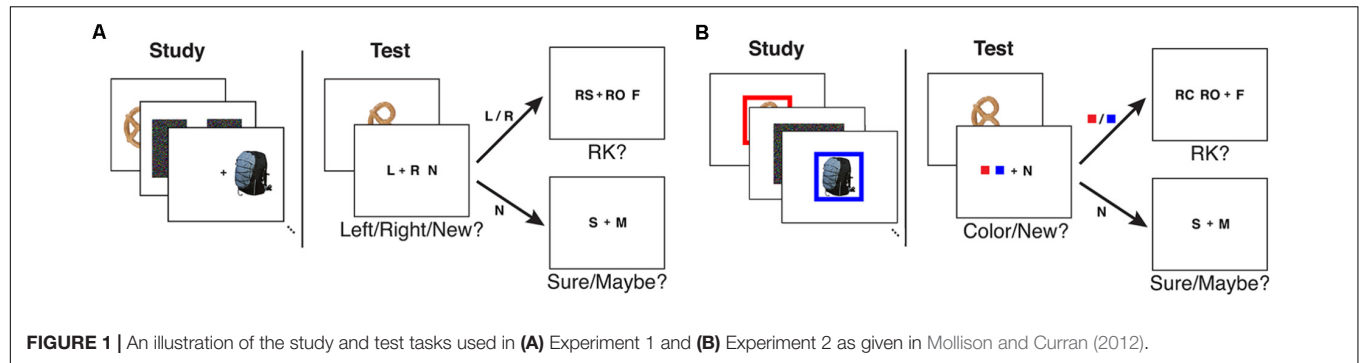
FIGURE 1 | An illustration of the study and test tasks used in **(A)** Experiment 1 and **(B)** Experiment 2 as given in Mollison and Curran (2012).
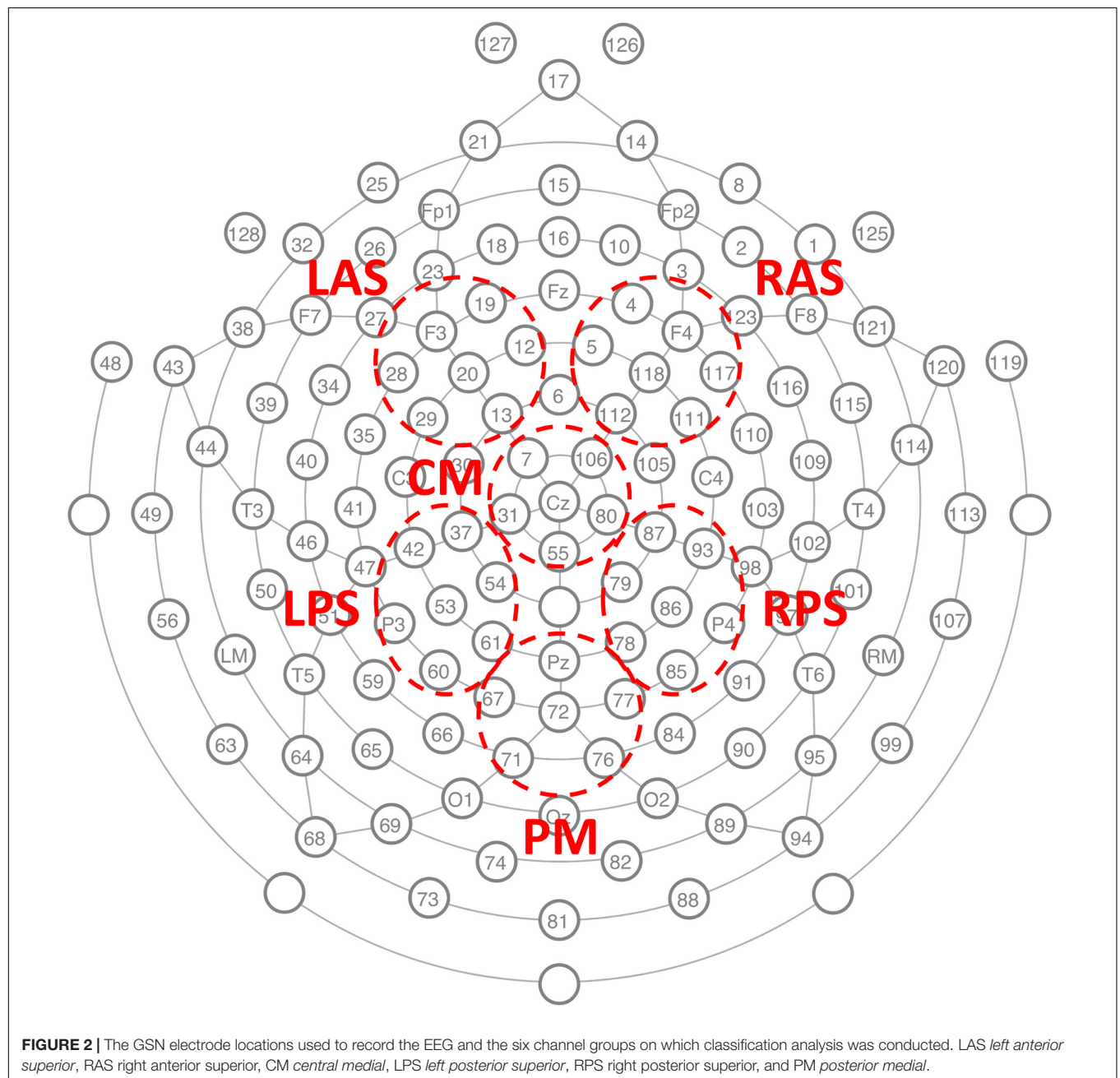
FIGURE 2 | The GSN electrode locations used to record the EEG and the six channel groups on which classification analysis was conducted. LAS *left anterior superior*, RAS right anterior superior, CM *central medial*, LPS *left posterior superior*, RPS right posterior superior, and PM *posterior medial*.

Experiment 3

Participants

The subjects were right-handed University of Colorado undergraduate students (ages 18–29, mean = 20.6) who volunteered for paid participation ($15 per hour) or course credit (21 male, 17 female). All subjects were native English speakers and had normal or corrected- to-normal vision.

### Experimental Paradigm and EEG Acquisition

The experiment was conducted in two separate sessions occurring on separate days. Each session consisted of four lists where two lists were the location source paradigm (as in Experiment 1) and two lists were the color source paradigm (as in Experiment 2). Only two frame colors (blue and yellow) were used for the color condition to match the number of location and color conditions across lists). For the first session, half of the subjects received the color condition in the even list numbers and the other half of the subjects received the color condition in the odd list numbers. The second session used the opposite order. The stimuli used in the two previous experiments were used for this experiment.

For both source conditions, a source indicator frame (color condition: blue/yellow frame, location condition: white frame on the left/right side of the screen) appeared on top of the visual Gaussian noise image prior to each study item presentation for 500 ms. Then the study item was presented inside the source indicator frame for 2000 ms followed by a slightly increased inter-stimulus interval (1125 ± 125 ms).

The timing of the recognition phase was the same as the previous experiments. However, a number of changes were made to the procedures. No color preview was given prior to test item presentation during the color condition lists. Also, the solid color squares used as source cues (as in Experiment 2) were changed to letters B and Y to better match the location conditions. Finally, both of the source responses (B and Y/L and R) were assigned to one hand and the new response (N) was assigned to the other hand. The key assignments were counterbalanced across subjects.

Electroencephalography was recorded with the same equipment as in the previous experiment except with a 500 Hz sampling rate and without the 0.1 Hz hardware high-pass filter.

Pre-processing

Electroencephalography epochs from the recognition phase of each experiment were extracted and recalculated to average reference. In order to address any possible deficiencies in the average reference method, a subset of analyses were repeated using the reference electrode standardization technique (REST) (Yao, 2001; Dong et al., 2017; Lei and Liao, 2017), which uses reconstructed equivalent sources to re-reference electrode signals relative to a reference at infinity. The lead field for using REST had 3000 potential sources corresponding to the 128-channel HydroCel Geodesic Sensor Net™ recording system used (Mollison and Curran, 2012).

Each epoch was filtered between 0.1 and 50 Hz using a 40 tap FIR filter and baseline corrected using data from $-200-0$ ms. Data from Experiment 3 were down-sampled to 250 Hz after the pre-processing procedure to match the sampling rate of Experiments 1 and 2.

Classification Problem

Classification analysis was conducted separately on Experiment 1, Experiment 2, location source blocks from Experiment 3 (denoted as Experiment 3-location or Exp 3-loc), and color source blocks from Experiment 3 (denoted as Experiment 3-color or Exp 3-col). The data from Experiment 3 were divided into the different source conditions in order to reveal any potential differences between the location and color conditions that may correspond to ERP differences observed in Mollison and Curran (2012). Before conducting classification, the trials were divided into five conditions (*SC*: source correct, *SI*: source incorrect, *CR*: correct rejection, *M*: miss, *FA*: false alarm) based on their source judgments (1st response) as illustrated in **Figure 3**. Note that in **Figure 3** and for the rest of the paper, RS refers to remember source which includes both remember side and remember color.

The classifiers were trained to find the projection function onto the vector perpendicular to the decision boundary (we sometimes refer to these vectors as planes) which is characterized by the choice of the training conditions. The behavioral conditions corresponding to correct item retrieval (SC and SI) and correct item rejection (CR) were selected for training. As a result, three different two-class binary classifiers (SC-CR, SI-CR, and SC-SI) with probability outputs ($0 \leq p \leq 1$) were trained to discriminate between pairs of behavioral conditions. These probability outputs given by the classifiers are denoted as *classifier scores* in this paper. The classifiers were trained on each individual subject and only the subjects with a minimum of 25 trials for each of the 2 trained conditions (SC, SI, and CR) were included in the analysis. For each classification problem, the classifier scores were also computed for the trials which were not included in the training procedure (non-training trials).

(1) SC-CR classifier

   The SC-CR classifier (trained to discriminate between SC and CR) was expected to find the projection which

FIGURE 3 | Categorization of the trials based on the subjects' source judgments (SC, source correct; SI, source incorrect; CR, correct rejection; M, miss; FA, false alarm) and subjective ratings (RS, remember source; RO, remember other; F, familiar; MN, maybe new; SN, sure new).
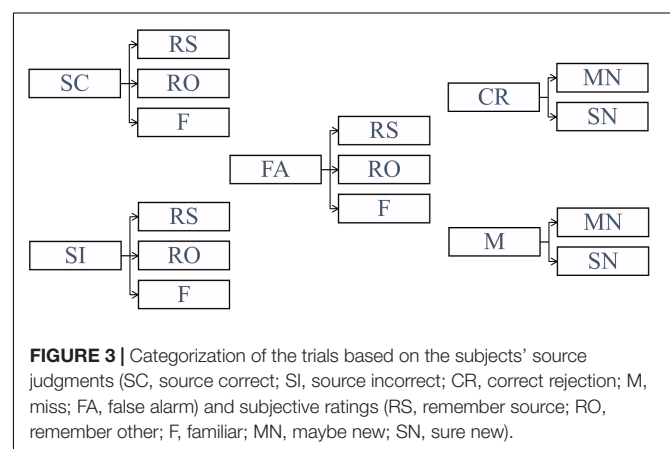
maximizes the difference in the amount of information retrieved from the study episode.

(2) SI-CR classifier

This classifier (trained to discriminate between SI and CR) was designed to discriminate between correctly retrieved old items (with incorrect source judgments) and the correctly rejected new items.

(3) SC-SI classifier

The SC-SI classifier (trained to discriminate between SC and SI) was designed to distinguish the correctly retrieved old items with correct source judgments from those with incorrect source judgments. Hence the classifier would extract the information on source memory retrieval.

Classification

The spatio-temporal structure of the ERPs was extracted based on previous findings on the old/new effect. Six channel groups were selected for evaluation (LAS, RAS, CM, LPS, RPS, and PM) as given in **Figure 2**. The average voltage for each channel group

was computed and the data between 300 and 800 ms after test item presentation were extracted to take advantage of the ERP effects related to memory retrieval. The dimensionality of these subsequences was reduced to 5 by averaging over 100 ms length non-overlapping windows. The features from all six channel groups were concatenated to build a 30-dimensional feature vector for each trial. A binary classifier using linear discriminant analysis (LDA) with automatic shrinkage (Ledoit and Wolf, 2004; Schaefer and Strimmer, 2005) was trained to classify these feature vectors (Lotte et al., 2007; Blankertz et al., 2011). In order to avoid any overfitting to the training data, the projections for the training conditions were computed using leave-two-out (one from each class) cross-validation. In order to train with balanced classes, trials from the majority class were randomly discarded (from training) to have equal numbers of trials in each class. These trials, however, were still used for evaluation of the classifier (using a classifier trained on all the selected

TABLE 1A | Classification results for Experiment 1.

Subject	SC-CR	SI-CR	SC-SI
102	0.5538	*0.4702*	*0.4693*
103	*0.4857*		
104	**0.6875**	0.5572	**0.6540**
106	**0.6720**	0.5455	0.5358
108	**0.6550**	0.4891	0.4949
109	0.5593		
110	**0.5947**	*0.4712*	*0.5*
112	*0.4953*		
113	**0.6667**	*0.328*	***0.6271***
114	**0.6746**	0.5575	**0.6442**
115	**0.5741**	0.5269	0.5517
116	0.5251	0.5183	0.4839
117	**0.5944**	*0.5148*	*0.5025*
118	**0.65**	*0.4762*	*0.5398*
119	**0.6154**	**0.5756**	0.5
120	**0.6172**	*0.5108*	***0.6090***
121	**0.5977**	0.5057	0.5356
122	0.5255	***0.6224***	*0.5420*
123	**0.6585**	0.5603	0.5368
124	*0.5649*	0.5577	0.6104
125	**0.6518**	0.5571	0.4373
126	**0.6955**	*0.4419*	***0.5981***
127	**0.6048**	*0.5530*	*0.5194*
128	**0.7474**	*0.4633*	*0.5302*
129	**0.5542**	*0.4328*	*0.4714*
Overall	0.6231	0.5090	0.5383
Overall with 50 trials/class cutoff	0.6290	0.5368	0.5397

Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance (p < 0.05). Results from subjects with less than 50 trials per condition are italicized.

TABLE 1B | Classification results for Experiment 2.

Subject	SC-CR	SI-CR	SC-SI
201	*0.5113*	*0.5581*	0.5298
202	***0.6860***	0.5915	*0.4713*
203	*0.5444*	0.4805	0.4773
204	**0.6018**	*0.48*	***0.5787***
205	**0.5766**	0.4303	0.5126
206	0.5524	*0.5816*	*0.4563*
207	**0.6204**	*0.4177*	*0.4912*
208	**0.6349**	0.4643	**0.5701**
209	***0.6222***	*0.5714*	0.5344
210	**0.6189**	0.5356	0.5193
211	**0.6**	0.4851	0.5193
212	*0.4964*	*0.5146*	0.5723
213	**0.5942**	**0.5922**	0.5071
214	**0.6553**	0.5659	0.5211
215	**0.6728**	**0.6554**	0.5270
216	0.5475	0.5315	0.5571
217	**0.5761**	0.5484	0.52
219	0.5506	*0.4859*	*0.4757*
220	**0.6174**	**0.6171**	0.5436
221	**0.5737**	0.4826	0.5345
222	0.5504	0.4758	0.5780
223	0.55	0.4943	0.4775
224	**0.5789**	**0.5822**	**0.6053**
225	0.5538	**0.5952**	0.5431
227	**0.6757**	**0.6**	0.5410
228	*0.5533*	*0.5784*	0.5287
229	0.4588	0.4262	0.4456
230	**0.5771**	**0.5805**	0.5171
Overall	0.5904	0.5383	0.5263
Overall with 50 trials/class cutoff	0.5945	0.5416	0.5311

Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance (p < 0.05). Results from subjects with less than 50 trials per condition are italicized.

TABLE 1C | Classification results for Experiment 3.

Subject	SC-CR (loc)	SC-CR (col)	SI-CR (loc)	SI-CR (col)	SC-SI (loc)	SC-SI (col)
310	**0.5665**	0.4833	*0.4538*	0.5	*0.5289*	0.5208
312	**0.6422**	**0.6257**	*0.5380*	0.5451	*0.4591*	0.5208
313	**0.57**	**0.5766**		0.5509		0.5498
315	0.5127	**0.5949**	0.5074	0.5505	0.5181	0.5064
317	**0.6039**	**0.6101**	0.5153	0.5	0.5019	0.4704
318	*0.5327*	*0.5469*	*0.6026*	*0.5185*	*0.5315*	*0.5041*
319	0.5550	*0.4706*	0.4581	*0.5399*	0.5444	0.4681
321	0.5613	**0.6291**	**0.5985**	0.5369	0.5076	**0.5647**
322	*0.5252*	*0.4818*	*0.4811*	*0.5076*	0.5170	0.4696
323	0.5615	0.5291	*0.4538*	0.5505	*0.5519*	0.5528
324	0.5244	0.5393	0.5407	0.5056	0.5430	0.5358
326	**0.6181**	**0.5700**	0.4928	0.4962	0.5516	0.4697
327	**0.6111**	0.5022		0.5586		0.5054
328	0.512	0.4971	*0.5515*	0.52	*0.5426*	0.5137
330	**0.6348**	**0.5638**	0.4462	0.4908	**0.6035**	0.4967
332	**0.5710**	**0.5698**		0.4632		0.4615
333	0.5593	0.5445		0.5724		0.5364
334	*0.5370*	0.4563	*0.5149*	**0.5984**	0.5195	0.4978
335	0.4963	**0.5714**		0.5		*0.4701*
336	**0.6262**	**0.6313**	*0.4545*	0.4527	***0.5685***	0.4963
337	**0.5638**	0.5481	0.5546	0.5037	0.5547	0.5
340	*0.4875*	**0.6138**	*0.3663*	0.4965	0.4785	0.5134
342	**0.5980**	0.5233	*0.4710*	0.5425	*0.4651*	0.4870
343	**0.6667**	*0.5179*	0.5652	***0.5870***	0.5059	0.4805
344	**0.5682**	0.5158	0.4930	0.4667	0.5081	0.4727
345	0.5101	0.4688	*0.5429*	0.5058	*0.5587*	0.5053
Overall	0.5736	0.5553	0.5049	0.5182	0.5261	0.5024
Overall with 50 trials/class cutoff	0.5789	0.5610	0.5153	0.5159	0.5286	0.5031

Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance ($p < 0.05$). Results from subjects with less than 50 trials per condition are italicized.

balanced training data). The data from the remaining conditions (e.g., Misses and False Alarms) were not used to evaluate the classifier, but were still projected onto the discriminative vector (learned from the entire balanced training set) for interpretative analysis.

Statistical Methods

The average classifier scores (for a given classification problem) across all subjects were compared across different behavioral conditions (SC, SI, CR, M, and FA). The classifier score is a projection of the high-dimensional EEG data onto a 1-dimensional vector which is representative of the given classification problem. Paired t-tests were conducted on the trial-by-trial classifier scores separately for the four available datasets to compare the classifier scores of the different retrieval/subjective rating conditions. A comparison was considered to be significant only when all four separate datasets gave p-values below 0.05 for the conditions of interest.

It is advantageous to also visualize the EEG features utilized by the classifiers for interpreting any effects identified from the multivariate analysis using the pattern classifiers. This was done by analyzing the classifier activation patterns representing which channel, time pairs were important for classification (Haufe et al., 2014). For each source type, the 30-dimensional classifier activation pattern vector for each subject was normalized to have length 1.

In order to identify features consistent across subjects, a cluster-based method for correction for multiple comparisons was used (Maris and Oostenveld, 2007). In this method, first each spatiotemporal pixel significantly different from zero ($p < 0.05$) was identified. Then the t-statistic of all significant flagged neighboring pixels with the same sign was summed and the maximum absolute value over all clusters taken. This value is compared to the distribution of max absolute cluster values obtained from a permutation distribution resulting from 10,000 random permutations of class labels for each subject. Temporal neighbors were temporally adjacent time windows. Spatial groups were considered neighbors if they contained adjacent electrodes from the cap layout (see **Figure 2**). Using this rule, LAS, CM, and RAS were all mutual neighbors; CM was also neighbors with LPS and RPS; LPS and RPS were also neighbors with PM.

RESULTS

Classifier Performance

Performance of the SC-CR classifier was computed based on classification of the SC and CR trials (SC-RS, SC-RO, SC-F, CR-SN, CR-MN). The significance of the performance of a classifier (whether it performs significantly over chance) was evaluated based on the number of test trials used for classification. The 95% confidence interval for the obtained accuracy was calculated using Wald intervals with small sample size adjustments (Agresti and Caffo, 2000) for each subject. Classification results were considered to be significantly over chance only when the interval did not include 50%. Results are given in **Table 1**. The overall classification accuracy for Experiment 1 (SC-CR) was 62% with 18 of 25 subjects having individual accuracies significantly over chance. When restricted to subjects with at least 50 trials in each class, the performance is somewhat better. The overall classification accuracy for Experiment 2 (SC-CR) was 59% with 17 of 28 subjects having individual accuracies significantly over chance. Experiment 3-loc (SC-CR) had an average accuracy of 57% and Experiment 3-col (SC-CR) had an average accuracy of 56%.

Figure 4 gives the ROC (receiver operating characteristic) curves for choosing different thresholds (between 0 and 1) to make decisions between classes 1 and 2 for all 3 classification problems. **Table 2** gives the area under these ROC curves. All results were above 0.5, however, there was a variability in performance across the different classification problems. The SC-CR classifiers showed the highest performance on all four datasets. It was also found that the datasets with recordings from multiple days (Exp 3-loc and Exp 3-col) showed a slight decrease in performance compared to the single session datasets. The SC-SI classification performs better for the location source datasets relative to the color source datasets in contrast with the SI-CR classifiers.

We redid some classifications using the reference electrode standardization technique (REST) (Yao, 2001; Dong et al., 2017). The performance of classifiers SC-CR in Experiment 1 using REST for re-referencing pre-processing showed comparative AUC (0.6571) and accuracy (0.6221) to that obtained with our usual average reference method (AUC of 0.6555 and accuracy of 0.6231). We then compared the REST method for the harder SC-SI classification in the two color source datasets, but this also resulted in no significant improvement in the classification results. Specifically for SC-SI in Exp2 (color) with REST we have AUC of 0.5206 (vs. 5357 with AR) and accuracy of 0.5191 (vs. 5263 with AR) and for SC-SI in Exp3 (color) with REST we have AUC of 0.5075 (vs. 5108) and accuracy of 0.5034 (vs. 5024).

Analysis of the Classifier Scores

The projection weights for a given classification problem can be used to project the EEG data onto a discriminative vector. In this paper, these projection values are denoted as the classifier scores. The relationship between the average classifier scores for the different behavioral conditions represents the characteristics of the different discriminative hyperplanes (Noh and de Sa, 2014). As described in section "Statistical Methods," the representation of the EEG data on the three different discriminative vectors were compared across the different behavioral conditions. The classifier scores were computed for each classification problem (as described in section "Classification Problem") and the average scores corresponding to the different behavioral conditions were compared. The results were compared across the four datasets and effects with $p < 0.05$ consistently across the different datasets were considered to be meaningful (the individual comparison results are given in **Table 3**).

The correct item memory conditions (SC, SI, and CR) showed similar patterns across the different projections where SC trials gave the highest scores and the CR trials showed the lowest scores. However, the relative distance between the three conditions varied across the different discriminative vectors. It was found that the SI condition was mapped closer to the CR condition on the SC-SI plane (see **Figure 5C**) while it was mapped closer to the SC condition on the SI-CR plane (see **Figure 5B**). It was also found that the difference between the SI and CR trials were only significant ($p < 0.05$ for all four datasets) on the SC-CR and SI-CR planes (see **Figure 5**).

The relative mapping of the error conditions (M and FA) with respect to the correctly retrieved/rejected conditions (SC, SI, and CR) gave different patterns for the different projection directions. Interestingly, the source correct (SC) trials and false

FIGURE 4 | The ROC curves for the three different classification problems (**A**: SC vs. CR; **B**: SI vs. CR; and **C**: SC vs. SI) are given separately for the four individual datasets.
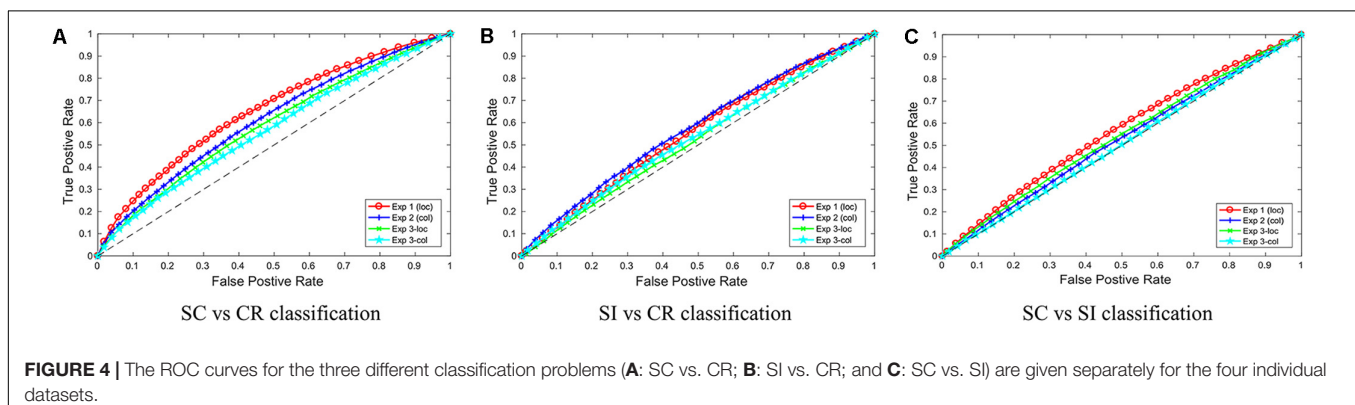
TABLE 2 | The average AUCs of the ROC curves for the three different classification problems on four experiments.

	SC-CR	SI-CR	SC-SI
Experiment 1 (loc)	0.6555	0.5586	0.5434
Experiment 2 (col)	0.6160	0.5779	0.5357
Experiment 3 (loc)	0.5916	0.5264	0.5375
Experiment 3 (col)	0.5726	0.5376	0.5108
Average	0.6089	0.5501	0.5319

Every AUC was computed using the projections of the trials, by leave-two-out training, in the selected balanced training data.

alarms (FA) were mapped to significantly different values on the SC-CR and SC-SI plane but not on the SI-CR plane (see **Table 4**). In contrast, the misses (M) gave values significantly lower than the two correct item retrieval conditions (SC and SI) when mapped onto the SC-CR and SI-CR plane.

A similar analysis was conducted considering the different subjective ratings given to the correct item retrieval/rejection trials (SC, SI, and CR). These responses consisted of remember source (RS), remember other (RO), and familiar (F) for the SC/SI conditions and sure (SN denoting sure new) and maybe (MN denoting maybe new) for the CR condition. The error conditions (FA and M) can be similarly projected. While the classifiers generally gave a monotonic decrease in classifier scores from the RS to SN conditions, there were interesting interactions with the memory retrieval conditions as illustrated in **Figure 6**.

Classifier Activation Patterns

The activation patterns which represent the features used by the classifiers (or the characteristics of the projection weights) were compared across the three different classification problems (see **Figure 7**). The activation patterns were computed for each subject and the average activation patterns were computed by averaging the values across all four datasets. A *t*-test was conducted on each of the features to illustrate which features showed similar effects across the different subjects. Cluster based analysis (Maris and Oostenveld, 2007) was then used to control for multiple comparisons. This revealed features with values significantly above/below zero across all the subjects available for analysis. The activation patterns are given as a 2-dimensional matrix with its corresponding channel groups and time segments (the times give the center of the interval) in **Figure 7** and the most significant clusters (with significance values) are shown in **Figure 8**.

The SC-CR classifier utilized temporal features from 300 to 800 ms. The SI-CR classifier only showed consistent patterns

between 300 and 700 ms and the SC-SI classifier showed consistent patterns between 400 and 800 ms for the two tasks with spatially presented contextual information (1 and 3-loc). In the two tasks with colored frames as context, there is not a strong activation pattern consistency across subjects for the SC-SI classifier. Interestingly the SC-SI (source memory) classifier has strong consistent activity across all spatial areas except PM when the source context is location. The SI-CR (item memory) classifiers have an early frontal activation when the source context is the colored outline, but a more parietal activation when the source context is the location.

The activation patterns for the three classifiers we created using the REST preprocessing (SC-CR Exp1, SC-SI Exp 2-col, SC-SI Exp 3-col) were similar to the analogous ones with average referencing.

Classifier Scores Evolution Over Time

In the activation patterns, the characteristics of projection weights in different time intervals and channel groups were shown. The classifier scores variation across time gives a clear insight about the evolution of the separation of classes over time. To obtain the scores only under the operation with weights between 300 and 400 ms in activation patterns, the grouped EEG data after 400 ms were set to zero, and the remaining computations remained the same. In brief, the data were set to zero after the considered intervals and the trained classifier was used to get the classifier scores.

Figure 9A shows that the scores of SC and CR trials start to be discriminable around 500–600 ms and separate further afterwards. **Figure 9B**, shows that with the SI-CR classifier, scores of SI and CR trials also start to separate around 500–600 ms. As for the SC-SI classifier in **Figure 9C**, the scores of SC trials become more separable from the scores of the SI trials after about 700 ms. Note that while the activation patterns for the SI-CR classifier show not much significant activation that is consistent between subjects after 600 ms, the classifier scores continue to separate, indicating that the activation patterns causing this separation are less consistent between subjects.

DISCUSSION

The results show that it is possible to predict successfully identified old vs. new items based on single-trial scalp EEG activity recorded during the retrieval episode. The prediction rate was higher for the location-source datasets and the average accuracy of the single-session datasets was higher compared

TABLE 3 | Comparison results between the classifier scores for the SC-CR classifier.

	SC vs. CR	SC vs. SI	SC vs. M	SC vs. FA	CR vs. SI	CR vs. M	CR vs. FA	SI vs. M	SI vs. FA	M vs. FA
Exp 1 (loc)	5.79E-110	4.11E-11	6.45E-47	9.77E-08	5.22E-15	0.24363	1.08E-06	5.56E-12	0.21081	5.07E-08
Exp 2 (col)	5.53E-49	0.002464	1.93E-14	0.00092829	3.29E-27	0.14601	2.93E-07	6.62E-07	0.51457	0.0022279
Exp 3-loc	6.86E-34	0.0054622	1.36E-25	4.43E-10	3.89E-06	0.86631	0.0043021	1.34E-07	0.065969	0.0012858
Exp 3-col	3.63E-27	9.87E-05	2.25E-09	2.27E-07	7.60E-11	0.76771	0.23946	0.0001848	0.00011229	0.14554

Paired t-tests between all possible pairs are given with their corresponding uncorrected p-values.
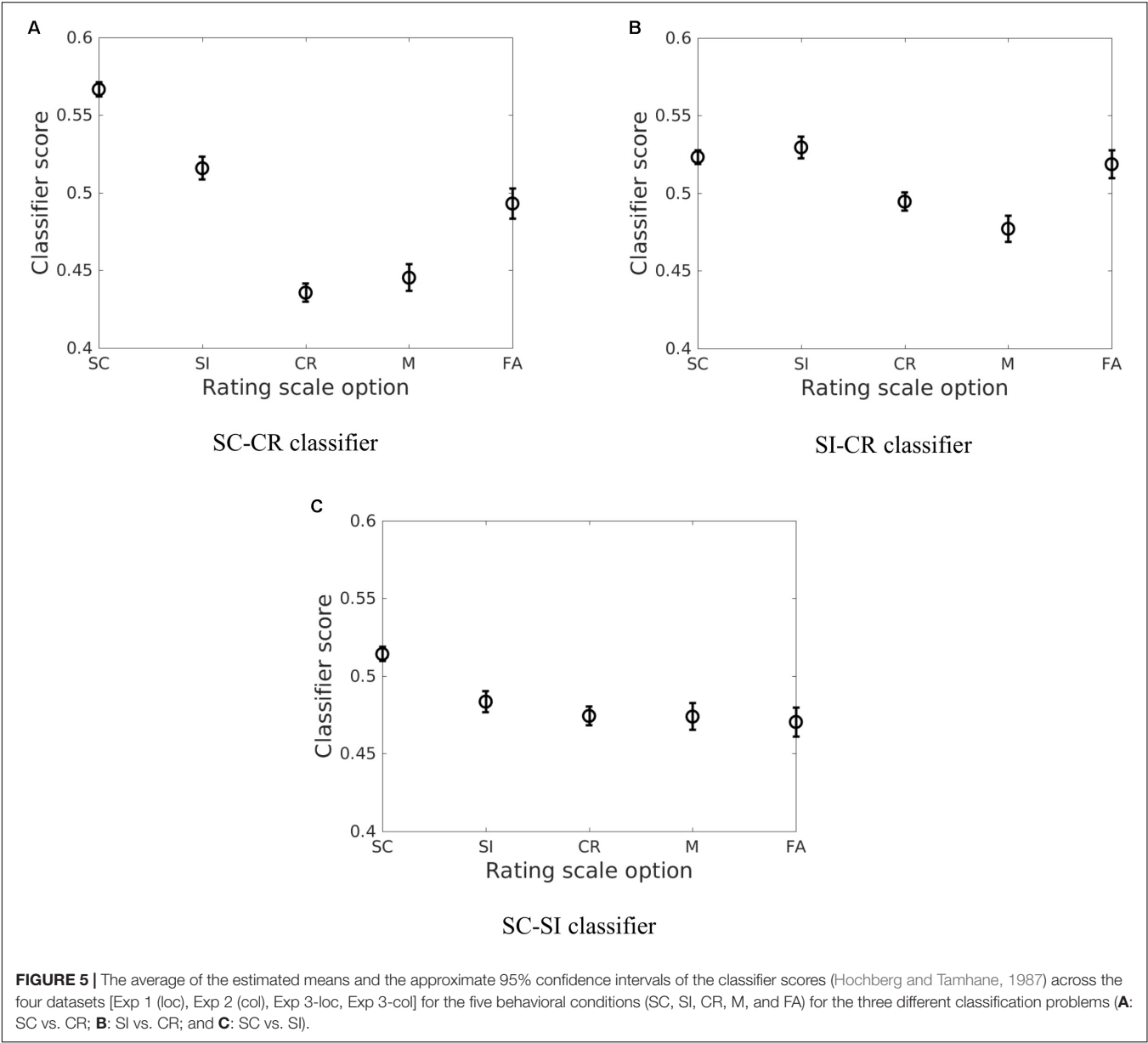
FIGURE 5 | The average of the estimated means and the approximate 95% confidence intervals of the classifier scores (Hochberg and Tamhane, 1987) across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col] for the five behavioral conditions (SC, SI, CR, M, and FA) for the three different classification problems (**A**: SC vs. CR; **B**: SI vs. CR; and **C**: SC vs. SI).

TABLE 4 | The uncorrected pairwise comparison results for the five behavioral conditions across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col].

	SC vs. CR	SC vs. SI	SC vs. M	SC vs. FA	SI vs. CR	CR vs. M	CR vs. FA	SI vs. M	SI vs. FA	M vs. FA
SC-CR	2.35E-193	1.75E-28	6.41E-83	4.86E-26	9.18E-49	0.055638	8.80E-14	2.07E-25	0.0504	6.10E-14
SI-CR	3.46E-10	0.12096	8.04E-10	0.52851	1.44E-11	0.023804	9.54E-05	9.27E-14	0.32681	3.38E-11
SC-SI	1.52E-23	6.84E-06	8.67E-11	1.36E-10	0.0043127	0.67362	0.19108	0.0098251	0.0014469	0.32367

The results for the different projections are given in separate rows.

to the multi-session datasets. The non-stationarity of the data between the two sessions (due to electrode position changes, impedance changes, or changes in brain-state) likely contributes to the drop in classification performance (Krauledat et al., 2007). Our analysis was restricted to time domain signals from specific channel groups known to be involved in frontal and parietal old/new effects. It is possible that accuracy could

be increased by using frequency domain information from multiple electrode location and frequency bands (see for example, Hammon and de Sa, 2007; Hammon et al., 2008; Velu and de Sa, 2013; Noh et al., 2014a; Mousavi et al., 2017).

The current analysis found that the projections of the temporal information from the EEG data onto different hyperplanes

FIGURE 6 | The average of the estimated means and the approximate 95% confidence intervals of the classifier scores (Hochberg and Tamhane, 1987) across the four datasets (Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col) when considering the breakdown by subjective ratings (RS, RO, F, MN, and SN) for the three different classification problems (**A**: SC vs. CR; **B**: SI vs. CR; **C**: SC vs. SI).
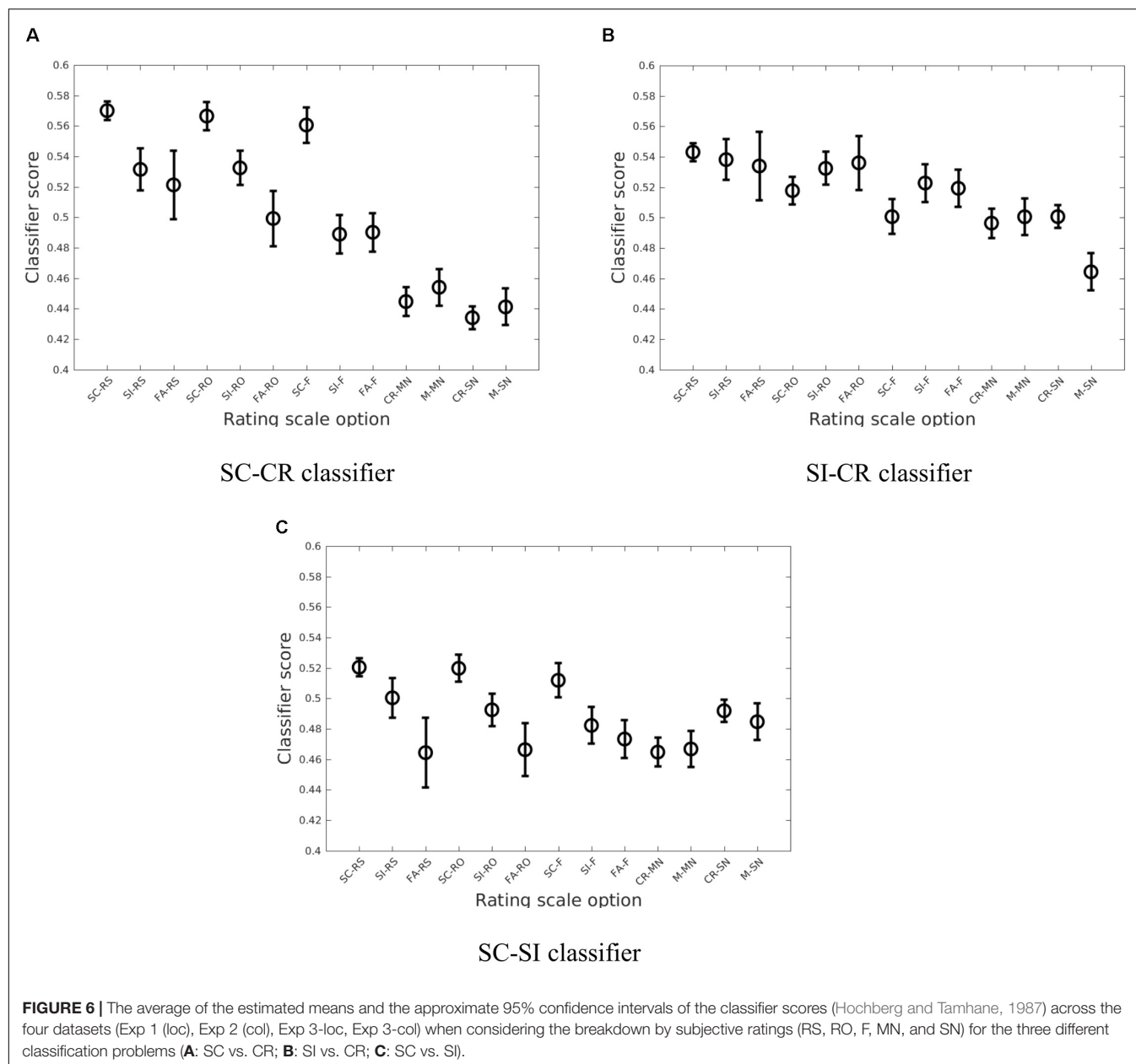
show different patterns. This was evident in the relationship between the behavioral conditions of interest. We focused on the patterns which were consistent across multiple subjects and multiple datasets to compare across the different classifiers. These results suggested that the classifier may be exploiting features which are more informative for discriminating between the two behavioral conditions selected for training. It was found that the SC-SI classifier performance on the two color source datasets (Experiment 2 and the color blocks from Experiment 3) was lower compared to the location datasets (Experiment 1 and the location blocks from Experiment 3). The activation patterns for the SC-SI classifier were also not significantly consistent across subjects for the color outline source. In Mollison and Curran (2012), it was found that

accurate/inaccurate judgments to the familiar responses were affected by source type where the SC trials with familiar ratings and SI trials with familiar ratings were significantly different only for the location-source datasets when comparison was conducted on a ROI centered at FCz. This suggests that the temporal information in the EEG signal may be less separable between SC and SI trials for the color datasets compared to the location datasets resulting in a lower classification performance.

The relationship between the correctly remembered conditions (where the classifier scores showed CR < SI < SC on all three discriminative vectors) suggests that these classifier scores may reflect the amount of information retrieved from the study episode. The difference in the amount of information
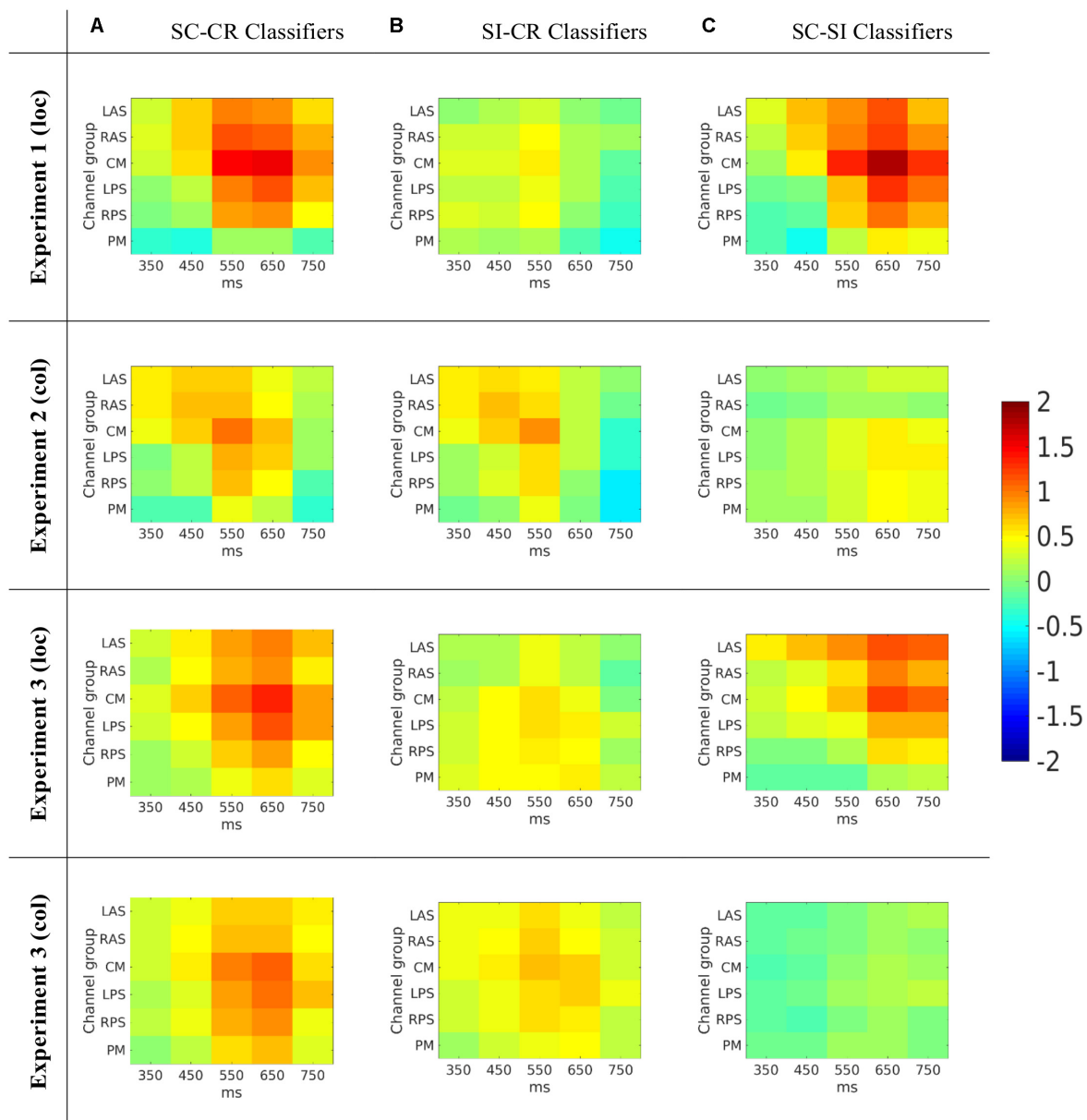
FIGURE 7 | (A) The average activation patterns from the SC-CR classifiers averaged across all available subjects. **(B)** The average activation patterns from the SI-CR classifier averaged across all available subjects. **(C)** The average activation patterns from the SC-SI classifiers averaged across all available subjects. Note that the numbers on the *x*-axes represent the mid-point of the 100 ms window used to compute the features.

retrieved from the study episode is maximal between conditions SC (when the correct item is retrieved from the study phase with the appropriate source information) and CR (when no information is retrieved from the study phase) which may be why the SC-CR classifier outperformed the other two classifiers. The drop in classifier performance for the SC-SI and SI-CR classifiers compared to the SC-CR classifier may be due to this innate relationship between the 3 behavioral conditions used for classifier training. The SI-CR classifier would primarily be able to utilize information related to differences between item retrieval vs. correct rejection to distinguish between the two classes. On

the contrary, the SC-SI classifier would only be able to utilize information related to source memory differences between correct source retrieval vs. incorrect source retrieval in order to distinguish between the SC and SI conditions.

The activation patterns (see **Figure 8**) indicated that the classifiers used features mostly around 400 to 800 ms and gave these features higher weights. The spatiotemporal distribution of predictive features associated with the (SI - CR) classifiers (early and more frontal) were somewhat consistent with the timing and location of the FN400 only in the color-source experiments [2 and 3(col)]. Likewise the spatiotemporal

FIGURE 8 | The average activation patterns from **Figure 7** but masked by the most significant cluster (with *p*-value for that cluster given below) (Maris and Oostenveld, 2007) for the three different classification problems (**A**: SC vs. CR; **B**: SI vs. CR; **C**: SC vs. SI). Note that the numbers on the *x*-axes represent the mid-point of the 100 ms window used to compute the features.
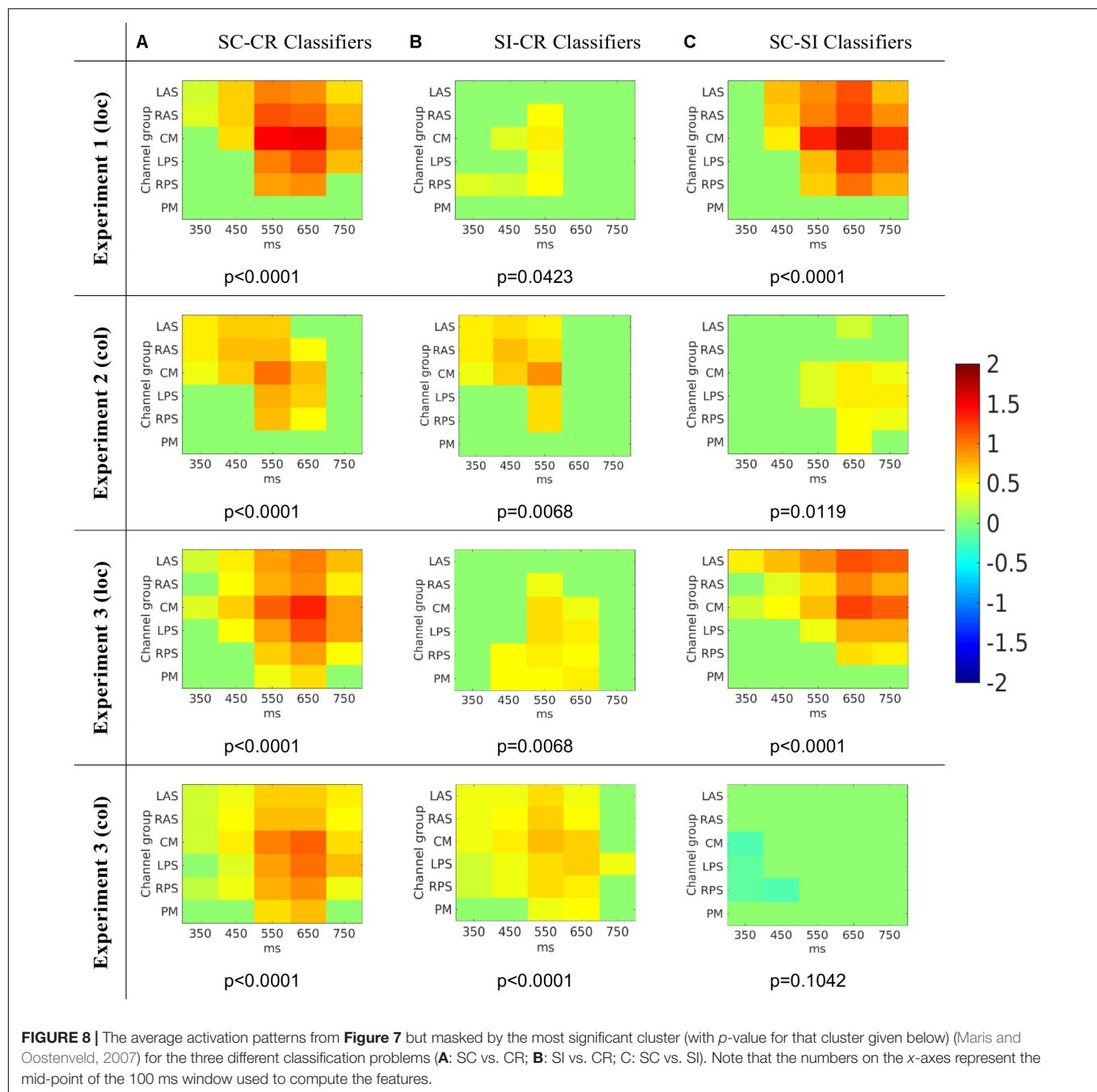
distribution of predictive features associated with the (SC-SI) classifiers (later and more parietal) were somewhat consistent with the timing and location of the parietal ERP old/new effect only in the color-source experiments. In the location-source experiments, the (SC-SI) classifier had significant contributions from both early (<500 ms) and late (500–800) time periods and frontal and parietal locations. This suggests that while the SI-CR classifier may be representative of the early frontal old/new effect and the SC-SI classifier representative of the later parietal old/new effect when color is the source information, the mapping is not as

appropriate when location is the source information. This is consistent with Mollison and Curran's (2012) observations suggesting that familiarity contributes to source recognition for location more so than for color. The activation patterns corresponding to the SC-CR classifier took advantage of the features across all time periods (see **Figure 8**) which most likely resulted in the largest distinction between the SC and CR condition.

Additionally, the multivariate classification approach showed that trial-by-trial variation in EEG corresponding to these ERP components are predictive of subjects' behavioral

FIGURE 9 | (A) The scores of all conditions across time by SC-CR classifiers. **(B)** The scores of all conditions across time by SI-CR classifiers. **(C)** The scores of all conditions across time by SC-SI classifiers.
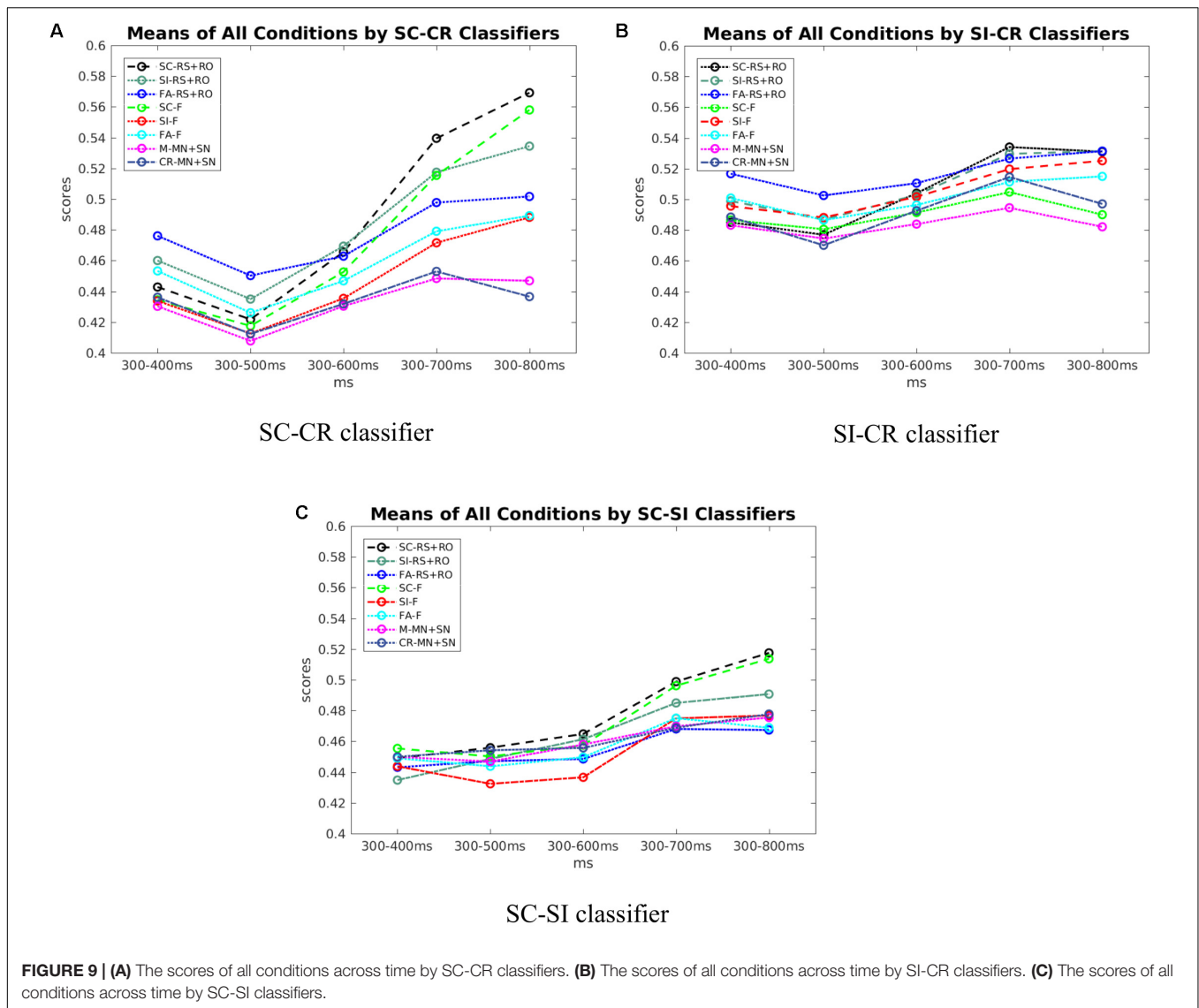
responses, which is consistent with the hypothesis that the underlying processes are influencing memory judgments. One previous study has similarly used logistic regression to predict performance on a city-size comparison task from single-trial EEG data corresponding to the FN400 (Rosburg et al., 2011). Their results showed that the relative familiarity of two cities, as indexed by single-trial FN400 measures, predicted which of the cities subjects judged as being more populous. Taken together with the current results, these classification approaches are important for establishing that EEG patterns which have been related to familiarity and recollection in ERP averages, can be shown to predict behavior on individual trials in both standard memory tasks as well as a decision making task that is influenced by memory. Overall, this strengthens the hypothesized links between these EEG patterns and behaviorally relevant memory processes.

The ERP studies of recognition memory often exclude error trials from analyses because of insufficient trials for stable

ERPs in these conditions. In their original study, Mollison and Curran (2012) excluded subjects with less than 15 artifact-free trials/condition/subject and 24% of subjects would have been excluded if errors were included in the analyses. One approach for increasing the false alarm rate has been to use lures that are similar to studied items (e.g., Curran, 2000; Curran and Cleary, 2003; Nessler et al., 2001). In these cases subjects are presumed to have a high false alarm rate because similar lures are as familiar as studied items, and the familiarity-related FN400 responds similarly to hits and false alarms to similar lures. It is also common to hypothesize that false alarms to even non-similar lures are driven by familiarity. For example, the Yonelinas (1994, 1997) dual process model of ROC curves explicitly assumes that recollection does not contribute to false alarms, which are only driven by familiarity. Few ERP studies have assessed false alarms from lures that were not similar to the studied items. If familiarity differentiates "no" (CR) and "yes" (FA) responses to new items, the FN400 should be more

TABLE 5 | The difference between the average classifier scores for the SC and FA conditions are given in the top three rows.

Classifier	Exp 1 (loc)	Exp 2 (col)	Exp 3-loc	Exp 3-col	Average
Difference between SC and FA					
SC-CR	0.0992	0.0630	0.0647	0.0546	0.0704
SI-CR	−0.0220	0.0099	−0.0042	0.0151	−0.0003
SC-SI	0.0832	0.0456	0.0479	0.0241	0.0502
Difference between SI and FA					
SC-CR	0.0121	0.0353	0.0135	0.0254	0.0216
SI-CR	−0.0173	0.0258	−0.0021	0.0096	0.0040
SC-SI	0.0244	0.0171	0.0121	0.0165	0.0175

The difference between the average classifier scores for the SI and FA conditions are given in the bottom three rows. Negative values indicate FA had larger values. The table shows that FA is mapped closer to SC and SI in the SI-CR classifier than the SC-CR and SC-SI classifiers.

TABLE 6 | The uncorrected pairwise comparison results for the five subjective rating options across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col].

	RS vs. RO	RS vs. F	RS vs. MN	RS vs. SN	RO vs. F	RO vs. MN	RO vs. SN	F vs. MN	F vs. SN	MN vs. SN
SC-CR	0.0031227	6.83E-27	8.57E-92	8.08E-158	5.62E-13	4.19E-66	2.31E-117	3.52E-28	1.97E-47	0.0017659

Only the trials with correct item judgments were included in this analysis.

positive to FA trials than CR trials. Although early studies that did not clearly differentiate the FN400 reported no differences between hits and false alarms (Wilding et al., 1995; Wilding and Rugg, 1996, 1997; Rubin et al., 1999), two studies that specifically focused on the FN400 did observe more positive FN400s to FA than to CR trials (Finnigan et al., 2002; Wolk et al., 2006). Wolk et al. (2006) included a very large number of test items, which resulted in an average of 105 FA trials/subject, but Finnigan et al. (2002) only averaged 12 trials/subject. The current multivariate analysis approach using pattern classifiers addresses this trial count issue by projecting the high dimensional EEG data onto a one-dimensional vector which is meaningful with respect to the experimental paradigm. The SI-CR classifier responded more strongly to FA trials than to CR, with FA being more similar to item hits (SC and SI), as would be expected if FA trials were driven by familiarity.

The relationship between the SC and FA conditions was particularly interesting. The difference between the two conditions were consistently larger across all four datasets on the SC-CR and SC-SI planes compared to the SI-CR plane as given in **Table 5** (and shown in **Figure 5**). This pattern was also evident between the SI and FA conditions, however, the distances between these two conditions were closer. Hence the representations with respect to the different classification boundaries suggest that SC and FA are more similar to each other on the SI-CR (item memory) plane compared to the other two representations. In other words, false alarms (on item information) may include information related to item retrieval while they do not include much information related to source retrieval (recollection).

The other type of error, misses (M), were generally similar to CR in all three classifiers. Both of these conditions reflect low levels of familiarity and recollection that lead to "no" responses. Previous studies have found 300–500 ms FN400 or 500–800 ms parietal old/new differences between hits and misses, but not

between CR trials and misses (Rugg et al., 1998; Curran and Hancock, 2007). Instead, Rugg et al. (1998) found differences between misses and CR were observed over posterior channels between 300 and 500 ms. The latter differences were interpreted as reflecting the activity of an implicit memory process because subjects were giving the same explicit "no" response to both old and new items, but the brain was still differentiating their memory status [although others dispute this definition of implicit memory, Voss and Paller (2008)]. Because our classifiers were trained to differentiate different levels of explicit memory, it makes sense that no major differences were observed between misses and CR in any of our results. Future work could be done to further investigate any differences by specifically involving misses in the classification training [see for example (Noh et al., 2014b)].

In summary, the present results showed that the classification analysis successfully extracts information related to retrieval strength from the EEG data. These results show that the classifier scores well represent the subjects' behavioral performance on source retrieval (the relationship between the SC, SI, and CR conditions in **Figure 5**) and indicate that EEG item-memory and source-memory responses may be more spatially widespread than previously thought and differ between source-types. The results also indicate that retrieval strength as reflected in the classifier scores follows the subjects' subjective ratings (**Figure 6** and **Table 6**). It was also found that the brain activity related to item memory/familiarity may be present during false item retrieval (FA trials) as well as during correct item retrieval (SC and SI trials).

AUTHOR CONTRIBUTIONS

TC and MM planned the EEG experiments. MM collected the data. EN, MM, TC, and VdS planned the initial analyses in this work. KL, EN, and VdS planned the temporal evolution and cluster analysis tests. EN and KL implemented the analyses.

All authors were involved in drafting and editing the work and are accountable for all aspects of the work.

REFERENCES

Agresti, A., and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences in proportions result from adding two successes and two failures. *Am. Stat.* 54, 280–288.

Anderson, J. R., Zhang, Q., Borst, J. P., and Walsh, M. M. (2015). The discovery of processing stages: extension of sternberg's method. *Psychol. Rev.* 123, 481–509. doi: 10.1037/rev0000030

Blankertz, B., Lemm, S., Treder, M. S., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of ERP components – a tutorial. *Neuroimage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Brady, T., Konkle, T., Alvarez, G., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci. U.S.A.* 105, 14325–14329. doi: 10.1073/pnas.0803390105

Curran, T. (2000). Brain potentials of recollection and familiarity. *Mem. Cogn.* 28, 923–938. doi: 10.3758/BF03209340

Curran, T., and Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cogn. Brain Res.* 15, 191–205. doi: 10.1016/S0926-6410(02)00192-1

Curran, T., and Hancock, J. (2007). The FN400 indexes familiarity-based recognition of faces. *Neuroimage* 36, 464–471. doi: 10.1016/j.neuroimage.2006.12.016

Dong, L., Li, F., Liu, Q., Wen, X., Lai, Y., Xu, P., et al. (2017). MATLAB toolboxes for reference electrode standardization technique (REST) of scalp EEG. *Front. Neurosci.* 11:601. doi: 10.3389/fnins.2017.00601

Finnigan, S., Humphreys, M. S., Dennis, S., and Geffen, G. (2002). ERP 'old/new' effects: memory strength and decisional factor(s). *Neuropsychologia* 40, 2288–2304. doi: 10.1016/S0028-3932(02)00113-6

Hammon, P. S., and de Sa, V. R. (2007). Pre-processing and meta-classification for brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 54, 518–525. doi: 10.1109/TBME.2006.888833

Hammon, P. S., Makeig, S., Poizner, H., Todorov, E., and de Sa, V. R. (2008). Predicting reaching targets from human EEG. *IEEE Signal Process. Mag.* 25, 69–77. doi: 10.1109/MSP.2008.4408443

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Hochberg, Y., and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York, NY: John Wiley & Sons, Inc. doi: 10.1002/9780470316672

Jafarpour, A., Fuentemilla, L., Horner, A. J., Penny, W., and Duzel, E. (2014). Replay of very early encoding representations during recollection. *J. Neurosci.* 34, 242–248. doi: 10.1523/JNEUROSCI.1865-13.2014

Krauledat, M., Schröder, M., Blankertz, B., and Müller, K.-R. (2007). Reducing calibration time for brain-computer interfaces: a clustering approach. *Adv. Neural. Inf. Process Syst.* 19, 753–760.

Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* 88, 365–411. doi: 10.1016/S0047-259X(03)00096-4

Lei, X., and Liao, K. (2017). Understanding the influences of EEG reference: a large-scale brain network perspective. *Front. Neurosci.* 2017:205. doi: 10.3389/fnins.2017.00205

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *J. Neural Eng.* 4, R1–R13. doi: 10.1088/1741-2560/4/2/R01

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Mollison, M. V., and Curran, T. (2012). Familiarity in source memory. *Neuropsychologia* 50, 2546–2565. doi: 10.1016/j.neuropsychologia.2012.06.027

Mousavi, M., Koerner, A. S., Zhang, Q., Noh, E., and de Sa, V. R. (2017). Improving motor imagery BCI with user response to feedback. *Brain Comput. Interfaces* 4, 74–86. doi: 10.1080/2326263X.2017.1303253

Nessler, D., Mecklinger, A., and Penney, T. (2001). Event related brain potentials and illusory memories: the effects of differential encoding. *Brain Res. Cogn. Brain Res.* 10, 283–301. doi: 10.1016/S0926-6410(00)00049-5

Noh, E., and de Sa, V. R. (2014). "Discriminative dimensionality reduction for analyzing EEG data," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Austin, TX, 1090–1095.

Noh, E., Herzmann, G., Curran, T., and de Sa, V. R. (2014a). Using single-trial EEG to predict and analyze subsequent memory. *Neuroimage* 84, 712–723. doi: 10.1016/j.neuroimage.2013.09.028

Noh, E., Mollison, M. V., Curran, T., and de Sa, V. R. (2014b). "Single-trial identification of failed memory retrieval," in *Proceedings of the 48th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 21–25. doi: 10.1109/ACSSC.2014.7094388

Ratcliff, R., Sederberg, P. B., Smith, T. A., and Childers, R. (2016). A single trial analysis of EEG in recognition memory: tracking the neural correlates of memory strength. *Neuropsychologia* 93, 128–141. doi: 10.1016/j.neuropsychologia.2016.09.026

Rosburg, T., Mecklinger, A., and Frings, C. (2011). When the brain decides: a familiarity-based approach to the recognition heuristic as evidenced by event-related brain potentials. *Psychol. Sci.* 22, 1527–1534. doi: 10.1177/0956797611417454

Rubin, S. R., Van Petten, C., Glisky, E. L., and Newberg, W. M. (1999). Memory conjunction errors in younger and older adults: event-related potential and neuropsychological evidence. *Cogn. Neuropsychol.* 16, 459–488. doi: 10.1080/026432999380889

Rugg, M. D., and Curran, T. (2007). Event-related potentials and recognition memory. *Trends Cogn. Sci.* 11, 251–257. doi: 10.1016/j.tics.2007.04.004

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., and Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature* 392, 595–598. doi: 10.1038/33396

Schaefer, J., and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4. doi: 10.2202/1544-6115.1175

Tsivilis, D., Allan, K., Roberts, J., Williams, N., Downes, J. J., and El-Deredy, W. (2015). Old-new ERP effects and remote memories: the late parietal effect is absent as recollection fails whereas the early mid-frontal effect persists as familiarity is retained. *Front. Hum. Neurosci.* 9:532. doi: 10.3389/fnhum.2015.00532

Tucker, D. M. (1993). Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalogr. Clin. Neurophysiol.* 87, 154–163. doi: 10.1016/0013-4694(93)90121-B

Velu, P. D., and de Sa, V. R. (2013). Single-trial classification of gait and point movement preparation from human EEG. *Front. Neurosci.* 7:84. doi: 10.3389/fnins.2013.00084

Voss, J. L., and Paller, K. A. (2008). Brain substrates of implicit and explicit memory: the importance of concurrently acquired neural signals of both memory types. *Neuropsychologia* 46, 3021–3029. doi: 10.1016/j.neuropsychologia.2008.07.010

Wilding, E. L. (2000). In what way does the parietal ERP old/new effect index recollection? *Int. J. Psychophysiol.* 35, 81–87. doi: 10.1016/S0167-8760(99)00095-1

Wilding, E. L., Doyle, M. C., and Rugg, M. D. (1995). Recognition memory with and without retrieval of context: an event-related potential study. *Neuropsychologia* 33, 743–767. doi: 10.1016/0028-3932(95)00017-W

Wilding, E. L., and Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Int. J. Psychophysiol.* 119, 889–905.

Wilding, E. L. and Rugg, M. D. (1997). Event-related potentials and the recognition memory exclusion task. *Neuropsychologia* 35, 119–128. doi: 10.1016/S0028-3932(96)00076-0

Wolk, D. A., Schacter, D. L., Lygizos, M., Sen, N. M., Holcomb, P. J., Daffner, K. R., et al. (2006). ERP correlates of recognition memory: Effects of retention interval and false alarms. *Brain Res.* 1096, 148–162. doi: 10.1016/j.brainres.2006.04.050

Yao, D. (2001). A method to standardize a reference of scalp EEG recordings to a point at infinity. *Physiol. Meas.* 22, 693–711. doi: 10.1088/0967-3334/22/4/305

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 1341–1354. doi: 10.1037/0278-7393.20.6.1341

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: the contribution of recollection and familiarity. *Mem. Cognit.* 25, 747–763. doi: 10.3758/BF03211318

Spectral Entropy Can Predict Changes of Working Memory Performance Reduced by Short-Time Training in the Delayed-Match-to-Sample Task

Yin Tian *†, Huiling Zhang†, Wei Xu, Haiyong Zhang, Li Yang, Shuxing Zheng and Yupan Shi

Bio-information College, Chongqing University of Posts and Telecommunications, Chongqing, China

Spectral entropy, which was generated by applying the Shannon entropy concept to the power distribution of the Fourier-transformed electroencephalograph (EEG), was utilized to measure the uniformity of power spectral density underlying EEG when subjects performed the working memory tasks twice, i.e., before and after training. According to Signed Residual Time (SRT) scores based on response speed and accuracy trade-off, 20 subjects were divided into two groups, namely high-performance and low-performance groups, to undertake working memory (WM) tasks. We found that spectral entropy derived from the retention period of WM on channel FC4 exhibited a high correlation with SRT scores. To this end, spectral entropy was used in support vector machine classifier with linear kernel to differentiate these two groups. Receiver operating characteristics analysis and leave-one out cross-validation (LOOCV) demonstrated that the averaged classification accuracy (CA) was 90.0 and 92.5% for intra-session and inter-session, respectively, indicating that spectral entropy could be used to distinguish these two different WM performance groups successfully. Furthermore, the support vector regression prediction model with radial basis function kernel and the root-mean-square error of prediction revealed that spectral entropy could be utilized to predict SRT scores on individual WM performance. After testing the changes in SRT scores and spectral entropy for each subject by short-time training, we found that 16 in 20 subjects' SRT scores were clearly promoted after training and 15 in 20 subjects' SRT scores showed consistent changes with spectral entropy before and after training. The findings revealed that spectral entropy could be a promising indicator to predict individual's WM changes by training and further provide a novel application about WM for brain–computer interfaces.

Keywords: spectral entropy, WM performance, SVR, prediction, classification, BCIs

INTRODUCTION

Working memory (WM) was originally defined as a cognitive mechanism responsible for the temporal maintain and manipulation of new and stored memory information (Baddeley, 2012). WM was considered a limited-capacity, short-term, information retention system (Ma et al., 2014). The original model of WM proposed by Baddeley included three subcomponents: the central

executive, the visuospatial sketch pad and the phonological loop (Baddeley and Hitch, 1994). Later, an additional component, namely the episodic buffer, was added to WM. This component could take information from the other three components and from long-term memory, from which a single episodic representation was created and then temporarily preserved in the buffer (Baddeley, 2000). Individuals exhibited varying abilities in WM. For example, one person was able to memorize more information and manipulate the information more effectively than others (Baddeley, 2003). There is no doubt that if someone is confused, his or her memory ability will decline.

Working memory performance prediction has become an interesting topic which received considerable attention in recent years (Rajji et al., 2015; Johannesen et al., 2016; Tumari and Sudirman, 2016). In previous attempts, numerous approaches were proposed to extract reliable features to predict individual WM performance and distinguish the difficulty levels of cognitive tasks, such as alpha power (Myers et al., 2014), absolute power (Klimesch et al., 2006), and wavelet entropy (Zarjam et al., 2013).

Previous studies related to WM using EEG found that upper alpha event-related desynchronization (ERD) and small power were associated with good performance during actual processing of the task (Klimesch et al., 2006). Autoregressive model (Nai-Jen and Palaniappan, 2004) and wavelet entropy (Zarjam et al., 2013) extracted from EEG signals were used frequently to measure and distinguish the levels of WM task difficulty. Similarly, time-frequency characteristics using wavelet transform were applied to evaluate mental workload in WM with EEG signals. It has been demonstrated that the appearance time and the total power extracted from wavelet analysis were effective features to measure mental workload (Murata, 2005). Combining magnetoencephalography and EEG (MEEG) recordings with source reconstruction techniques showed that synchrony was enhanced with increasing memory loads among the frontoparietal regions during memory retention while the individual WM capacity could be forecasted by phase synchronization in a network among frontoparietal and visual regions (Palva et al., 2010). An event-related functional magnetic resonance imaging (fMRI) study found that better WM performance in a Sternberg-type delayed match WM task could be predicted by greater temporo-parietal junction (TPJ) and default mode network (DMN) deactivation during the encoding period (Anticevic et al., 2010).

In recent years, the rapid development of neuroscience facilitated the improvements of brain–computer interfaces (BCIs) which were information transfer systems transforming brain intention into control commands without involving peripheral neural pathways (Mühl et al., 2014; Putze et al., 2014). Originally, BCIs were designed to provide a new way for patients with impaired motor functions to communicate with others. Generally, motor imagery and evoked visual potentials received a lot of attentions in BCIs (Ferrez and Millán, 2008; George and Lécuyer, 2010). Lately, BCIs have become available to anyone who wants or needs them and have been envisaged to monitor cognitive state, such as attention, fatigue, and emotions (Frey et al., 2013; Mühl et al., 2014, 2015; Chanel and Mühl, 2015).

Moreover, the WM-related EEG signal was utilized in BCIs (Putze and Schultz, 2010; Mühl et al., 2014).

To date, the above measures have widely been investigated in the analysis of the WM loads (i.e., various difficulty levels) (Murata, 2005; Mühl et al., 2014; Dimitriadis et al., 2015), but the estimation or prediction of individual WM performance has rarely been involved especially in BCIs. Previous literatures also tried to find ways to improve one's WM ability which could be reflected by predicting performance in carrying out a wide range of cognitive tasks (Chouhan et al., 2015). To some extent, WM performance might be a symbol of individual's WM ability when carrying out the memory tasks.

Usually memory has been regarded as a personal constant character. However, recent studies revealed that it could be promoted by adaptive and extended training (Klingberg, 2010). The density of cortical dopamine receptors changed after training through test (Mcnab et al., 2009). Previous findings showed that using WM-related fMRI, the training-induced variations were linked to the increased activity in prefrontal and parietal regions (Olesen et al., 2003). Research on children with attention deficit hyperactivity disorder (ADHD) also suggested that WM impairments might be overcome by training and stimulant medication on WM (Holmes et al., 2010).

In this work, we attempted to use a proper and objective feature of retention period in WM EEG as a biomarker to predict individual's WM performance. The existing EEG-based studies, to our best knowledge, demonstrated that power spectrum of low frequency resting-state EEG was associated with individual's WM performance. Changes in brain activity could be reflected by the changes of power spectral density (PSD) during performing cognitive tasks (Weiss, 1992; Pachou et al., 2008).

Recently, different entropy concepts have been applied to describe the order state of sequences (Bruhn et al., 2000, 2001; Xu et al., 2013; Zhang et al., 2015). Among them, Shannon entropy has been shown as an effective measure for the predictability of EEG series in describing anesthetic drug effect (Bruhn et al., 2001; Zhang et al., 2015). However, Shannon entropy was not normalized to the total power of EEG. Consequently, the absolute value of Shannon entropy might vary among individual, which hampered the applications in clinical areas. To overcome this drawback, spectral entropy was developed by using the Shannon entropy to the Fourier-transformed signals (Vanluchene et al., 2004). Therefore, the spectral entropies were regarded as features to distinguish high-performance from low-performance groups in WM with SVM classifier and further predicted subjects' individual WM performance by short-time training with support vector regression (SVR) prediction model in the current study. We assume that spectral entropy could be applied to predict individual WM performance and provide a novel approach for the study of BCI in the future.

MATERIALS AND METHODS

Ethics Statement

The experiment was approved by the ethical committee of Chongqing University of Posts and Telecommunications.

FIGURE 1 | An example of the stimulus sequence in the experiment. Each trial started with a fixation cross (0.5° × 0.5°) at the center of the monitor flashing for 50 ms. Next, the memory array with two, four or eight letters (0.5° × 0.5°) was presented for 200 ms. After 3000 ms retention interval, a probe item was presented for 100 ms and subjects responded with a button press, pointing out if the probe item once was presented in the memory array.

Written informed consent was signed prior to participating in the study. Subjects received a monetary compensation after experiment. None of them had cognitive impairments, mental and neurological disorders.

Subjects

Twenty healthy and right-handed male subjects (mean age 21 years old) participated in the experiment. Subjects were asked to keep relaxed and to restrain wide movements as much as possible during the whole experiments. Subjects were requested to perform a continuous delayed match task consisted of two sessions at three levels of task difficulty. The two sessions were exactly equal and the only distinction between the two sessions was that subjects carried out the first tasks without training but they performed the second tasks after WM training.

Stimuli and Design

Figure 1 showed an example of the stimulus sequence. A fixation cross (0.5 × 0.5°; at the center of the monitor) was displayed throughout the entire block of trials. Each trial started with the fixation cross flashing for 50 ms. Following that, the memory array which was randomly consisted of two, four, or eight letters (0.5° × 0.5°) was presented for 200 ms with the same appearing frequency of two, four, and eight letters. After 3000 ms retention interval, the test array was presented on the screen for 100 ms as a probe item. Then subjects responded with a button press, indicating whether the probe item was in the memory array. Subjects pressed the key "F" with their left index fingers for the absence of the probe item from memory array, and pressed the key "J" with their right index fingers for the attendance of the probe item in the memory array. The number between the absence and the attendance of the probe items in the memory array was equal. Response accuracy (RACC) and speed were equally stressed in the instructions.

Subjects were required to maintain central fixation and minimize eye blinks and body motion throughout the recordings.

The experiment consisted of two sessions per subject and each session was composed of 60 trials with three kinds of memory loads (two, four, or eight letters). Stimuli were presented and behavioral results were recorded using E-prime software[1].

EEG Recordings and Preprocessing

EEG recording was accomplished by using a 64-channel NeuroScan system (Quik-Cap, band pass: 0.05–100 Hz, sampling rate: 1000 Hz, impedances < 5 kΩ) with a vertex reference. To monitor ocular movements and eye blinks, Electro-Oculogram (EOG) signals were simultaneously recorded from four surface electrodes, one pair placed over the higher and lower eyelid and the other pair placed 1 cm lateral to the outer corner of the left and right orbit.

The data was re-referenced to the infinity reference (IR) (Yao, 2001; Tian and Yao, 2013) using the software REST[2]. In the study, EEG was segmented from 100 ms before the onset of the memory array to 100 ms after the subjects' response onset. In other words, the retention and retrieval stages of WM were extracted for the following preprocessing. EOG and Electromyography (EMG) were excluded by Blind Source Separation (BSS) (Negro et al., 2016). Other noise was removed by automatic artifact rejection (±100 μν). The data was baseline corrected using 100 ms EEG signal before the memory array onset. Then the EEG recordings were filtered with a band-pass of 0.5–45 Hz. After the above preprocessing, 60 trials were obtained for each subject under three memory loads (two, four, and eight items). The retention stage (3000 ms) of each trial was extracted for subsequent analysis.

Behavioral Analyses

When performing time limit tasks incorporating both reaction time (RT) and RACC, subjects either sacrificed accuracy in

---

[1] http://www.pstnet.com/eprime.cfm
[2] www.neuro.uestc.edu.cn/rest

exchange for response speed, or exchanged response speed for high accuracy (Mordkoff and Egeth, 1993; Breukelen, 2005). Here, we utilized the Signed Residual Time (SRT) scores (Maris and Han, 2012) for speed accuracy trade-off (Schmitt and Scheirer, 1977) to represent subjects' behavioral performance. The SRT scoring rule was defined as follows (Schmitt and Scheirer, 1977; Maris and Han, 2012; van Rijn and Ali, 2017):

$$\sum_i (2\text{RACC}_i - 1)(MT - t_i) \qquad (1)$$

where the subscript i denoted the item index, $\text{RACC}_i = \{0,1\}$ was the response accuracy (RACC: 0 for incorrect and 1 for correct) for a single trial, and MT denoted the maximum allowable response time for item i. t_i denoted response latency. The total scores were simply the sum of the scores in each item. In other words, for a correct response, subjects earned the residual time as score, and for an incorrect response, subjects lost the residual time as score.

For testing the left–right hand effect and memory-load effect on behavioral WM performance, 2 (hand: left vs. right) × 3 (memory load: 2 items vs. 4 items vs. 8 items) repeated measures analysis of variances (ANOVAs) were conducted. The dependent variables were the SRT scores in intra-sessions (session 1, session 2) and the change rates of subjects' SRT scores in inter-sessions (from session 1 to session 2). Greenhouse–Geisser correction was used when the sphericity assumption was violated in repeated measures ANOVAs (Greenhouse and Geisser, 1959), and a factor had more than two levels in the current study to protect against Type I errors (Kisley et al., 2005). In addition, false discovery rate (FDR) was utilized to the correction of multiple comparisons.

Spectral Entropy

Spectral entropy based on Shannon entropy in physics, quantifying the regularity/randomness of the power spectrum during a given period of time, was used to establish the biomarker for WM performance in the current study. The methodological details were similar to those adopted in a previous study on the prediction of BCI performance (Zhang et al., 2015).

The retention period for each trial was extracted to calculate PSD $P_{sd(f)}$ via the Welch's method (Akbar et al., 2016). The PSD of a time series was defined as the distribution of power as a function of frequency. The normalized PSD was defined as the $P_{sd(f)}$ divided by the total power to obtain probability density function.

$$\widehat{Psd}\ (f) = \frac{Psd\ (f)}{\sum_{f=0.5}^{f=45} Psd\ (f)} \qquad (2)$$

where $\widehat{Psd}_{(f)}$ was the normalized PSD of $P_{sd(f)}$. We estimated spectral entropy based on the PSD within 0.5–45 Hz. The entropy of $\widehat{Psd}_{(f)}$ was generated by using the following equation:

$$SEn = -k \sum_{f=0.5}^{f=45} \widehat{Psd}\ (f) \log\ (\widehat{Psd}\ (f)) \qquad (3)$$

where $k = 1$. The base of the logarithm was 10 and the unit of SEn was dit (i.e., Decimal Digit) in the current study (Schneider,

2007). In effect, spectral entropy reflected the uniformity of the power spectrum distribution. The greater the spectral entropy was, the more uniform the power spectral distribution was.

Scalp Spatial Distribution for Highest Relationship

Spearman's rank correlation coefficient was widely applied to measure the monotonic relationships. It was defined according to the following equation (Gautheir, 2001; Sedgwick, 2014):

$$r = 1 - \frac{6 \sum_{i=1}^{m} d_i^2}{m\ (m^2 - 1)} \qquad (4)$$

where each variable was ordered respectively from lowest to highest. d_i was the difference between two ranks for paired variables x_i and y_i. m was the number of data pairs. In the current study, the relationships between SRT and spectral entropy on each channel were separately measured by Spearman's correlation. From these, 64 spearman's coefficients were observed. Then, we constructed the fingerprint figures and the scalp spatial distribution according to the spearman's coefficients on each channel (the processing steps as shown in **Figure 2**). Based on the above method, the electrode site for highest correlation between SRT and spectral entropy was identified to carry out the next steps (see below).

Grouping Rules and Classification

Subjects were divided into two groups according to their standard z scores of SRT. Here, z-score was defined as:

$$z_i = \frac{x_i - \mu}{\sigma} \qquad (5)$$

where x_i $(i = 1,2,3\ldots,20)$ was the SRT scores of the ith subject, μ was the mean of the 20 subjects' SRT scores and σ was the standard deviation of SRT scores between 20 subjects in the current study. Subjects whose z values of SRT were above zero were allocated to the group with high WM performance, and the rest were assigned to the group with low WM performance. The high and low memory performance groups were defined as positives and negatives, respectively.

The evaluation of generalization performance for SVM classifier was a required step after using the SVM to classify the high from low memory performance groups. SVM was developed by Vapnik based on statistics learning theory (SLT) (Netherlands, 2008). As a result of its excellent generalization performance, SVM has been applied in a wide variety of issues, such as text classification, images classification, hand writing recognition and gene classification. Furthermore, SVM had the feature of empirical risk minimization (ERM) and global optimum solution (Netherlands, 2008). If the SVM classifier could reflect the relationship between features and the class labels very well, then the classifier was considered that it could predict the classes of new samples with good performance. Therefore, classification accuracy (CA), sensitivity (SE), specificity (SP), and area under ROC curve (AUC) were utilized to evaluate the classification performance of SVM classifier (Galar et al., 2012). At the same time, leave-one-out cross-validation (LOOCV) was applied to
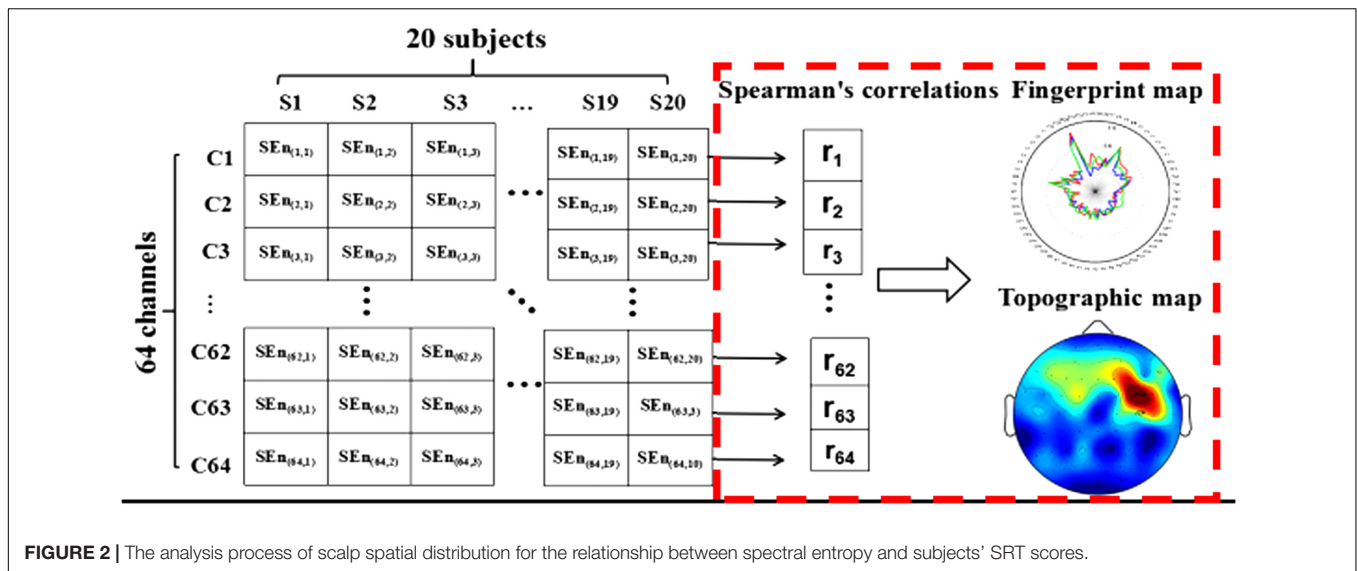
FIGURE 2 | The analysis process of scalp spatial distribution for the relationship between spectral entropy and subjects' SRT scores.

evaluate the generalization performance of SVM for a relatively small sample size in the present study.

(1) The percentage of the number of samples predicted correctly in the test set over the total samples, CA, was calculated as follows:

$$CA = \frac{TP - TN}{TP + TN + FP + FN} \quad (6)$$

where true positive (*TP*) was the number of high-performance samples correctly predicted and true negative (*TN*) was the number of low-performance samples correctly predicted. False positive (*FP*) denoted the number of high-performance samples incorrectly predicted and false negative (*FN*) denoted the number of low-performance samples incorrectly predicted.

(2) SE and SP were calculated by the following formulae, respectively:

$$SE = \frac{TP}{TP + FN} \quad (7)$$

$$SP = \frac{TN}{TN + FP} \quad (8)$$

SE referred to the ratio of correctly classified high-performance samples to the total population of high-performance samples, whereas *SP* was the ratio of correctly classified low-performance samples to the total population of low-performance samples.

(3) AUC was defined as the area under ROC curves, which was discovered and proved to be better than CA to evaluate the predictive performance of classification learning algorithms (Jin and Ling, 2005). Moreover, AUC was indeed a statistically consistent and more discriminating measure than CA (Ling et al., 2003). Originally, the ROC curves were introduced to evaluate machine learning algorithms (Provost et al., 1997). On ROC curves, *TP* was plotted on the *Y* axis and *FP* was plotted on the *X* axis. It described the classifiers' performance across the entire range independent of class distributions (Provost et al., 1997; Jin and Ling, 2005). However, often there was no clear

dominating relation between two ROC curves in the entire range. Therefore, AUC was introduced to provide a good "summary" for the performance of the learning algorithms based on ROC.

To examine the stability of the SVM classification performance, the classification was performed respectively under intra- and inter-session conditions. In the intra-session, LOOCV was utilized separately for the first and second sessions (Vehtari et al., 2016). One subject was picked out as a test sample, while the rest of the subjects were regarded as the train samples for obtaining the classification model. The above steps were repeated until every subject has been allocated as a test sample for one time. During the inter-session prediction, subjects in the first session were regarded as test samples and subjects in the second session were regarded as the training samples. Then subjects in the first session were considered as training samples and subjects in the second session were considered as test samples.

Predicting WM Performance

The subject's spectral entropies derived from WM data at the retention period were selected to predict the corresponding SRT scores. According to the Section "Grouping Rules and Classification," the spectral entropies of the channel achieving the highest correlations between SRT and spectral entropy were applied to SVR to develop predictive model for a relatively small sample size (Ju et al., 2014) respectively in session 1, session 2, and merged session (i.e., session 1 + session 2). Then, LOOCV was utilized to test the stability of the model and the performance of the predictions. Detailed information for the LOOCV procedure for prediction was described as follows:

We assumed that there were n samples in the dataset. One sample was picked out as a testing set, and the rest of the samples were regarded as training sets to develop the predictive model. The above steps were repeated until every subject had been assigned as a test sample for one time and eventually n SVR models were obtained. The correlation coefficients between

FIGURE 3 | Scalp spatial distribution for correlations between the normalized spectral entropy and subjects' SRT scores. **(A)** Fingerprint maps. **(B)** Topographic maps. The color bar denoted the correlation values between SRT scores and spectral entropy. **(C)** Correlations between SRT scores and spectral entropy on channel FC4 separately in session 1, session 2, and merged session.
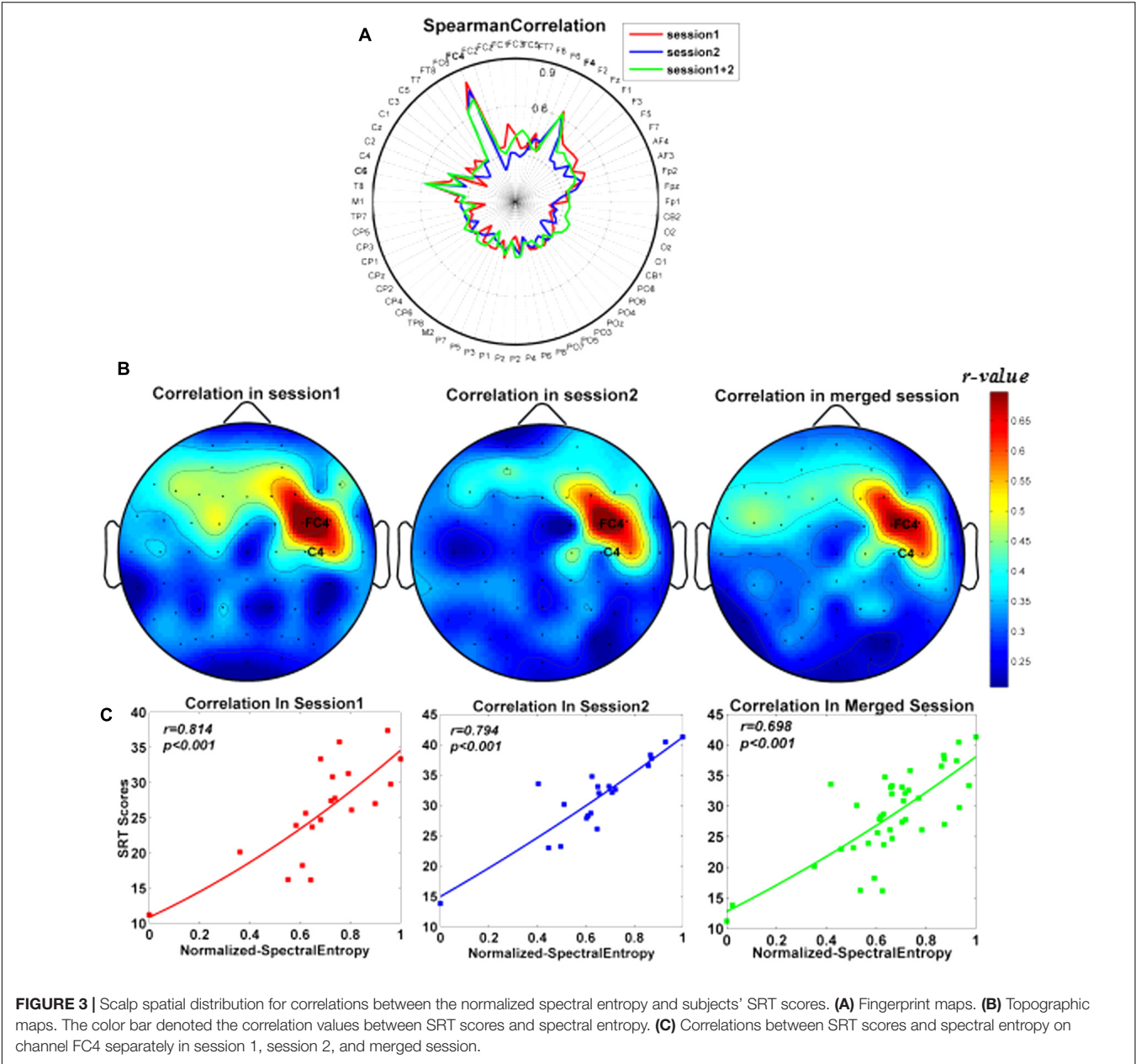
TABLE 1 | The number of subjects in different conditions.

Session		Groups		Changes		Change rates	
		High	**Low**	**Increase**	**Decrease**	**Consistent**	**Inconsistent**
Intra	Session 1	11	9				
	Session 2	12	8				
Inter				16	4	15	5

"Groups" denoted the grouping results about the number of subjects separately in session 1 and session 2 according to standard z scores of subjects' SRT. "Changes" denoted the changes (i.e., SRT scores increased or SRT scores decreased) in the number of subjects after training (i.e., from session 1 to session 2). "Change Rates" denoted the number of subjects whose change rates of SRT showed the consistent and inconsistent change with spectral entropy after training.

TABLE 2 | The classification results of SVM classifier.

		CA	AUC	SE	SP
Intra-session	Session 1	0.950	0.976	1.000	0.900
	Session 2	0.850	0.918	0.800	0.900
Inter-session	Session 1*	0.950	0.973	0.898	1.000
	Session 2*	0.900	0.942	0.900	0.900

CA, classification accuracy; AUC, area under ROC curves; SE, sensitivity; SP, specificity. The session 1 with asterisk (*) denoted the classification results of session 1 with WM data in session 2 as training samples. The session 2 with asterisk (*) denoted the classification results of session 2 with WM data in session 1 as training samples.

predicted and actual SRT scores, together with root-mean-square error of prediction (RMSEP), were calculated to evaluate the prediction performance of the SVR model (Ju et al., 2014; Kichonge et al., 2015).

Change Rates of SRT Scores and Spectral Entropy between Sessions

For a single subject, behavioral scores usually varied by training. Thus we attempted to study whether the spectral entropy of WM data at retention stage could predict changes in memory performance before and after training.

Since the subjects' SRT scores and spectral entropy were in different unit scales, a new measure called Change Rate (CR) was defined as follow (Zhang et al., 2015):

$$CR = 2 \times (TA - TB)/(TA + TB) \times 100\% \qquad (9)$$

where TB denoted the subjects' SRT scores or spectral entropy of WM data at retention period before training (i.e., session 1). TA was the subjects' SRT scores or spectral entropy predictors at retention period after training (i.e., session 2).

RESULTS

Behavioral Performance

For SRT scores, the significant main effects of memory load were separately observed in intra-sessions (session 1: $F = 218.99$, $P < 0.001$, $\eta_p^2 = 0.920$; session 2: $F = 162.32$, $P < 0.001$, $\eta_p^2 = 0.895$), indicating that different memory-load tasks affected the behavioral performance on SRT scores regardless of before training and after training. Non-significant main effects of hand were found in intra-sessions (session 1: $F = 0.435$, $P = 0.518$, $\eta_p^2 = 0.01$; session 2: $F = 1.51$, $P = 0.234$, $\eta_p^2 = 0.074$), indicating that SRT scores were not affected by left–right hand effect in intra-sessions. There were non-significant interactions between hand and memory load (session 1: $F = 4.24$, $P = 0.053$, $\eta_p^2 = 0.183$; session 2: $F = 0.92$, $P = 0.405$, $\eta_p^2 = 0.046$).

For the change rates of SRT scores from session 1 to session 2, there were non-significant main effects of hand ($F = 3.636$, $P = 0.072$, $\eta_p^2 = 0.161$) and memory loads ($F = 0.028$, $P = 0.973$, $\eta_p^2 = 0.002$), indicating that the change rates reduced by training were not affected by both hand and memory-load effects. There were non-significant interactions between hand and memory load ($F = 3.99$, $P = 0.060$, $\eta_p^2 = 0.169$).

Relationship between SRT Scores and Spectral Entropy

The correlations between SRT scores and the spectral entropy on 64 channels were illustrated in the fingerprint map (**Figure 3A**) and the scalp topographic map (**Figure 3B**). Among them, the spectral entropy on channel FC4 showed the strongest correlation with subjects' SRT scores (session 1: $r = 0.814$, $P < 0.001$; session 2: $r = 0.761$, $P < 0.001$; and merged session: $r = 0.698$, $P < 0.001$; FDR correction, and also shown in **Figure 3C**).

Therefore, the spectral entropies on channel FC4 from WM EEG recording at the retention stage were then used as features for SVM classifiers to distinguish the high-performance group from the low-performance group, and further used to predict subjects' SRT scores with SVR prediction model.

Intra- and Inter-Session Classification for SRT

In the first session (i.e., before training), 11 subjects were divided into high-performance group and nine subjects were divided into low-performance group; while for the second session (i.e., after training), 12 of 20 subjects were assigned to high-performance group and eight subjects were assigned to low-performance group (**Table 1**).

The CA was 95 and 85% for the first and second session with spectral entropy features, respectively. The sensitivity (SE) and specificity (SP) at the optimal operating point, as well as the resulting AUC were showed for the first and second session, respectively. The grouping results were showed in **Table 2**.

For intra-session prediction, the resulting CA, AUC, SE, and SP were respectively 0.950, 0.976, 1.000, and 0.900 for session 1 with spectral entropy as classification feature. CA, AUC, SE, and SP were respectively 0.850, 0.918, 0.800, and 0.900 for session 2 with spectral entropy as classification feature.

For inter-session classification, CA, AUC, SE, and SP were respectively 0.950, 0.973, 0.898, and 1.000 for session 1 with session 2 as training samples. The resulting CA, AUC, SE, and SP were respectively 0.900, 0.942, 0.900, and 0.900 for session 2 with session 1 as training samples.

The classification results were illustrated in **Figure 4** and **Table 2**. The red line demonstrated the ROC curves of session 1, and the blue line showed the ROC of session 2.

SRT Predicted by the Spectral Entropy

Support vector regression prediction model and LOOCV revealed that the spectral entropy could be utilized to predict
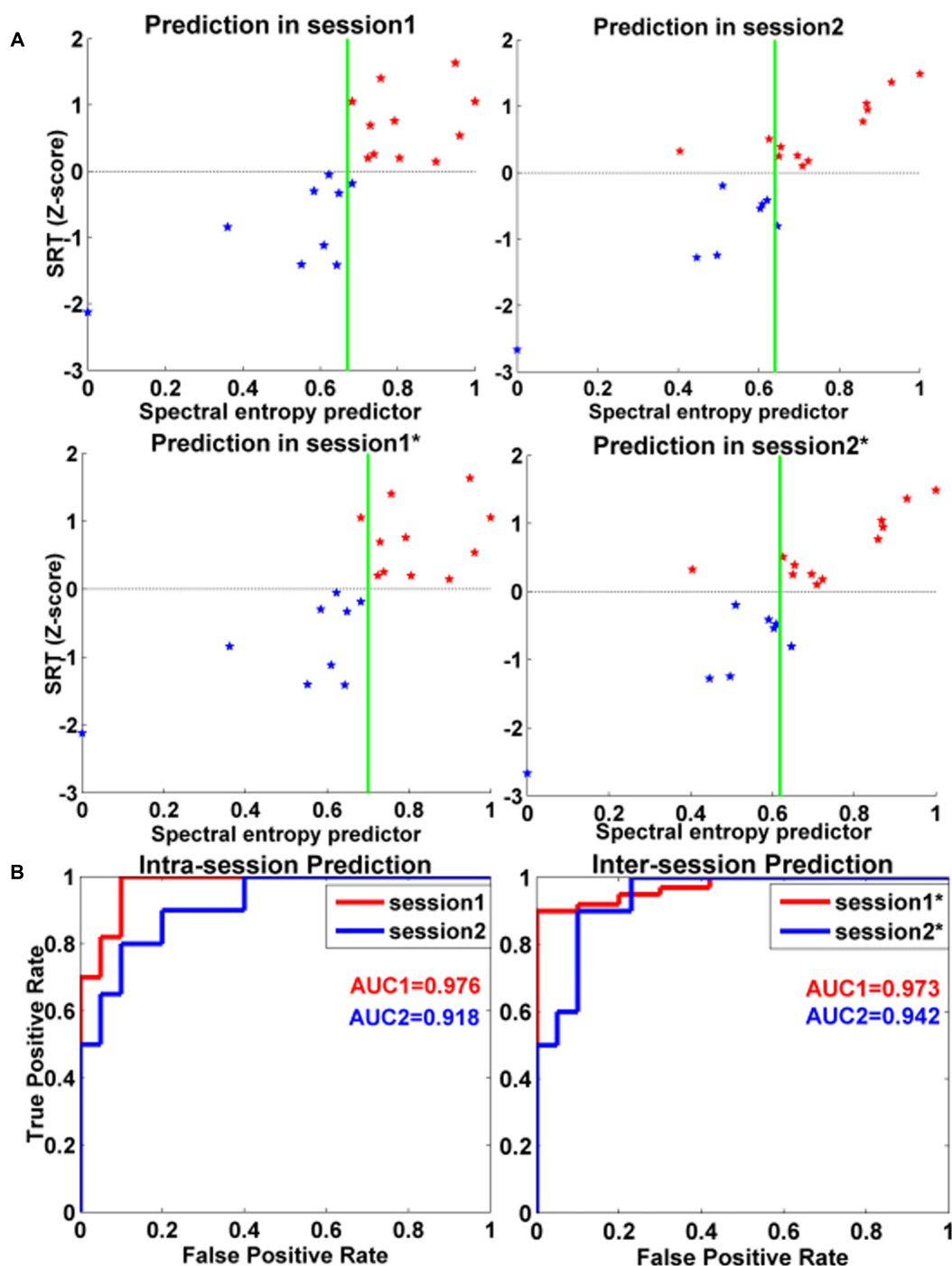
FIGURE 4 | Intra- and inter-session classifications for distinguishing high-performance and low-performance groups. **(A)** Classification plot for intra-session and inter-session. The red star denoted the high-performance group, and the blue star denoted the low-performance group. The green line was the classification boundaries. **(B)** ROC curves for spectral entropy predictors for intra-session and inter-session in classifying the two different performance groups in WM tasks. The session 1 with asterisk (*) denoted the classification results of session 1 with WM data in session 2 as training samples. The session 2 with asterisk (*) denoted the classification results of session 2 with WM data in session 1 as training samples. The abscissa represented the false positive rate, and the ordinate denoted the true positive rate. The red line denoted the ROC curves of session 1, and the blue line represented the ROC curves of session 2. AUC: area under curve.

FIGURE 5 | Prediction performance of SVR model. The figures illustrated the correlations between original SRT scores and predicted SRT scores respectively in session 1, session 2, and merged session.
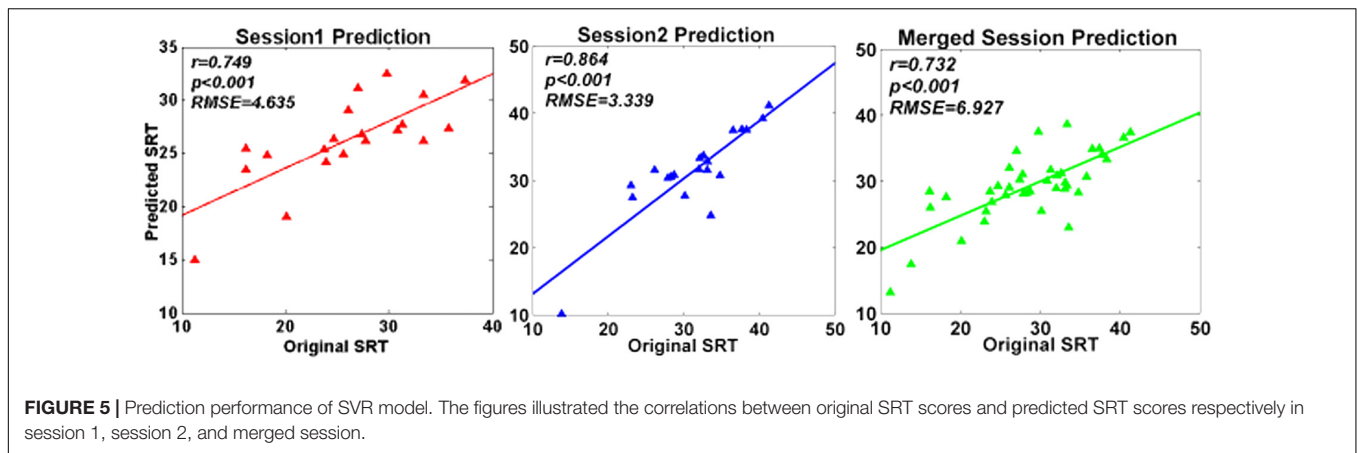
individual WM performance on SRT scores in the current study. The RMSEP after doing LOOCV were 4.635 (session 1), 3.339 (session 2), and 6.972 (merged session), respectively. As illustrated in **Figure 5**, the predicting SRT scores were significantly correlated with the original SRT scores (session 1: $r = 0.749$, $P < 0.001$; session 2: $r = 0.864$, $P < 0.001$; merged session: $r = 0.732$, $P < 0.001$; FDR correction).

Consistent Changes in Spectral Entropy and SRT Scores between Sessions

Within single subject, we explored whether the increased (or decreased) changes in spectral entropy could be predictive of the increased (or decreased) changes in SRT scores. **Figure 6A** showed that 16 in 20 subjects' SRT scores increased by short-time training. **Figure 6B** showed that 15 in 20 subjects' SRT scores consistently varied with spectral entropy predictor before and after training. The results also could be seen in **Table 1**.

DISCUSSION

The present study utilized the spectral entropy to predict individual WM performance changes reduced by short-time training during carrying out the delayed-match-to-sample tasks. We found that: (1) the spectral entropy features extracted from FC4 channel during retention stage of WM was strongly related to subjects' SRT scores; (2) the averaged CA to distinguish high-performance from low-performance group in WM tasks was 90.0 and 92.5% for intra-session and inter-session, respectively; (3) SVR with LOOCV revealed that spectral entropy could predict individual WM performance; (4) 16 out of 20 subjects' SRT scores increased and 15 in 20 subjects' SRT scores were consistent with the changes in spectral entropies by short-time training.

Spatial Distribution for WM Performance

As shown in **Figure 3B**, the spatial distribution of the *r*-values (i.e., the relationship between spectral entropies and behavioral scores) focused on the right frontal area, which was consistent with the previous findings that the frontal cortex might play a prominent part in WM tasks (Gentili et al., 2015; Thürer

et al., 2016). The previous study showed that right inferior frontal junction (rIFJ) as a prefrontal cortex (PFC) control region mediated the causal connection between top–down modulation in the service of attentional goals and WM performance (Gazzaley and Nobre, 2012). Among these channels (**Figure 3A**), the spectral entropy of channel FC4 from WM at retention period generated the strongest correlation ($r = 0.698$) in comparison to other channels in the merged session. Likewise, in the two separate sessions, the highest correlation coefficients also appeared on channel FC4 before training ($r = 0.814$) and after training ($r = 0.794$), respectively. The good classification effect of SVM showed that the spectral entropies on channel FC4 from WM EEG recording at the retention stage might be a dependable biomarker to classify two memory groups successfully.

In the current study, there was no significant effect on the change rates of subjects' SRT scores before and after training when different WM items (i.e., two, four, and eight) were loaded. Thus, subjects' SRT scores were utilized to reflect individual WM performance while carrying out the WM tasks regardless of memory load. As illustrated by **Figure 3C**, the spectral entropy indexes of the subjects were significantly related to individual SRT scores in all sessions including separate sessions and merged session, indicating that spectral entropy indexes might be applied to predict individual WM performance (i.e., SRT scores).

For the further analysis, SVR prediction model combined with LOOCV was established to estimate the predictive ability of spectral entropy indexes on WM performance separately in session 1, session 2, and the merged session (**Figure 5**). The resultant RMSEP after using LOOCV, as well as the high correlation between original SRT scores and predicted SRT scores, demonstrated that prediction models constructed by SVR were effective. The spectral entropy obtained from channel FC4 could be a biomarker to predict individual WM performance.

WM Training

Converging evidence revealed that one's memory ability was absolutely not innate and could be promoted through proper training (Klingberg et al., 2005; Holmes et al., 2010; Klingberg, 2010). Moreover, this improved performance was related to training-induced plasticity from the intracellular level to

functional organization of the cortex for WM (Klingberg, 2010). Vogt et al. (2009) conducted a WM training study on multiple sclerosis. They found that the patient's memory was improved effectively and the spread of bad mood was delayed after receiving the special training for memory (Vogt et al., 2009).
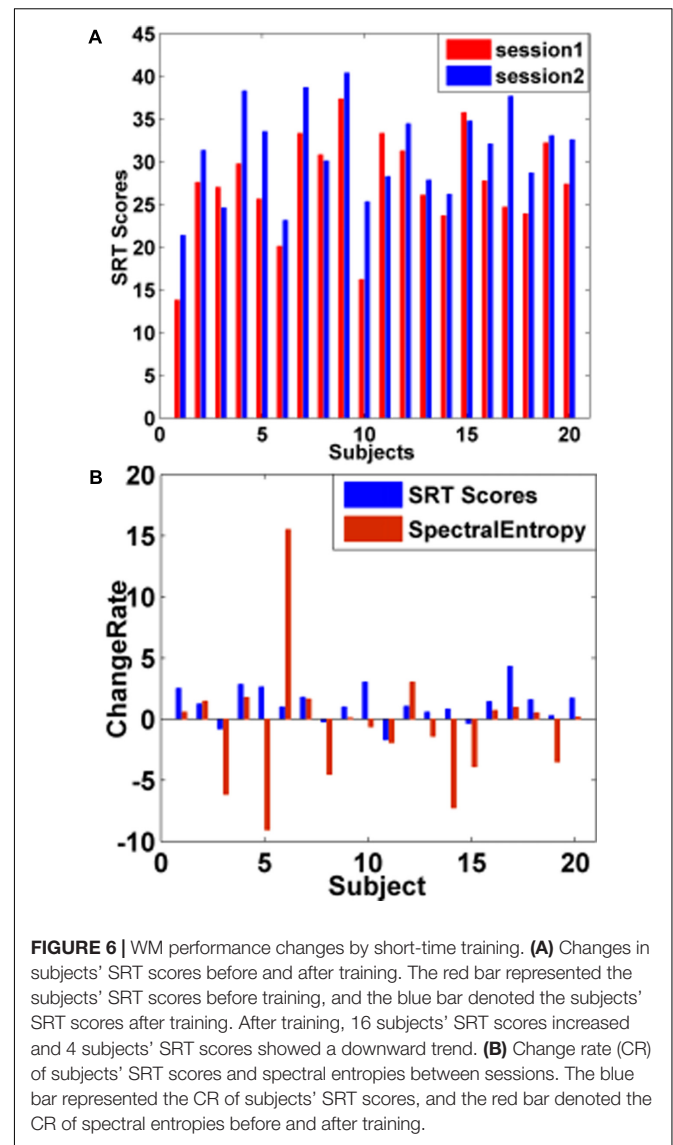
As illustrated by **Figure 6A** and **Table 1**, 16 subjects' behavioral scores were obviously increased after training, indicating that individuals' WM performance could be promoted effectively through training, consistent with the previous study (Klingberg et al., 2005; Holmes et al., 2010; Klingberg, 2010). The CR of 15 in 20 subjects' SRT scores increased (or declined) with the increase (or decline) of the spectral entropy before and after training, which demonstrated that the variations in the spectral entropy predictor could be predictive of the WM performance variations. The findings revealed the consistent changes in SRT scores and spectral entropy by training (**Figure 6B** and **Table 1**).

WM in BCIs

Previously, P300 in motor imagery tasks and steady state evoked visual potentials (SSVEP) were frequently applied in BCIs (Parra et al., 2003; Dal et al., 2009; George and Lécuyer, 2010). Recently, BCIs also were designed to recognize human emotions (Garcia-Molina et al., 2013; Chanel and Mühl, 2015) and to monitor WM load (Sánchez et al., 2015), while there are still little research on the detection of WM performance or individual memory in BCIs. Consequently, it is of great significance to find a reliable feature for the monitor of individual WM performance in BCIs. The development of predictors on WM performance could recognize the potential memory impairment subjects, assisting them in preparation for the positive life and avoiding the frustration from disordered memory. On the other hand, the relevant study might in turn be heuristic for making effective training strategies for those subjects with low WM performance. Moreover, spectral entropy could be a potential instructive biomarker applied in the detection of schizophrenia, depression, ADHD and two-way affective disorder patients and meanwhile provide a new thought for BCIs on the feature extraction of WM.

Limitations

As shown in **Figure 6A**, 16 in 20 subjects' SRT scores obviously was improved after training, while there were still four subjects' SRT scores representing a downward trend which might be induced by individuals' state: either their mental state such as easy to feel tense when carrying out the unfamiliar WM tasks or staying unsuitable to experimental environment, or their physical state such as exhaustion without rest well before experiment or easy to be tired when doing the "dull" experiment. For small sample size in the current study, we just utilized LOOCV and four different evaluation indexes of the classifier's generalization to avoid overfitting as much as possible. It is noted that extracting classification features (like channels) outside of the LOOCV method still could induce overestimation problems. Therefore, we verified the robustness of the selected channel. And the correlation between spectral entropies and WM-related SRT scores of all subjects was highest on channel FC4 regardless of session 1, session 2 and merged session, suggesting the relationship was robust. In addition, spectral entropy was used

FIGURE 6 | WM performance changes by short-time training. **(A)** Changes in subjects' SRT scores before and after training. The red bar represented the subjects' SRT scores before training, and the blue bar denoted the subjects' SRT scores after training. After training, 16 subjects' SRT scores increased and 4 subjects' SRT scores showed a downward trend. **(B)** Change rate (CR) of subjects' SRT scores and spectral entropies between sessions. The blue bar represented the CR of subjects' SRT scores, and the red bar denoted the CR of spectral entropies before and after training.

to predict the individual SRT scores thus indirectly reflecting the WM performance. In the future, a direct way represented the online WM-related BCI would be explored by optimizing classifiers, expanding sample size and improving experimental paradigm.

CONCLUSION

In the current study, we first proposed to use spectral entropy as a feature applied in the classification of WM to distinguish high-performance from low-performance groups in the delayed-match-to-sample task. The resulting RMSEP for the SVR prediction models, as well as the high correlations between original SRT scores and predicted SRT scores, demonstrated that the spectral entropies on channel FC4 could implicitly predict individual WM performance. The changes in the spectral entropy can be predictive of changes in behavioral scores

for individual WM performance. This study could provide theoretical foundation for researchers in the establishment of enhanced training strategies for memory impairment humans BCIs feedback system on memory state with spectral entropy as feature.

AUTHOR CONTRIBUTIONS

Conceived, designed the experiments and wrote the manuscript: YT. Performed the experiments, analyzed the data and wrote the first draft: HuZ. Contributed reagents/materials/analysis tools: WX, HyZ, LY. Discussed the experiment design,

analyzed the data and discussed the experiment results: SZ, YS.

ACKNOWLEDGMENTS

REFERENCES

Akbar, Y., Khotimah, S. N., and Haryanto, F. (2016). Spectral and brain mapping analysis of EEG based on Pwelch in schizophrenic patients. *J. Phys.* 694:012070. doi: 10.1088/1742-6596/694/1/012070

Anticevic, A., Repovs Gshulman, G. L., and Barch, D. M. (2010). When less is more: TPJ and default network deactivation during encoding predicts working memory performance. *Neuroimage* 49, 2638–2648. doi: 10.1016/j.neuroimage.2009.11.008

Baddeley, A. (2000). Working memory: some recent developments. *Int. J. Psychol.* 35, 203–203. doi: 10.1037/10523-117

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Baddeley, A. D., and Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology* 8, 485–493. doi: 10.1037/0894-4105.8.4.485

Breukelen, G. J. P. V. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika* 70, 359–376. doi: 10.1007/s11336-003-1078-0

Bruhn, J., Lehmann, L. E., Röpcke, H., Bouillon, T. W., and Hoeft, A. (2001). Shannon entropy applied to the measurement of the electroencephalographic effects of desflurane. *Anesthesiology* 95, 30–35. doi: 10.1097/00000542-200107000-00010

Bruhn, J., Röpcke, H., and Hoeft, A. (2000). Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia. *Anesthesiology* 92, 715–726. doi: 10.1097/00000542-200003000-00016

Chanel, G., and Mühl, C. (2015). Connecting brains and bodies: applying physiological computing to support social interaction. *Interact. Comput.* 18, 53–59. doi: 10.1093/iwc/iwv013

Chouhan, T., Panse, A., Smitha, K. G., and Vinod, A. P. (2015). "A comparative study on the effect of audio and visual stimuli for enhancing attention and memory in brain computer interface system," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2015)*, Hong Kong, 3104–3109. doi: 10.1109/SMC.2015.539

Dal, S. B., Matteucci, M., and Mainardi, L. (2009). Online detection of P300 and error potentials in a BCI speller. *Comput. Intell. Neurosci.* 2010:307254. doi: 10.1155/2010/307254

Dimitriadis, S. I., Sun, Y., Kwok, K., Laskaris, N. A., Thakor, N., and Bezerianos, A. (2015). Cognitive workload assessment based on the tensorial treatment of EEG estimates of cross-frequency phase interactions. *Ann. Biomed. Eng.* 43, 977–989. doi: 10.1007/s10439-014-1143-0

Ferrez, P. W., and Millán, J. D. R. (2008). "Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy," in *Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course*, Graz, 197–202.

Frey, J., Mühl, C., Lotte, F., and Hachet, M. (2013). "Review of the use of electroencephalography as an evaluation method for human-computer interaction," in *Proceedings of the International Conference on Physiological Computing Systems*, Lisbon, 214–223.

Galar, M., Fernandez, A., and Barrenechea, E. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 42, 463–484. doi: 10.1109/TSMCC.2011.2161285

Garcia-Molina, G., Tsoneva, T., and Nijholt, A. (2013). Emotional brain-computer interfaces. *Int. J. Autonomous Adaptive Commun. Syst.* 6, 9–25. doi: 10.1504/IJAACS.2013.050687

Gauthier, T. D. (2001). Detecting trends using spearman's rank correlation coefficient. *Environ. Forensics* 2, 359–362. doi: 10.1080/713848274

Gazzaley, A., and Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135. doi: 10.1016/j.tics.2011.11.014

Gentili, R. J., Bradberry, T. J., Oh, H., Costanzo, M. E., Kerick, S. E., Contrerasvidal, J. L., et al. (2015). Evolution of cerebral cortico-cortical communication during visuomotor adaptation to a cognitive-motor executive challenge. *Biol. Psychol.* 105, 51–65. doi: 10.1016/j.biopsycho.2014.12.003

George, L., and Lécuyer, A. (2010). "An overview of research on "passive" brain-computer interfaces for implicit human-computer interaction," in *Proceedings of the International Conference on Applied Bionics and Biomechanics (ICABB)*, Venice.

Greenhouse, S. W., and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24, 95–112. doi: 10.1007/BF02289823

Holmes, J., Gathercole, S. E., Place, M., Dunning, D. L., Hilton, K. A., and Elliott, J. G. (2010). Working memory deficits can be overcome: Impacts of training and medication on working memory in children with ADHD. *Appl. Cogn. Psychol.* 24, 827–836. doi: 10.1002/acp.1589

Jin, H., and Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowledge Data Engin.* 17, 299–310. doi: 10.1109/TKDE.2005.50

Johannesen, J. K., Bi, J., Jiang, R., Kenney, J. G., and Chen, C.-M. A. (2016). Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults. *Neuropsychiatr. Electrophysiol.* 2:3. doi: 10.1186/s40810-016-0017-0

Ju, X., Cheng, M., Xia, Y., Quo, F., and Tian, Y. (2014). Support vector regression and time series analysis for the forecasting of Bayannur's total water requirement. *Proc. Comput. Sci.* 31, 523–531. doi: 10.1016/j.procs.2014.05.298

Kichonge, B., John, R. G., Tesha, T., and Iddi, S. N. M. (2015). Prediction of tanzanian energy demand using support vector machine for regression (SVR). *Int. J. Comput. Appl.* 109, 34–39. doi: 10.5120/19172-0643

Kisley, M. A., Davalos, D. B., Engleman, L. L., Guinther, P. M., and Davis, H. P. (2005). Age-related change in neural processing of time-dependent stimulus features. *Cogn. Brain Res.* 25, 913–925. doi: 10.1016/j.cogbrainres.2005.09.014

Klimesch, W., Doppelmayr, M., and Hanslmayr, S. (2006). Upper alpha ERD and absolute power: their meaning for memory performance. *Prog. Brain Res.* 159, 151–165. doi: 10.1016/S0079-6123(06)59010-7

Klingberg, T. (2010). Training and plasticity of working memory. *Trends Cogn. Sci.* 14, 317–324. doi: 10.1016/j.tics.2010.05.002

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlstrom, K., et al. (2005). Computerized training of working memory in children with ADHD-A randomized, controlled trial. *J. Am. Acad. Child Adolesc. Psychiatry* 44, 177–186. doi: 10.1097/00004583-200502000-00010

Ling, C. X., Huang, J., and Zhang, H. (2003). "AUC: a statistically consistent and more discriminating measure than accuracy," in *Proceedings of the* 18th International Joint Conference on Artificial intelligence, Acapulco, 519–524.

Ma, W. J., Husain, M., and Bays, P. M. (2014). Changing concepts of working memory. *Nat. Neurosci.* 17, 347–356. doi: 10.1038/nn.3655

Maris, G., and Han, V. D. M. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y

Mcnab, F., Varrone, A., Farde, L., Jucaite, A., Bystritsky, P., Forssberg, H., et al. (2009). Changes in cortical dopamine D1 receptor binding associated with cognitive training. *Neuroimage* 323, 800–802. doi: 10.1126/science.1166102

Mordkoff, J. T., and Egeth, H. E. (1993). Response time and accuracy revisited: converging support for the interactive race model. *J. Exp. Psychol. Hum. Percept. Perform.* 19, 981–991. doi: 10.1037/0096-1523.19.5.981

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interf.* 1, 66–84. doi: 10.1080/2326263X.2014.912881

Mühl, C., Heylen, D., and Nijholt, A. (2015). "Affective brain-computer interfaces: neuroscientific approaches to affect detection," in *Oxford Handbook of Affective Computing*, eds R. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas (Oxford: Oxford University Press), 217–232.

Murata, A. (2005). An attempt to evaluate mental workload using wavelet transform of EEG. *Hum. Fact.* 47, 498–508. doi: 10.1518/001872005774860096

Myers, N. E., Stokes, M. G., Walther, L., and Nobre, A. C. (2014). Oscillatory brain state predicts variability in working memory. *J. Neurosci. Offic. J. Soc. Neurosci.* 34, 7735. doi: 10.1523/JNEUROSCI.4741-13.2014

Nai-Jen, H., and Palaniappan, R. (2004). "Classification of mental tasks using fixed and adaptive autoregressive models of EEG signals," in *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '04)*, San Francisco, CA, 507. doi: 10.1109/IEMBS.2004.1403205

Negro, F., Muceli, S., Castronovo, A. M., Holobar, A., and Farina, D. (2016). Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation. *J. Neural Eng.* 13:026027. doi: 10.1088/1741-2560/13/2/026027

Netherlands (2008). *Support Vector Machine (SVM)*. Dordrecht: Springer.

Olesen, P. J., Westerberg, H., and Klingberg, T. (2003). Increased prefrontal and parietal activity after training of working memory. *Nat. Neurosci.* 7, 75–79. doi: 10.1038/nn1165

Pachou, E., Vourkas, M., Simos, P., Smit, D., Stam, C. J., Tsirka, V., et al. (2008). Working memory in schizophrenia: an EEG study using power spectrum and coherence analysis to estimate cortical activation and network behavior. *Brain Topogr.* 21, 128–137. doi: 10.1007/s10548-008-0062-5

Palva, J. M., Monto, S., Kulashekhar, S., and Palva, S. (2010). Neuronal synchrony reveals working memory networks and predicts individual memory capacity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 7580–7585. doi: 10.1073/pnas.0913113107

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2003). Response error correction–a demonstration of improved human-machine performance using real-time EEG monitoring. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 11, 173–177. doi: 10.1109/TNSRE.2003.814446

Provost, F. J., Fawcett, T., and Kohavi, R. (1997). "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the Fifteenth International Conference on Machine Learning*, San Francisco, CA, 445–453.

Putze, F., Hesslinger, S., Tse, C. Y., Huang, Y., Herff, C., Guan, C., et al. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Front. Neurosci.* 8:373. doi: 10.3389/fnins.2014.00373

Putze, F., and Schultz, T. (2010). "Utterance selection for speech acts in a cognitive tourguide scenario," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, Chiba, 3014–3017.

Rajji, T. K., Mulsant, B. H., Davies, S., Kalache, S. M., Tsoutsoulas, C., Pollock, B. G., et al. (2015). Prediction of working memory performance in schizophrenia by plasma ratio of clozapine to N-desmethylclozapine. *Am. J. Psychiatry* 172, 579–585. doi: 10.1176/appi.ajp.2015.14050673

Sánchez, A. M., Gaume, A., Dreyfus, G., and Vialatte, F. B. (2015). "A cognitive brain-computer interface prototype for the continuous monitoring of visual working memory load," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, Salerno, 1–5.

Schmitt, J. C., and Scheirer, C. J. (1977). Empirical approaches to information processing: speed-accuracy tradeoff functions or reaction time. *Acta Psychol.* 41, 321–325. doi: 10.1016/0001-6918(77)90025-7

Schneider, T. D. (2007). *Information Theory Primer with an Appendix on Logarithms PDF Version*. Available at: http://alum.mit.edu/www/toms/papers/primer/primer.pdf

Sedgwick, P. (2014). Spearman's rank correlation coefficient. *BMJ* 349:g7327. doi: 10.1136/bmj.g7327

Thürer, B., Stockinger, C., Focke, A., Putze, F., Schultz, T., and Stein, T. (2016). Increased gamma band power during movement planning coincides with motor memory retrieval. *Neuroimage* 125, 172–181. doi: 10.1016/j.neuroimage.2015.10.008

Tian, Y., and Yao, D. (2013). Why do we need to use a zero reference? Reference influences on the ERPs of audiovisual effects. *Psychophysiology* 50, 1282–1290. doi: 10.1111/psyp.12130

Tumari, S. Z. M., and Sudirman, R. (2016). Performance prediction of working memory retention using event-related potential signal towards children. *J. Teknol.* 78, 153–160.

van Rijn, P. W., and Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *Br. J. Math. Statist. Psychol.* 70, 317–345. doi: 10.1111/bmsp.12101

Vanluchene, A. L., Vereecke, H., Thas, O., Mortier, E. P., Shafer, S. L., and Struys, M. M. (2004). Spectral entropy as an electroencephalographic measure of anesthetic drug effect: a comparison with bispectral index and processed midlatency auditory evoked response. *Anesthesiology* 101, 34–42. doi: 10.1097/00000542-200407000-00008

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27:1433. doi: 10.1007/s11222-016-9709-3

Vogt, A., Kappos, L., Calabrese, P., Stöcklin, M., Gschwind, L., Opwis, K., et al. (2009). Working memory training in patients with multiple sclerosis - comparison of two different training schedules. *Restor. Neurol. Neurosci.* 27, 225–235. doi: 10.3233/RNN-2009-0473

Weiss, V. (1992). The relationship between short-term memory capacity and EEG power spectral density. *Biol. Cybern.* 68, 165–172. doi: 10.1007/BF00201438

Xu, P., Hu, X., and Yao, D. (2013). Improved wavelet entropy calculation with window functions and its preliminary application to study intracranial pressure. *Comput. Biol. Med.* 43, 425–433. doi: 10.1016/j.compbiomed.2013.01.022

Yao, D. (2001). A method to standardize a reference of scalp EEG recordings to a point at infinity. *Physiol. Meas.* 22, 693–711. doi: 10.1088/0967-3334/22/4/305

Zarjam, P., Epps, J., Chen, F., and Lovell, N. H. (2013). Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Comput. Biol. Med.* 43, 2186–2195. doi: 10.1016/j.compbiomed.2013.08.021

Zhang, R., Xu, P., Chen, R., Li, F., Guo, L., Li, P., et al. (2015). Predicting inter-session performance of SMR-based brain–computer interface using the spectral entropy of resting-state EEG. *Brain Topogr.* 28, 680–690. doi: 10.1007/s10548-015-0429-3

Different Topological Properties of EEG-Derived Networks Describe Working Memory Phases as Revealed by Graph Theoretical Analysis

Jlenia Toppi[1,2], Laura Astolfi[1,2], Monica Risetti[2], Alessandra Anzolin[1,2], Silvia E. Kober[3,4], Guilherme Wood[3,4] and Donatella Mattia[2]*

[1] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy,
[2] Neuroelectrical Imaging and Brain-Computer Interface Laboratory, Fondazione Santa Lucia IRCCS, Rome, Italy,
[3] Department of Psychology, University of Graz, Graz, Austria, [4] BioTechMed-Graz, Graz, Austria

Several non-invasive imaging methods have contributed to shed light on the brain mechanisms underlying working memory (WM). The aim of the present study was to depict the topology of the relevant EEG-derived brain networks associated to distinct operations of WM function elicited by the Sternberg Item Recognition Task (SIRT) such as encoding, storage, and retrieval in healthy, middle age (46 ± 5 years) adults. High density EEG recordings were performed in 17 participants whilst attending a visual SIRT. Neural correlates of WM were assessed by means of a combination of EEG signal processing methods (i.e., time-varying connectivity estimation and graph theory), in order to extract synthetic descriptors of the complex networks underlying the encoding, storage, and retrieval phases of WM construct. The group analysis revealed that the encoding phase exhibited a significantly higher *small-world* topology of EEG networks with respect to storage and retrieval in all EEG frequency oscillations, thus indicating that during the encoding of items the global network organization could "*optimally*" promote the information flow between WM sub-networks. We also found that the magnitude of such configuration could predict subject behavioral performance when memory load increases as indicated by the negative correlation between Reaction Time and the *local efficiency* values estimated during the encoding in the alpha band in both 4 and 6 digits conditions. At the local scale, the values of the *degree index* which measures the degree of in- and out- information flow between scalp areas were found to specifically distinguish the *hubs* within the relevant sub-networks associated to each of the three different WM phases, according to the different role of the sub-network of regions in the different WM phases. Our findings indicate that the use of EEG-derived connectivity measures and their related topological indices might offer a reliable and yet affordable approach to monitor WM components and thus theoretically support the clinical assessment of cognitive functions in presence of WM decline/impairment, as it occurs after stroke.

Keywords: EEG, brain networks, working memory, sternberg task, connectivity, graph theory

INTRODUCTION

The working memory (WM) is a non-unitary construct that involves the temporary maintenance and manipulation of information either recently acquired or retrieved from long-term storage (Baddeley, 1996). The Baddeley's model is one of the most recognized among the several current models describing the operating principles of WM (D'Esposito and Postle, 2015). It encompasses diverse separable but interacting subsystems such as: 2 unimodal storage sub-systems (phonological loop for verbal material and visuo-spatial sketchpad for visuo-spatial material), a flexible system (central executive) which is responsible for the control and regulation of the storage sub-systems and a multimodal system with limited capacity storage (episodic buffer) that allows the interaction between the various components of WM and the interface with long-term memory (LTM) (Baddeley, 2000, 2010).

The Sternberg Item Recognition Task (SIRT; Sternberg, 1966) has been largely used in cognitive neuroscience to assess WM capacity in terms of storage and data retrieval (Nosofsky et al., 2011; Corbin and Marquer, 2013). It allows for a segregation of encoding, executive maintenance and retrieval processes (not manipulation) regarded as central within the multi-component model of WM. The SIRT is also relatively free from practice effects (Kristofferson, 1972). The SIRT was initially introduced to investigate the neurophysiological processes at the basis of WM by means of indirect behavioral measures (Sternberg, 1966, 1969). Its application was extended later into the field of neuroimaging techniques, functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG) aiming at directly measuring the neural correlates underpinning WM processes (Rypma et al., 1999; Cairo et al., 2004; Payne and Kounios, 2009; Keren-Happuch et al., 2012).

In this regard, several fMRI studies have shown that verbal WM processing in adult humans requires the involvement of a large network of areas which includes bilateral dorso-lateral, prefrontal, left inferior frontal, middle and superior frontal areas, premotor and supplementary motor areas as well as inferior parietal and superior temporal areas, the insula and parts of the cerebellum (Smith and Jonides, 1998; Marvel and Desmond, 2010; Luis et al., 2015). Further studies using SIRT found specific patterns of activation for each of the three phases of WM (encoding, storage, and retrieval) which were also sensitive to WM load levels (Rypma and D'Esposito, 1999; Cairo et al., 2004; Chen and Desmond, 2005; Marvel and Desmond, 2010; Thürling et al., 2012; Vergauwe et al., 2015).

Evidence for specific brain oscillatory responses elicited during the different phases of WM emerged from EEG and MEG studies using the SIRT. In particular, the maintenance (storage) phase of verbal SIRT was associated with oscillatory power in theta (4–8 Hz) predominantly over the frontal midline and left temporal-parietal sites (Payne and Kounios, 2009; Brookes et al., 2011; Kottlow et al., 2015) as well as in alpha (8–13 Hz) power over the parietal midline, the parieto-occipital and left tempo-parietal regions (Jensen et al., 2002; Scheeringa et al., 2009; Heinrichs-Graham and Wilson, 2015; Xie et al., 2016). Changes

in EEG power spectra such as an increase of bilateral frontal delta, frontal-midline theta, and temporo-parietal alpha and a decrease of beta and gamma activities in frontal and occipital areas have been observed as function of WM load (Jensen and Tesche, 2002; Hwang et al., 2005; Payne and Kounios, 2009; Axmacher et al., 2010; Brookes et al., 2011; Roux and Uhlhaas, 2014; Zakrzewska and Brzezicka, 2014; Maurer et al., 2015; Gurariy et al., 2016). Delta power also varies as a function of stimulus type (so-called old-new effect; Kayser et al., 2007; Mathes et al., 2012).

To fully understand brain functions, functional neuroimaging methods have been also applied to investigate the dynamics within networks of brain areas that underlie specific cognitive processes (such as WM), and how a brain damage-induced disruption of neural circuits could account for behavioral impairments (Honey and Sporns, 2008; Cramer et al., 2011; Grefkes and Fink, 2011, 2014). In this regard, functional connectivity estimation was applied to track age-related changes in brain connectivity in a group of children and adolescents performing a modified version of the SIRT (van den Bosch et al., 2014). Task-related networks were identified for encoding (including left motor area, right prefrontal, parietal, and occipital cortex cerebellum) and recognition (including anterior and posterior cingulate cortex, right motor area, cerebellum, left parietal, and prefrontal cortex) phases and their load-induced modulation also correlated with age (Woodward et al., 2013; van den Bosch et al., 2014).

In this study, we take advantage of high temporal resolution of EEG technique and its relatively low-cost and easiness to use to isolate salient descriptors of WM processes as elicited by SIRT, in healthy middle age condition. To this purpose, a combined approach based on EEG-derived connectivity patterns and graph theory (Baccalá and Sameshima, 2001; Milde et al., 2010; Rubinov and Sporns, 2010; Astolfi et al., 2013) was adopted. We expected such combined approach to return quantitative measurements specific for the three different WM phases (encoding, storage and retrieval) and sensitive to different memory workload. The relationship between extracted neurophysiological indices and subject memory performance was also assessed to explore to what extent the estimated EEG networks topology would account for memory behavior. Here, we targeted middle-age population (i.e., between 40 and 50 years) to limit possible confounding effects on the stability of EEG network measures as due to changes in memory task-related neural activity that may emerge from middle age (i.e., fourth decade) onward (Aine et al., 2006, 2011; Grady et al., 2006; Mattay et al., 2006; MacPherson et al., 2014).

The ultimate goal is to provide affordable (EEG-based) computational instruments to measure the electrophysiological dynamics at the level of brain networks that underpin theoretical models of WM processes. As such, this EEG-based network approach could also serve as an objective counterpart of the behavioral assessment of WM impairments which often occur in acquired brain lesions (e.g., stroke) and to ground future cognitive rehabilitative strategy design (Kober et al., 2015, 2017).

MATERIALS AND METHODS

Participants

Seventeen healthy subjects (age: 46 ± 5 years old; 6 males; education: 14.8 ± 3 years; see **Table 1**) were enrolled in the study. All participants except one were right-handed with normal or corrected-to-normal vision. No participant reported a history of neurological or psychiatric diseases; in addition, they were all screened for intake medications and none was receiving any pharmacological treatment affecting cognitive functions. Participants underwent some subtests (Similarities, Information, Coding, Picture Completion, Mosaic Test) from the German adaptation of the Wechsler Adult Intelligence Scale (WAIS III, Von Aster et al., 2006), for a general screening of the cognitive functions and also a deep evaluation of the memory functions. In particular, for the evaluation of the verbal and visuo-spatial memory, subjects performed the Corsi Block Tapping Test (CBTT) (Corsi and Michael, 1972), the Visual and Verbal Memory Test (VVM 2) (Schellig and Schächtele, 2009), the Digit Span (Härting, 2000), the Verbal Learning Memory Test (VLMT), the Nonverbal Learning Test (NVLT), the Verbal Learning Test (VLT) (Sturm and Willmes, 1999). All the subjects obtained normal scores in all the investigated cognitive domains (**Table 1**). This study was carried out in accordance with the recommendations provided in the declaration of Helsinki. All participants provided written informed consent according to the convention of Helsinki. The ethics committee of the University of Graz approved the study. All participants received monetary reward for their participation to the study.

Data Acquisition and Experimental Paradigm

The EEG potentials were recorded from 60 scalp electrodes embedded in a lycra cap, with a left mastoid reference and a ground at Fpz. Horizontal and vertical electro-oculogram (EOG) signals were recorded from 3 electrodes in total, two placed on the outer canthi of the eyes and one below the right eye, respectively. EEG signals were amplified (BrainAmp; Brain Products GmbH, Munich, Germany) and filtered by means of a [0.01–100] Hz band-pass filter prior to digitization at 500 Hz. Electrode impedances were kept below 5 and 10 kOhms for the EEG and EOG recording, respectively.

After 2 min of resting EEG acquisition (of eyes open and close), each subject performed the Sternberg task (Sternberg, 1966). Accordingly, the experimental procedure to deliver the paradigm was as follows (**Figure 1**): first, a series of digits was visually presented to the participants who had to memorize it (encoding phase); then, the participants had to retain the memorized information for a fixed period (storage phase) and finally, participants had to retrieve such memorized content in a brief time interval (retrieval phase). In particular, participants were asked to remember a set of unique digits (between 1 and 9), and then a probe stimulus in the form of a digit was presented. Subjects were instructed to answer, as quickly as possible, whether the probe was in the previously presented set of digits or not. The size of the initial set of digits determined the WM load required to the subject to execute the task (4 digits → easy,

low workload; 6 digits → difficult, high workload). Each trial started with a 2 s presentation of a fixation cross in the middle of the screen. Afterwards, a "memory set" of four (e.g., 5682) or 6 digits (e.g., 146372) was presented (1 s) to allow for memorization (encoding phase). The presentation of the digit series was then followed by a fixation cross, displayed for 2 s (storage period). A single probe digit was then presented for 250 ms (retrieval phase) followed by a fixation cross presented for 1,250 ms. Afterwards, the question "yes or no?" appeared on the screen (maximum duration of 1,500 ms) and the participant had to answer: if the probe was a member of the preceding digit series, the participant had to press the left button (Target Condition) whereas, the right button had to be pressed in case the probe was not part of the series (NoTarget Condition). A new trial would start upon participant answer. The probability of Target condition (36 trials in total) was 0.5 and the digits contained in each trial were presented in a randomized order. The conditions 4/6 digits and Target/NoTarget were also randomized along the recording session.

Behavioral Data

We collected reaction times (RT) and the percentage of correct answers for each subject and each condition (Target/NoTarget and 4–6 digits). To examine any effect of task-related complexity and task-related trials on the subject behavioral performances, 2 separate two-way repeated measures ANOVAs with digits number (DIGITS; 4 or 6) and target type (TYPE; Target/NoTarget) as within main factors were performed considering the percentage of correct answers and reaction times (RTs) parameters as dependent variables.

Data Pre-processing

EEG signals were downsampled to 100 Hz (with anti-aliasing low-pass filter) to optimize the following connectivity analysis and then band-pass filtered in the range (1–45) Hz in order to isolate the EEG spectral content of interest. Independent Component Analysis (ICA—fastICA algorithm) was used to remove ocular artifacts (i.e., the blinks-related IC was removed on the basis of its temporal content and spatial distribution—mainly located over frontal scalp areas). EEG traces were segmented in relation with the specific timing of the paradigm, (0–4,500) ms (period of interest) according to the onset of the first screen containing the digits series and classified according to different conditions (Target_4digits, NoTarget_4digits, Target_6digits, NoTarget_6digits). Only trials correctly executed were included in the analysis. Residual artifacts were then removed by means of a semi-automatic procedure based on a threshold criterion (±80 µV). Only the artifacts-free trials were used for further analysis (no less than 30 trials per condition were considered for final analysis). The entire pre-processing procedure was performed by means of Brain Vision Analyzer 1.0 software (Brain Products GmbH).

Time-Varying Connectivity Estimation

Pre-processed EEG signals were subjected to a time-varying connectivity estimation process for each subject and each experimental condition (4/6 digits, Target/NoTarget).

TABLE 1 | Table reporting the demographic data of the participants and the results of the cognitive screening.

Code		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Gender		F	M	F	M	F	F	M	F	F	F	M	F	M	M	F	F	F
Age		43	41	49	47	49	43	40	43	44	63	45	56	45	49	42	40	51
Handedness		R	R	R	R	R	R	R	R	R	R	L	R	R	R	R	R	R
Highest education		U	HS	HS	HS	CS	U	U	U	HS	CS	U	HS	U	HS	HS	HS	HS
Cognitive tests	CBTT forward task	10	8	8	8	8	8	8	8	6	8	5	7	10	11	11	6	8
	Digit span forward task	11	12	10	8	11	9	10	13	9	10	10	11	9	13	9	12	11
	VVM2 – City map 1	40	51	49	59	34	51	47	42	47	65	51	43	49	51	51	64	56
	VVM2 – Construction 1	34	55	44	57	41	55	48	46	50	71	50	55	57	60	60	46	58
	CBTT backwards task	6	7	6	6	7	5	7	6	6	6	8	5	8	8	10	7	7
	Digit span backwards task	7	10	7	7	10	7	8	10	10	7	10	10	7	10	9	13	8
	VLT	61	54	43	30	32	34	48	50	46	56	52	57	31	61	63	57	46
	NVLT	46	37	41	47	40	44	43	45	45	43	38	37	37	38	44	58	41
	VLMT T—ΣDg1–5	60	63	51	63	60	63	54	48	67	58	54	60	33	63	67	67	63
	VLMT T—Dg7	52	62	33	62	41	56	44	48	60	52	41	61	33	56	62	52	56
	VLMT T—Dg5-Dg7	40	58	33	58	45	52	45	45	58	49	34	58	40	51	58	45	55
	VLMT T—W-F	54	54	48	54	53	54	53	48	54	53	45	62	33	54	54	48	48
	VVM2—City map 2	54	48	32	59	28	41	41	49	41	72	49	58	41	48	54	60	62
	VVM2—Construction 2	32	54	42	54	35	47	38	47	54	71	56	51	62	56	66	42	48
	WAIS-IV Similarities	57	47	50	63	67	47	57	47	53	60	57	60	57	57	60	57	53
	WAIS-IV Picture completion	50	50	67	63	50	43	50	40	57	53	67	50	57	57	67	50	63
	WAIS-IV Mosaic test/block design	67	60	67	63	53	43	63	40	47	50	60	53	70	50	63	63	60
	WAIS-IV Number-symbol test	40	43	40	67	50	47	50	33	40	70	50	47	73	57	57	40	73
	WAIS-IV Information	57	57	70	57	60	57	60	50	53	63	63	47	63	50	60		60

F, female; M, male; R, right; L, left; U, university; HS, high school; CS, compulsory school.
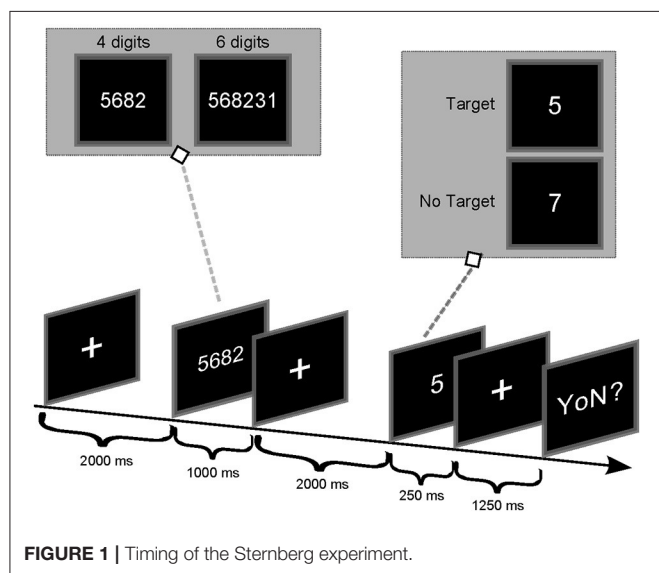
FIGURE 1 | Timing of the Sternberg experiment.

Several time-domain or frequency-domain measures were developed to estimate connectivity from EEG data in terms of correlation/coherence, statistical dependencies, or causal interaction among data series. Some of these measures are deemed to be less prone to volume conduction effects (Haufe et al., 2013), such as imaginary coherence (Nolte et al., 2004) or phase slope index (Nolte et al., 2008) but return non-causal, undirected measures, and are based on a pairwise approach, which results in high rates of false positives when the network complexity increases (Kus et al., 2004). Here, we employed a time-varying adaptation of Partial Directed Coherence (PDC), a spectral multivariate estimator which provides with the directed influences between any given pair of signals in a multivariate data set (Baccalá and Sameshima, 2001; Astolfi et al., 2006; Toppi et al., 2016b). Such time-varying adaptation is based on the General Linear Kalman Filter (GLKF) (Milde et al., 2010) which is able to follow temporal dynamics of brain networks with high temporal resolution in high density EEG data (Toppi et al., 2016a). We used it to estimate the relationships between signals for all frequency samples in the range (1–45) Hz and for all the samples in the time interval (0–4,500) ms.

The connectivity patterns contrasted with the baseline period were estimated for each time sample and averaged in the five frequency bands-of-interest and in three time intervals (periods-of-interest). The frequency bands were individually defined according to the Individual Alpha Frequency (IAF-10 ± 0.9 Hz), as determined by means of the Fast Fourier Transform spectra of 2 min resting EEG (recorded before SIRT execution) over posterior leads (parietal, parieto-occipital, and occipital). The following frequency bands were then considered: Delta (IAF-8/IAF-6), Theta (IAF-6/IAF-2), Alpha (IAF-2/IAF+2), Beta (IAF+2/IAF+14) and Gamma (IAF+15/IAF+30; Klimesch, 1999). The three periods-of-interest correspond to: (0–1,000) ms

(encoding phase); (1,000–3,000) ms (storage phase) and (3,000–4,500) ms (retrieval phase). The analysis was conducted only on Target condition.

Any relevant changes in the time-varying connectivity matrices related to the different experimental conditions were evaluated by means of statistical comparisons (independent samples t-test) performed for each task condition (Target_4digits, NoTarget_4digits, Target_6digits, NoTarget_6digits) between each post-stimulus time window (encoding, storage, retrieval) and the baseline period. The baseline period was the time interval (-1000-0) ms preceding the appearance of the digits series (subjects fixing a cross on the screen). Time samples were used as observations for statistical test. The test was repeated for each frequency band and each subject. The significance level was set at 5%. A False Discovery Rate (FDR) was conducted for multiple comparison correction (Benjamini and Yekutieli, 2001).

The analysis pipeline was performed in Matlab environment (MATLAB, 2011).

Graph Indices

The main global and local properties of the estimated networks were quantified by means of indices derived from the graph theory. Such indices are defined on the basis of a binary adjacency matrix G, obtained by comparing each entry of the connectivity matrix A with its corresponding threshold as follows:

$$G_{ij}(f,t) = \begin{cases} 1 \rightarrow A_{ij}(f,t) \geq \tau_{ij}(f,t) \\ 0 \rightarrow A_{ij}(f,t) < \tau_{ij}(f,t) \end{cases} \quad (1)$$

where G_{ij} and A_{ij} represent the entry (i,j) of the adjacency matrix G and the PDC matrix A, respectively, and τ_{ij} is the corresponding threshold. As mentioned above, we adopted a statistical approach for the threshold definition in order to avoid the attribution of false properties to the networks due to the application of an empirical thresholds (Toppi et al., 2012). Accordingly, the threshold τ_{ij} corresponds to the 95th percentile (corrected for multiple comparisons by FDR) of the PDC distribution obtained for the baseline condition.

The estimated adjacency matrices were then, used to extract local and global indices as described below. To avoid network-size effects, each global index was normalized by its corresponding value obtained from 100 random graphs generated by fixing the connection density of the original network. Random graphs were thus used, as reference level for the description of global properties of WM networks (see below).

Local and global indices were computed by means of routines provided in Brain Connectivity Toolbox developed for Matlab environment (Rubinov and Sporns, 2010).

General Properties of the Network

The human brain can be viewed as a large-scale complex network that is simultaneously segregated and integrated via specific connectivity patterns (Bullmore and Sporns, 2009). We selected the three indices—local and global efficiency and small-worldness—that are widely utilized to describe the general topological properties of a network, thus reflecting the integration and segregation of the information flow between areas (Sporns, 2013b).

Global Efficiency (GE)

The GE is a global measure (considering all the connections in the whole-network) of how efficiently a network exchanges information internally. It is defined as the average of the inverse of the geodesic distance (shortest path between two nodes in the network) and it represents the efficiency of the communication between all the nodes within the network (Latora and Marchiori, 2001). It can be defined as follows:

$$GE = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \quad (2)$$

where N represents the number of nodes in the graph and d_{ij} the geodesic distance between i and j.

Local Efficiency (LE)

The LE is a measure of the fault tolerance of a network. It verifies whether the communication between nodes is still efficient when a node is removed from the network. The higher the LE, the greater the robustness of the network at local scale.

The LE is the average of the global efficiencies computed on each sub-graph S_i belonging to the network and it represents the efficiency of the communication between all the nodes around the node i in the network (Latora and Marchiori, 2001). It can be defined as follows:

$$LE = \frac{1}{N} \sum_{i=1}^{N} E_g(S_i) \quad (3)$$

where N represents the number of nodes in the graph and S_i the sub-graph obtained deleting the ith row and the ith column from the original adjacency matrix.

Small-Worldness (SW)

It has been suggested that the human networks are organized to optimize efficiency, due to a small-world topology allowing simultaneous global and local parallel information processing (Bassett and Bullmore, 2006). SW is a measure of a network global organization in terms of its integration and segregation properties. Small-world topology is typical of networks highly segregated (nodes organized according to clusters) and highly integrated (high communication speed between electrodes).

A network G is defined as small-world network if $L_G \geq L_{rand}$ and $C_G \gg C_{rand}$ where L_G and C_G represent the characteristic path length (Sporns et al., 2004) and the clustering coefficient (Fagiolo, 2007) of a generic graph and L_{rand} and C_{rand} represent the correspondent quantities for a random graph (Watts and Strogatz, 1998). On the basis of this definition, a measure of small-worldness of a network can be introduced as follows:

$$SW = \frac{C_G / C_{rand}}{L_G / L_{rand}} \quad (4)$$

being a *small-world* network if S > 1 (Humphries and Gurney, 2008).

Local Properties of the Network

The topology of the networks was further investigated by computing the degree index for each scalp electrode to characterize the (local) level of in- and out- information flows exchanged within the network.

Degree

The degree of a node is the number of connections involving it. As such the degree is the simplest index identifying *hubs* in graphs. In directed networks, the indegree is the number of inward links and the outdegree is the number of outward links (Sporns et al., 2004). It can be defined as follows:

$$k_f = \sum_{j \in N, j \neq f} g_{fj} + \sum_{i \in N, i \neq f} g_{if} \qquad (5)$$

where k_f is the degree of node f and g_{ij} represents the entry ij of the adjacency matrix G. The degree of a specific electrode was normalized with respect to the network density, in order to capture local changes and not a general increase/decrease of the network density.

All the extracted global and local indices were subjected to a two-way ANOVA with memory phases (PHASES: Encoding, Storage, Retrieval) and digits number (DIGITS: 4, 6) as main *within-subject* factors. Duncan's *post-hoc* test was used to verify differences between the ANOVA levels. FDR was further applied to correct for multiple ANOVAs. Furthermore, to explore the relationship between the indices extracted for each memory phase and the relative behavioral data (correct answers rate, reaction time) a Pearson correlation analysis was performed. FDR was applied to reduce type I errors due to multiple correlations.

RESULTS

Behavioral Results

The overall behavioral data obtained from each subject is reported in **Table 2**. All the participants showed a percentage of correct answers above 80% (except for subject 5 in 6 Digits who was removed from the analysis) and reaction times (RTs) comprised between 250 and 700 ms for the 4 SIRT conditions. The variability ranges observed for the two behavioral parameters are in agreement with literature and comparable with those reported in other studies (Sternberg, 1966; Cummins and Finnigan, 2007; Tuladhar et al., 2007).

The two-way ANOVA revealed that the percentages of correct answers significantly decreased (94 ± 3 to 87 ± 5%) when the subjects were challenged with the condition of 6 digits with respect to 4 [main factor DIGITS $p = 0.00007$, $F_{(1, 15)} = 28.15$, MSE = 672.8]. The RTs were also significantly longer in the condition 6 digits with respect to 4 digits condition [main factor DIGITS $p = 0.021$, $F_{(1, 15)} = 6.54$, MSE = 13336]. Furthermore, the NoTarget condition was associated to significantly longer RTs as compared to those obtained during the Target condition (470 ± 126 vs. 500 ± 130 ms) in the 6 digits case [main factor TYPE $p = 0.013$, $F_{(1, 15)} = 7.83$, MSE = 12269]. Subject 5 was excluded from the EEG analysis because of low accuracy performance (30% error rate).

General Properties of the Networks

The results of the two-way ANOVA for the Local Efficiency (LE), Global efficiency (GE), and Small-Worldness (SW) indices with respect to memory phases and WM load are reported in **Table 3** for the five frequency bands.

As shown in **Figure 2**, the LE index mean value ($n = 16$) relative to alpha band was significantly higher in the *Encoding* with respect to both *Storage* and *Retrieval* phases (**Figure 2A**). An opposite trend was observed for the GE index (**Figure 2B**) that was significantly higher in the *Retrieval* as compared to both *Encoding* and *Storage* (**Figure 2B**). Finally, the SW index (**Figure 2C**) was significantly higher in the *Encoding* as compare with *Storage* and *Retrieval*. Similar significant results were found for the three indices in the other frequency bands (**Table 3**).

We found significant differences between 4 and 6 digits conditions for the LE and the SW indices (**Figures 2A,C**). In particular, the LE and SW showed significantly higher values for the 6 with respect to 4 digits only during *Encoding* in alpha (**Figures 2A,C**). Similar results were found in gamma band (**Table 3**). No significant differences between 4 and 6 digits were found for the GE.

Furthermore, the LE index computed for alpha band and relative to the *Encoding* phase negatively correlated with RTs obtained from both 4 ($r = -0.7026$, $p = 0.0024$) and 6 ($r = -0.7048$, $p = 0.0023$) digits cases.

Local Properties of the Networks

The degree index was computed for each electrode and each subject and then averaged within the experimental group for the three PHASES and the two DIGIT conditions [Grand Average (GA) Degree Maps]. A spatial representation of such index is reported in the topographical maps of **Figure 3** for 4 digits (**Figure 3A**) and 6 digits (**Figure 3B**) cases.

The visual inspection of the GA Degree Maps relative to the 4 digits condition revealed that the three WM phases were associated with distinct connectivity networks for each frequency band oscillation (**Figure 3A**). During the *Encoding*, we observed a connectivity pattern which mainly included (high degree index) the central midline, the bilateral frontal areas and the bilateral posterior areas in the delta and theta frequency bands. In the alpha band, such patterns were mainly represented over the frontal midline, the left frontal areas and the right hemisphere from frontal to parietal areas. In beta and gamma oscillation ranges, the patterns were prevalent over the bilateral fronto-temporal areas.

Storage was consistently associated with a high involvement (high degree) of the bilateral fronto-temporal areas, the frontal midline and the right posterior areas in the delta, theta and alpha bands. Bilateral fronto-temporal areas, left tempo-parietal areas and right central areas have an importal role in the beta band. In gamma band, we found an involvement of bilateral fronto-temporal areas and frontal midline.

The *Retrieval* phase showed a connectivity pattern mainly involving (high degree) frontal-central midline, left fronto-tempo-parietal areas, right frontal areas and occipital areas in the delta, theta, and alpha bands. In beta band, we found a high

TABLE 2 | Mean values of the percentage of correct answers and relative reaction time (RTs) obtained from each participant.

Subj #	Correct answers (%)				Reaction time (ms)			
	Target		NoTarget		Target		NoTarget	
	4 digits	6 digits	4 digits	6 digits	4 digits	6 digits	4 digits	6 digits
1	94	81	94	94	356.91	346.97	422.76	469.35
2	94	78	97	92	373.97	386.68	382.66	444.73
3	100	81	97	86	548.39	468.03	502.80	521.00
4	97	100	94	89	306.37	321.53	305.26	382.34
5	92	72	97	97	411.48	498.77	405.94	511.97
6	94	81	97	89	619.24	632.31	597.43	620.78
7	86	86	94	83	425.81	458.03	483.50	499.70
8	94	86	89	83	473.76	460.26	607.34	554.80
9	83	89	86	78	430.63	509.16	433.10	504.71
10	97	81	92	89	778.11	785.28	704.91	722.66
11	97	97	97	97	319.51	331.20	295.80	371.71
12	97	92	86	94	413.97	347.45	381.87	398.74
13	94	94	92	94	282.38	250.29	265.58	285.88
14	92	81	92	81	444.70	573.41	608.03	595.31
15	97	94	97	89	460.71	474.68	559.97	516.81
16	92	83	94	89	616.64	626.97	580.24	705.78
17	92	83	94	89	616.64	626.97	580.24	705.78
MEAN	94	86	94	89	463.48	476.35	477.50	518.36
STD	4,2	7,6	3,8	5,6	133.01	140.16	128.55	124.12

Missing answers (RT = 0) were excluded.

involvement of bilateral fronto-tempo-parietal areas and parieto-occipital midline. An important role of bilateral fronto-temporal areas and central areas resulted in gamma band.

The averaged patterns obtained for the 6 digits condition are illustrated in **Figure 3B**. The qualitative (visual inspection) analysis of 6 digits condition revealed a general superimposition with the areas mainly involved in the 4 digits condition.

On the basis of these findings (**Figure 3**), we selected eight scalp areas (macro-areas) symmetrically distributed over the left and right sides and computed the respective average degree index. The following macro-areas were considered: Left Frontal (Fp1, AF7, F7), Frontal Midline (AFz, Fz, FCz), Right Frontal (Fp2, AF8, F8), Left Temporal (FT7, T7, TP7), Right Temporal (FT8, T8, TP8), Left Parietal (PO7, O1), Right Parietal (PO8, O2), Occipital (Oz).

The results of the two-way ANOVA on degree index with respect to the memory phases and WM load are reported in **Table 4** for each macro-area and frequency band. The schematic representation of **Figure 4** summarizes the trends obtained for the macro-areas degree across the three memory phases in the five frequency bands (irrespective of DIGIT factor). In particular, the areas distinctive for the Encoding were the bilateral frontal areas in the delta band, the right frontal and right parietal areas in theta band, right parietal area in alpha band, left, and right frontal and right temporal areas in beta and gamma bands. The Storage was instead characterized by right frontal area in both delta and theta bands and frontal midline in gamma band. The retrieval

involved occipital area in alpha, left parietal, and occipital areas in beta band and frontal midline, left parietal and occipital areas in gamma band.

DISCUSSION

This study applied a graph theory—driven approach to complex causality patterns derived from EEG recordings with the aim to identify distinct topological properties of the neural networks associated to encoding, storage, retrieval phases of WM as elicited during visual SIRT performed by healthy, middle age adults. We found that, during the encoding phase, the global network exhibited a small-world topology (in all frequency bands) thus, indicating a network configuration accounting for both global information transfer and local processing. The requirement of such *optimal* configuration specifically for item encoding appeared further corroborated by the negative correlation between local efficiency and behavioral task performance. The small-world configuration of the whole network persisted across maintenance and rehearsal of encoded items but it showed a descendent trend. At the local scale, the degree of information flow between scalp regions was selective for each of the three different WM phases, in that *network hubs* were consistent with the different role of brain regions in different WM phases. Overall, these EEG findings provide evidence that the complexity of functional networks underpinning the model of multicomponent WM (Baddeley, 2012; Eriksson et al., 2015)

Global index	Band	Phases (d.f. = 2,30)	Digits (d.f. = 1,15)	Phases × digits (d.f. = 2,30)
Local efficiency	δ	**29.24****	**4.05***	2.81
	θ	**27.58****	0.02	0.01
	α	**30.38****	**4.66***	**3.4***
	β	**79.55****	0.11	0.3
	γ	**177.77****	**9.62***	**8.75***
Global efficiency	δ	**41.35****	**5.48***	**4.68***
	θ	**21.51****	1.2	0.21
	α	**39.31****	0.03	0.58
	β	**10.64****	2.98	0.51
	γ	**9.19****	0.95	1.81
Small-Worldness	δ	**57.44****	2.09	2.16
	θ	**65.32****	0.71	1.1
	α	**48.84****	**4.56***	**4.07***
	β	**148.16****	0.07	0.002
	γ	**122.52****	0.61	0.51

FDR correction for multiple ANOVAs was applied. Significant results are highligthed in bold.

may be represented by synthetic EEG indices which preserve the selectivity of the dynamics occurring during encoding, maintenance and rehearsal of memory items.

Behavioral Results

The behavioral results obtained from our sample of healthy, middle age adults are in line with what was reported by previous studies conducted in healthy adults and where the SIRT was applied to investigate working memory (WM) processes (Sternberg, 1966; Cummins and Finnigan, 2007; Tuladhar et al., 2007). As expected, WM loads (4, 6 digits conditions) had a significant effect on the response time and accuracy for both Target and NoTarget probes in our sampled population. These WM load-related behavioral changes have been previously ascribed to a serial scanning of memorized elements required in order to recall the memorized material (Majerus et al., 2006).

Global Organization of the WM Network

The complex human brain networks have been found to have a "small-world" topology which is characterized by a high local specialization and a high global integration to sustain a high efficiency at a low wiring cost (Sporns, 2013a).

The significant effect of phase factor on all the global indices (global and local efficiency, small-worldness; **Table 3**) indicated that a small-word topology of the networks was present in all three WM phases with the highest value associated with encoding with respect to storage and retrieval (**Figure 2**; in alpha band). Such descendent trend was evident for lower to higher frequency oscillations (**Table 3**).

The finding of a descendent trend in the network small-worldness in all frequency bands may reflect a general network
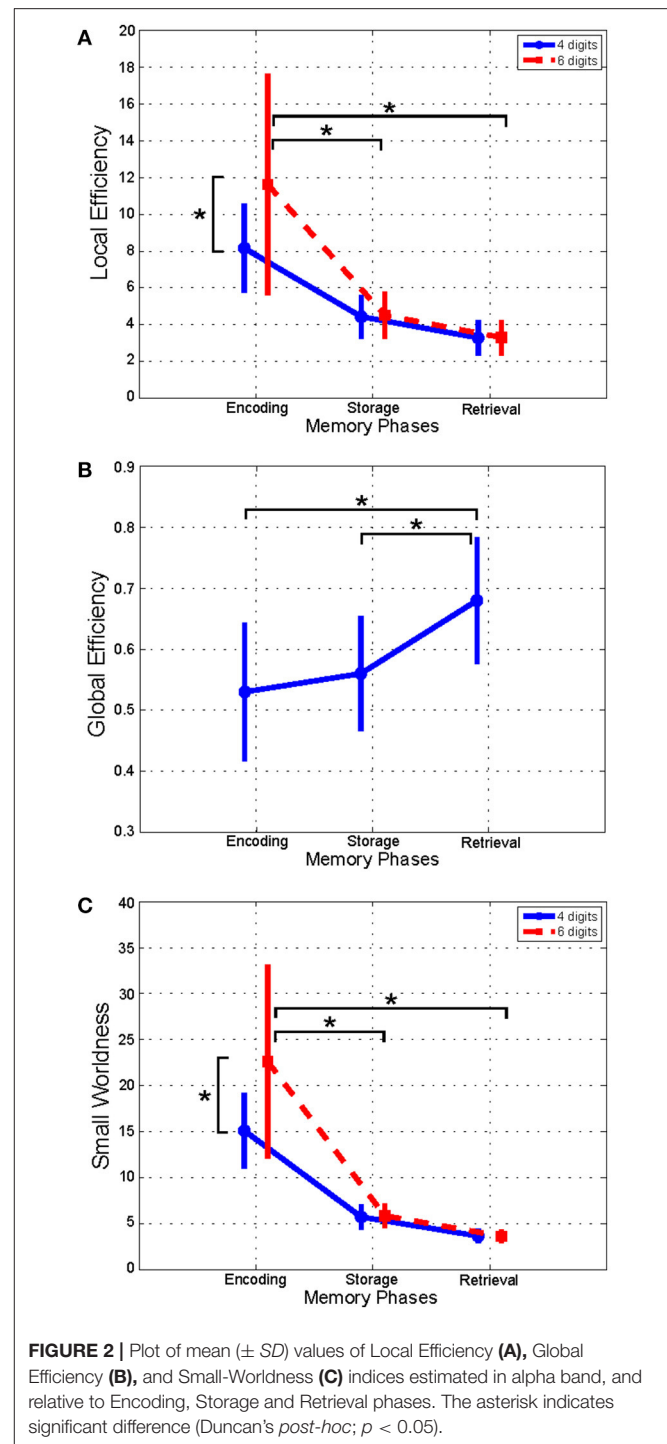
FIGURE 2 | Plot of mean (± *SD*) values of Local Efficiency **(A),** Global Efficiency **(B),** and Small-Worldness **(C)** indices estimated in alpha band, and relative to Encoding, Storage and Retrieval phases. The asterisk indicates significant difference (Duncan's *post-hoc*; $p < 0.05$).

tendency to reduce local segregation (see also local efficiency decrease) in favor of global integration (see global efficiency increase) of the information exchange between/within the different brain regions as cognitive processing evolves from encoding to retrieval. Such dynamics in the topological re-arrangement is consistent with the recently released global workspace theory (Baars and Franklin, 2003; Baars et al., 2013) postulating that the networks structure reorganizes across
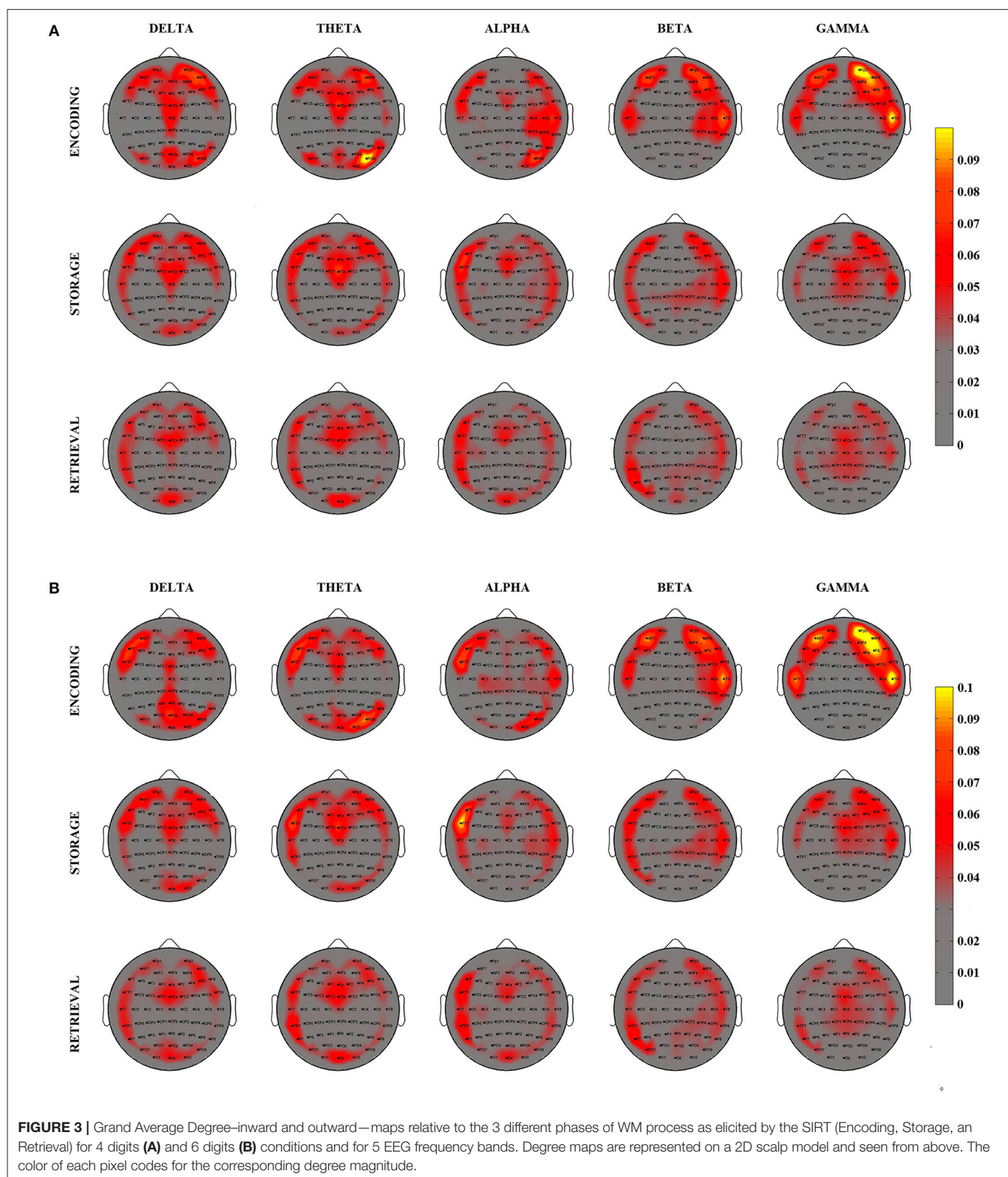
FIGURE 3 | Grand Average Degree–inward and outward—maps relative to the 3 different phases of WM process as elicited by the SIRT (Encoding, Storage, an Retrieval) for 4 digits **(A)** and 6 digits **(B)** conditions and for 5 EEG frequency bands. Degree maps are represented on a 2D scalp model and seen from above. The color of each pixel codes for the corresponding degree magnitude.

the temporal evolution of WM cognitive processing (Bola and Sabel, 2015). As such, the dynamics observed in the EEG-derived network(s) would reinforce the assumption at the base of multi-store model revised by Baddeley (2012) that the several cognitive components (i.e., central executive, episodic buffer (s), phonological loop and visuo-spatial sketchpad) are

TABLE 4 | Results of two-way repeated measures ANOVA on local indices (F-values, **p <0.001, *p < 0.01).

Local index	Band	Phases (d.f. = 2,30)	Digits (d.f. = 1,15)	Phases × digits (d.f. = 2,30)
Left Frontal Degree	δ	**4.89***	0.34	0.66
	θ	2.67	0.37	2.89
	α	2.12	1.31	1.65
	β	**4.11***	0.38	0.14
	γ	**26.16****	4.77	1.48
Frontal Midline Degree	δ	0.13	1.34	1.06
	θ	1.42	0.72	0.38
	α	0.11	0.65	0.65
	β	2.31	0.49	0.05
	γ	**6.39***	0.01	0.83
Right Frontal Degree	δ	**12.81****	3.05	**6.92***
	θ	**4.98***	0.12	1.05
	α	0.58	0.13	0.08
	β	**12.16****	2.23	0.74
	γ	**20.56****	**6.51***	3.22
Left Temporal Degree	δ	0.66	4.49	0.83
	θ	3.01	2.33	0.46
	α	1.39	**7.62***	2.84
	β	0.78	2.02	0.57
	γ	0.53	0.001	0.75
Right Temporal Degree	δ	1.72	0.38	1.49
	θ	1.5	0.19	0.73
	α	1.53	0.34	0.93
	β	**9.14****	0.38	2.52
	γ	**15.35****	0.09	0.27
Left Parietal Degree	δ	0.64	0.9	0.68
	θ	1.11	0.12	1.55
	α	0.46	0.93	0.42
	β	**7.55***	1.15	0.79
	γ	**26.74****	**4.89***	2.08
Occipital Degree	δ	0.45	0.2	0.71
	θ	1.22	0.22	1.55
	α	**5.63***	1.16	1.04
	β	**8.95****	0.43	2.49
	γ	**6.42***	0.49	**5.04***
Right Parietal Degree	δ	2.87	0.96	0.59
	θ	**11.19****	0.09	0.09
	α	**5.75***	0.009	0.14
	β	0.17	0.01	0.2
	γ	0.14	0.001	0.17

FDR correction for multiple ANOVAs was applied. Significant results are highligthed in bold.

not "crystallized" but have a tendency to be "fluid" as well the capacity they sub-serve (attention, temporary storage…; Baddeley, 2012). In particular, this dynamic observed for the
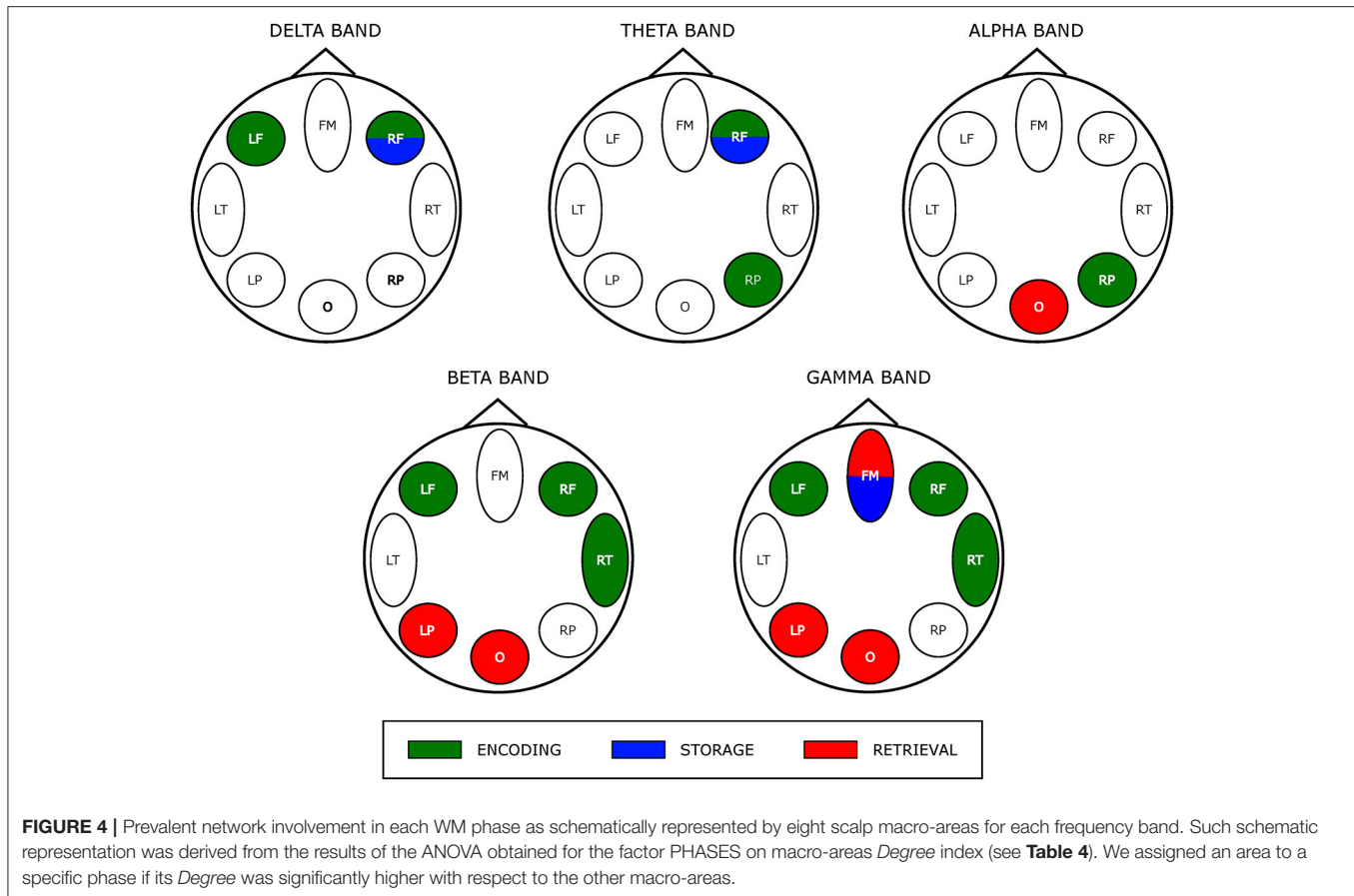
topology networks might reflect the operational mode of the "episodic buffer" component of the Baddeley model (Baddeley, 2010). This "buffer" serves as an intermediary between the storage subsystems with different codes (i.e., phonological loop and visuo-spatial sketchpad) whose content is bounded by the buffer into unitary multi-dimensional representations. Thus, one can speculate that a tendency toward a more global vs. local integration network topology (ie, the decrease of small-worldness across WM phases) would "optimally" serve the function of the episodic buffer by favoring the information flow between WM networks (i.e., 2 storage subsystems).

The dynamic changings toward a more globally integrated network(s) interplay with less specialized segregation could also be effective in sustaining more recent WM models (for review see Eriksson et al., 2015). In this recent reappraisal of WM functioning, the content of working memory would be defined by an interaction between selective perceptual (visual, auditory…) information process (operated via a selective attention) and LTM representations being in a particular state of "accessibility" that requires a largely persistent activity of specialized networks controlled by attentional processes. Thus, a whole brain network with high global information transfer (integration) would better "sustain" an optimal interplay between locally specialized networks (see below, local organization of WM subnetworks).

The encoding process directly influences the precision and accuracy of subsequent WM representations (Awh and Vogel, 2008; Rutman et al., 2010). The well-known limitation in the capacity to simultaneously encode objects requires efficient mechanisms to operate a selection of only the most relevant objects from the immediate environment to be represented in WM by restricting those irrelevant from consuming capacity (Vogel et al., 2005; Chun, 2011). In this regard, several evidence indicate that a successful encoding information into WM is the result of an interplay between brain circuits underlying selective attention processes and perceptual (e.g., visual) object representation (for a review Gazzaley and Nobre, 2012) that, in a more recent vision would trigger LTM object representation (Eriksson et al., 2015). A small-world topology could well account for this complex network interplay by supporting both specialized and integrated information processing in the whole brain connectivity. It comprises both high segregated (or modular) processing (high clustering) and distributed (or integrated) processing (short path length; Bassett and Bullmore, 2006). The observed high small-world network during WM encoding (with respect to maintenance and retrieval) would fit with the necessity of the brain to combine the functioning of specialized (segregated) modules with a number of inter-modular links integrating those modules.

To further corroborate this interpretation, it is notable to mention that a disruption of an *optimal* small-world network organization has been described in schizophrenic patients (Fornito et al., 2012) who exhibited an impairment of WM performance accounted by a decreased efficiency in item encoding (Cairo et al., 2006; Koch et al., 2009). As yet, altered oscillatory dynamics during encoding of information have been reported in normal and pathological aging associated with

FIGURE 4 | Prevalent network involvement in each WM phase as schematically represented by eight scalp macro-areas for each frequency band. Such schematic representation was derived from the results of the ANOVA obtained for the factor PHASES on macro-areas *Degree* index (see **Table 4**). We assigned an area to a specific phase if its *Degree* was significantly higher with respect to the other macro-areas.

cognitive decline (Aine et al., 2011; Kirova et al., 2015; Proskovec et al., 2016).

We also found that the network *optimal* topology defined by a high local efficiency and small-worldness, increased significantly as a function of WM load increase (4 vs. 6 digits) only during encoding. This WM load-induced modulation of network topology reinforces the above interpretation of a high network modularity required during encoding. Recent neurophysiological evidences support the idea that visual WM capacity limitation (i.e., the so-called *set size effect* Luck and Vogel, 1997) begins with neural attention resource allocation at encoding (Gurariy et al., 2016). Of note, the WM load-induced increase in the values of the network global indices was selectively observed in the alpha and gamma frequency bands. Alpha oscillations have been hypothesized to play an active role in protecting WM items from non-relevant information (Jensen and Mazaheri, 2010) by suppressing distracting sensory information (Romei et al., 2010). More specifically, the increase of WM load is associated to an increase of alpha-band coherence between midline parietal and left temporal/parietal sites during encoding (Payne and Kounios, 2009). From the behavioral view point, we found that the network local efficiency estimated in alpha band and relative to encoding varied as function of the RTs (negative correlation; **Figure 3**). The existence of such correlation exclusively in alpha band is in accordance with previous evidence of a correlation between changes in α-power spectrum and behavioral performance

during encoding (Klimesch, 1999; Bashivan et al., 2014). The increase in the values of the indices describing the global network organization during encoding was also selective for the gamma frequency. As such this finding is in line with evidence of a direct correlation between the changes in the gamma power spectrum amplitude and the number of items to be memorized (Howard et al., 2003; Roux et al., 2012; Roux and Uhlhaas, 2014).

Local Organization of the Working Memory Networks

The computation of local degree maps (GA illustrated in **Figure 3**) allowed for the identification of *hubs* within the WM network(s) activated across the three phases elicited by the visual SIRT. As expected, encoding, storage and retrieval WM phases were consistently characterized by a main involvement of bilateral *frontal* and *temporal* regions in all frequency oscillations while an *anterior-to-posterior midline* pattern was prevalent in the low (delta/theta) EEG frequency range. In addition, a bilateral *parieto-occipital connectivity* was observed mainly in theta/delta oscillations during the encoding, while storage/retrieval phase were characterized by a prevalent *left temporo-parietal and right fronto-parietal connectivity* in alpha/beta bands. These patterns were sensitive to WM load increase.

Consistently with the view of WM as emerging from the dynamic interplay of several brain regions, recent evidences indicate that the integrity of white matter pathways connecting

the dorsolateral frontal cortex, parietal cortex, and temporal cortex correlates with working-memory performance (Charlton et al., 2010). Within this large network of areas, the prefrontal cortex has been suggested to be crucial for executive demands such as the maintenance of resilient information during WM, updating WM content, and shifting (Nee et al., 2013). Together with the prefrontal areas, the parietal cortex is also causally involved in WM functioning, being associated with executive aspects (superior parietal cortex) of WM (Collette et al., 2005) and the implementation of selective attentional control (Awh et al., 2006). Interestingly, parietal cortex activity correlates with WM capacity in that its activity increases as the number of items to remember increases (Vogel and Machizawa, 2004). According to computational modeling (O'Reilly, 2006), the basal ganglia (striatum) would play a key role in controlling (filtering) when the prefrontal cortex representations should be maintained vs. updated. Moreover, the above mentioned parietal load effect negatively correlated with basal ganglia activity (McNab and Klingberg, 2008).

Our connectivity patterns expressed as locally distributed *hubs* of information flow between scalp regions (i.e., local degree index) well reflect the main interpretational mapping of WM processes to brain regions, thus highlighting the accuracy of our EEG network estimation approach in providing indices which can specifically describe the distributed topography of the networks involved in WM task solving. The spectral features of the estimated local topological indices further corroborate their selectivity in capturing the (local) functional dynamics underlying WM processing.

There are a number of evidence that an interplay between rhythmic activity at low (delta/theta) and high (beta/gamma) frequency has been suggested to enable WM item encoding and maintenance in humans (for review see Roux and Uhlhaas, 2014). Particularly, the gamma band would be relevant for active maintenance of WM information, whereas theta band would be involved in the temporal organization of WM items. The relevance of alpha oscillation would reside in filtering task non-relevant information.

As schematically illustrated in **Figure 4**, we actually found that *Encoding* networks were mainly described by *hubs* (encoding—related local degree indices contrasted against those of storage and retrieval time series) located within bilateral frontal and right fronto-temporal scalp area in low (delta/theta) and in the high frequency oscillation range (beta/gamma), respectively.

As discussed above, gamma band activity plays a role in maintenance of visual (and others sensory) WM items (Tallon-Baudry et al., 1998; Kaiser et al., 2008) which spatially occurs within the prefrontal cortex in association with parietal cortex (Eriksson et al., 2015). In addition, EEG/MEG source localization studies pointed out that gamma oscillatory activity changes (increased power) is mainly localized over frontal (and parietal) regions (Palva and Palva, 2012). The exclusive bilateral frontal areas involvement for delta activity-related hubs would be consistent with the role of sustained delta activity in inhibiting interferences that might affect task performance (Harmony, 2013; Kleen et al., 2016).

Finally, the observed lateralization of network hubs toward right parietal area in alpha/theta during encoding might be related to spatial WM (Owen et al., 2005) elicited by a visual SIRT. Contemporary views of alpha/theta range of frequency suggest that it reflects the allocation of spatial attention to the memoranda, as well as the suppression of distracting information (Corbetta and Shulman, 2002; Asplund et al., 2010; Klimesch, 2012).

The *storage* partially overlaps *encoding* network hub map as it involved (right) frontal area in delta/theta frequency range. Although the correspondence in neural activity between encoding and maintenance still remains debatable (Gazzaley et al., 2004; Woodward et al., 2006; Chang et al., 2007) recent work by Cohen et al. (2014) provides empirical evidence for an overlap between encoding and maintenance processes as a critical element of WM (Postle, 2006; D'Esposito, 2007). A gamma-related frontal midline hub also was observed (in common with *retrieval*) that also could reflect the interplay between subcortical structures (i.e., basal ganglia) and (pre) frontal areas cortex responsible for maintenance of relevant WM items (Kaiser et al., 2009; Roux et al., 2012).

During *retrieval* we observed a selective distribution of network hubs within occipital area in alpha, beta and gamma bands as well as within the left parietal region in high frequency oscillations (beta/gamma band). Such parieto-occipital engagement could account for visual stimulus presentation and visual information processing during retrieval (Voytek et al., 2010). Moreover, neuronal synchronization in the gamma band over occipital areas has been associated to subject ability during encoding and retrieval memory phases (Osipova et al., 2006).

As expected local degree indices varied as function of WM load. Specifically, the left temporal degree increased as function of WM load in alpha band. This finding is consistent with a role of (left) temporal region in sub-lexical phonological processing of visual material (Howard et al., 1992; Price, 1998). During Sternberg tasks sequential encoding would activate the phonological loop to support the maintenance of sequenced WM items by means of subvocal rehearsal (silent speech; Henson et al., 2000; Barry et al., 2011). We found that frontal and parietal local degree indices were sensitive to WM load in gamma oscillations that well reflect the direct correlation between gamma power magnitude and the number of items to be memorized and the role of sustained activity in frontal and parietal areas for maintenance of memoranda (Howard et al., 2003; Roux et al., 2012; Roux and Uhlhaas, 2014).

The present EEG-derived network findings still await for further consolidation which requires a large sample of different age population to be involved and a network model testing procedure such as challenging the distributed network hub mapping both internally (e.g., leave-one-out approach to verify the fault tolerance of the network to the removal of one node) and/or externally (e.g., with non-invasive technique to induced virtual lesions such as rTMS).

Upon consolidation, our EEG-derived network estimation approach may on the one hand, break new ground in the WM function theoretical modeling and on the other it would offer a

valuable and affordable method to improve clinical assessment and evaluate treatment efficacy of cognitive disorders occurring after brain lesions (i.e., stroke).

AUTHOR CONTRIBUTIONS

JT: EEG experimental data analysis management; manuscript writing; LA: experimental design definition, EEG derived brain network data analysis supervision and validation; MR: interpretation of behavioral data; AA: implementation of methodological pipeline (connectivity estimation and graph theory approach); SK: EEG and behavioral data collection; GW: experimental design definition and supervision of data collection; DM: responsible for study; study design and management; overall data interpretation; manuscript writing management.

FUNDING

REFERENCES

Aine, C. J., Sanfratello, L., Adair, J. C., Knoefel, J. E., Caprihan, A., and Stephen, J. M. (2011). Development and decline of memory functions in normal, pathological and healthy successful aging. *Brain Topogr.* 24, 323–339. doi: 10.1007/s10548-011-0178-x

Aine, C. J., Woodruff, C. C., Knoefel, J. E., Adair, J. C., Hudson, D., Qualls, C., et al. (2006). Aging: compensation or maturation? *Neuroimage* 32, 1891–1904. doi: 10.1016/j.neuroimage.2006.05.005

Asplund, C. L., Todd, J. J., Snyder, A. P., and Marois, R. (2010). A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nat. Neurosci.* 13, 507–512. doi: 10.1038/nn.2509

Astolfi, L., Cincotti, F., Mattia, D., Marciani, M. G., Baccalà, L. A., de Vico Fallani, F., et al. (2006). Assessing cortical functional connectivity by partial directed coherence: simulations and application to real data. *IEEE Trans. Biomed. Eng.* 53, 1802–1812. doi: 10.1109/TBME.2006.873692

Astolfi, L., Toppi, J., Wood, G., Kober, S., Risetti, M., Macchiusi, L., et al. (2013). "Advanced methods for time-varying effective connectivity estimation in memory processes," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka), 2936–2939.

Awh, E., and Vogel, E. K. (2008). The bouncer in the brain. *Nat. Neurosci.* 11, 5–6. doi: 10.1038/nn0108-5

Awh, E., Vogel, E. K., and Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience* 139, 201–208. doi: 10.1016/j.neuroscience.2005.08.023

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., and Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc. Natl. Acad. Sci. U.S.A.* 1007, 3228–3233. doi: 10.1073/pnas.0911531107

Baars, B. J., and Franklin, S. (2003). How conscious experience and working memory interact. *Trends Cogn. Sci.* 7, 166–172. doi: 10.1016/S1364-6613(03)00056-1

Baars, B. J., Franklin, S., and Ramsøy, T. Z. (2013). Global workspace dynamics: cortical "binding and propagation" enables conscious contents. *Front. Psychol.* 4:200. doi: 10.3389/fpsyg.2013.00200

Baccalá, L. A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.* 84, 463–474. doi: 10.1007/PL00007990

Baddeley, A. (1996). The fractionation of working memory. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13468–13472. doi: 10.1073/pnas.93.24.13468

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* 4, 417–423. doi: 10.1016/S1364-6613(00)01538-2

Baddeley, A. (2010). Working memory. *Curr. Biol.* 20, R136–R140. doi: 10.1016/j.cub.2009.12.014

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29. doi: 10.1146/annurev-psych-120710-100422

Barry, J. G., Sabisch, B., Friederici, A. D., and Brauer, J. (2011). Encoding: the keystone to efficient functioning of verbal short-term memory. *Neuropsychologia* 49, 3636–3647. doi: 10.1016/j.neuropsychologia.2011.09.018

Bashivan, P., Bidelman, G. M., and Yeasin, M. (2014). Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. *Eur. J. Neurosci.* 40, 3774–3784. doi: 10.1111/ejn.12749

Bassett, D. S., and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* 12, 512–523. doi: 10.1177/1073858406293182

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Bola, M., and Sabel, B. A. (2015). Dynamic reorganization of brain functional networks during cognition. *Neuroimage* 114, 398–413. doi: 10.1016/j.neuroimage.2015.03.057

Brookes, M. J., Wood, J. R., Stevenson, C. M., Zumer, J. M., White, T. P., Liddle, P. F., et al. (2011). Changes in brain network activity during working memory tasks: a magnetoencephalography study. *Neuroimage* 55, 1804–1815. doi: 10.1016/j.neuroimage.2010.10.074

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575

Cairo, T. A., Liddle, P. F., Woodward, T. S., and Ngan, E. T. C. (2004). The influence of working memory load on phase specific patterns of cortical activity. *Brain Res. Cogn. Brain Res.* 21, 377–387. doi: 10.1016/j.cogbrainres.2004.06.014

Cairo, T. A., Woodward, T. S., and Ngan, E. T. C. (2006). Decreased encoding efficiency in Schizophrenia. *Biol. Psychiatry* 59, 740–746. doi: 10.1016/j.biopsych.2005.08.009

Chang, C., Crottaz-Herbette, S., and Menon, V. (2007). Temporal dynamics of basal ganglia response and connectivity during verbal working memory. *Neuroimage* 34, 1253–1269. doi: 10.1016/j.neuroimage.2006.08.056

Charlton, R. A., Barrick, T. R., Lawes, I. N. C., Markus, H. S., and Morris, R. G. (2010). White matter pathways associated with working memory in normal aging. *Cortex* 46, 474–489. doi: 10.1016/j.cortex.2009.07.005

Chen, S. H. A., and Desmond, J. E. (2005). Cerebrocerebellar networks during articulatory rehearsal and verbal working memory tasks. *Neuroimage* 24, 332–338. doi: 10.1016/j.neuroimage.2004.08.032

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia* 49, 1407–1409. doi: 10.1016/j.neuropsychologia.2011.01.029

Cohen, J. R., Sreenivasan, K. K., and D'Esposito, M. (2014). Correspondence between stimulus encoding- and maintenance-related neural processes underlies successful working memory. *Cereb. Cortex* 24, 593–599. doi: 10.1093/cercor/bhs339

Collette, F., Van der Linden, M., Laureys, S., Delfiore, G., Degueldre, C., Luxen, A., et al. (2005). Exploring the unity and diversity of the neural substrates of executive functioning. *Hum. Brain Mapp.* 25, 409–423. doi: 10.1002/hbm.20118

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Corbin, L., and Marquer, J. (2013). Is Sternberg's memory scanning task really a short-term memory task? *Swiss. J. Psychol.* 72, 181–196. doi: 10.1024/1421-0185/a000112.

Corsi, P. M., and Michael, P. (1972). *Human Memory and the Medial Temporal Region of the Brain. McGill University Montreal.* (Accessed May 22, 2017).

Cramer, S. C., Sur, M., Dobkin, B. H., O'Brien, C., Sanger, T. D., Trojanowski, J. Q., et al. (2011). Harnessing neuroplasticity for clinical applications. *Brain J. Neurol.* 134, 1591–1609. doi: 10.1093/brain/awr039

Cummins, T. D., and Finnigan, S. (2007). Theta power is reduced in healthy cognitive aging. *Int. J. Psychophysiol.* 66, 10–17. doi: 10.1016/j.ijpsycho.2007.05.008

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philos. Trans. R. Soc. B Biol. Sci.* 362, 761–772. doi: 10.1098/rstb.2007.2086

D'Esposito, M., and Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* 66, 115–142. doi: 10.1146/annurev-psych-010814-015031

Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F., and Nyberg, L. (2015). Neurocognitive architecture of working memory. *Neuron* 88, 33–46. doi: 10.1016/j.neuron.2015.09.020

Fagiolo, G. (2007). Clustering in complex directed networks. *Phys. Rev.* 76:026107. doi: 10.1103/PhysRevE.76.026107

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *Neuroimage* 62, 2296–2314. doi: 10.1016/j.neuroimage.2011.12.090

Gazzaley, A., and Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135. doi: 10.1016/j.tics.2011.11.014

Gazzaley, A., Rissman, J., and D'Esposito, M. (2004). Functional connectivity during working memory maintenance. *Cogn. Affect. Behav. Neurosci.* 4, 580–599. doi: 10.3758/CABN.4.4.580

Grady, C., Springer, M., Hongwanishkul, D., McIntosh, A., and Winocur, G. (2006). Age-related changes in brain activity across the adult lifespan. *J. Cogn. Neurosci.* 18, 227–241. doi: 10.1162/jocn.2006.18.2.227

Grefkes, C., and Fink, G. R. (2011). Reorganization of cerebral networks after stroke: new insights from neuroimaging with connectivity approaches. *Brain* 134, 1264–1276. doi: 10.1093/brain/awr033

Grefkes, C., and Fink, G. R. (2014). Connectivity-based approaches in stroke and recovery of function. *Lancet Neurol.* 13, 206–216. doi: 10.1016/S1474-4422(13)70264-3

Gurariy, G., Killebrew, K. W., Berryhill, M. E., and Caplovitz, G. P. (2016). Induced and evoked human electrophysiological correlates of visual working memory set-size effects at encoding. *PLoS ONE* 11:e0167022. doi: 10.1371/journal.pone.0167022

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Front. Integr. Neurosci.* 7:83. doi: 10.3389/fnint.2013.00083

Härting, C. (2000). *Wechsler-Gedächtnistest - Revidierte Fassung : WMS-R ; Deutsche Adaptation der Revidierten Fassung der Wechsler Memory Scale von David Wechsler.* Huber.

Haufe, S., Nikulin, V. V., Müller, K.-R., and Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: a simulation study. *Neuroimage* 64, 120–133. doi: 10.1016/j.neuroimage.2012.09.036

Heinrichs-Graham, E., and Wilson, T. W. (2015). Spatiotemporal oscillatory dynamics during the encoding and maintenance phases of a visual working memory task. *Cortex* 69, 121–130. doi: 10.1016/j.cortex.2015.04.022

Henson, R. N., Burgess, N., and Frith, C. D. (2000). Recoding, storage, rehearsal and grouping in verbal short-term memory: an fMRI study. *Neuropsychologia* 38, 426–440. doi: 10.1016/S0028-3932(99)00098-6

Honey, C. J., and Sporns, O. (2008). Dynamical consequences of lesions in cortical networks. *Hum. Brain Mapp.* 29, 802–809. doi: 10.1002/hbm.20579

Howard, D., Patterson, K., Wise, R., Brown, W. D., Friston, K., Weiller, C., et al. (1992). The cortical localization of the lexicons. positron emission tomography evidence. *Brain J. Neurol.* 115(Pt 6), 1769–1782. doi: 10.1093/brain/115.6.1769

Howard, M. W., Rizzuto, D. S., Caplan, J. B., Madsen, J. R., Lisman, J., Aschenbrenner-Scheibe, R., et al. (2003). Gamma oscillations correlate with working memory load in humans. *Cereb. Cortex* 13, 1369–1374. doi: 10.1093/cercor/bhg084

Humphries, M. D., and Gurney, K. (2008). Network "small-world-ness": a quantitative method for determining canonical network equivalence. *PLoS ONE* 3:e0002051. doi: 10.1371/journal.pone.0002051

Hwang, G., Jacobs, J., Geller, A., Danker, J., Sekuler, R., and Kahana, M. J. (2005). EEG correlates of verbal and nonverbal working memory. *Behav. Brain Funct.* 1:20. doi: 10.1186/1744-9081-1-20

Jensen, O., Gelfand, J., Kounios, J., and Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) Increase with memory load during retention in a short-term memory task. *Cereb. Cortex* 12, 877–882. doi: 10.1093/cercor/12.8.877

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4:186. doi: 10.3389/fnhum.2010.00186

Jensen, O., and Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *Eur. J. Neurosci.* 15, 1395–1399. doi: 10.1046/j.1460-9568.2002.01975.x

Kaiser, J., Heidegger, T., Wibral, M., Altmann, C. F., and Lutzenberger, W. (2008). Distinct gamma-band components reflect the short-term memory maintenance of different sound lateralization angles. *Cereb. Cortex* 18, 2286–2295. doi: 10.1093/cercor/bhm251

Kaiser, J., Lutzenberger, W., Decker, C., Wibral, M., and Rahm, B. (2009). Task- and performance-related modulation of domain-specific auditory short-term memory representations in the gamma-band. *Neuroimage* 46, 1127–1136. doi: 10.1016/j.neuroimage.2009.03.011

Kayser, J., Tenke, C. E., Gates, N. A., and Bruder, G. E. (2007). Reference-independent ERP old/new effects of auditory and visual word recognition memory: joint extraction of stimulus- and response-locked neuronal generator patterns. *Psychophysiology* 44, 949–967. doi: 10.1111/j.1469-8986.2007.00562.x

Keren-Happuch, E., Chen, S.-H. A., Ho, M.-H. R., and Desmond, J. E. (2012). A meta-analysis of cerebellar contributions to higher cognition from PET and fMRI studies. *Hum. Brain Mapp.* 35, 593-615. doi: 10.1002/hbm.22194

Kirova, A.-M., Bays, R. B., and Lagalwar, S. (2015). Working memory and executive function decline across normal aging, mild cognitive impairment, and Alzheimer's disease. *Biomed. Res. Int.* 2015:748212. doi: 10.1155/2015/748212

Kleen, J. K., Testorf, M. E., Roberts, D. W., Scott, R. C., Jobst, B. J., Holmes, G. L., et al. (2016). Oscillation phase locking and late erp components of intracranial hippocampal recordings correlate to patient performance in a working memory task. *Front. Hum. Neurosci.* 10:287. doi: 10.3389/fnhum.2016.00287

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Klimesch, W. (2012). α-band oscillations, attention, and controlled access to stored information. *Trends Cogn. Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Kober, S. E., Schweiger, D., Witte, M., Reichert, J. L., Grieshofer, P., Neuper, C., et al. (2015). Specific effects of EEG based neurofeedback training on memory functions in post-stroke victims. *J. Neuroeng. Rehabil.* 12:107. doi: 10.1186/s12984-015-0105-6

Kober, S. E., Witte, M., Neuper, C., and Wood, G. (2017). Specific or nonspecific? Evaluation of band, baseline, and cognitive specificity of sensorimotor rhythm- and gamma-based neurofeedback. *Int. J. Psychophysiol.* 120, 1–13. doi: 10.1016/j.ijpsycho.2017.06.005

Koch, K., Wagner, G., Schultz, C., Schachtzabel, C., Nenadic, I., Axer, M., et al. (2009). Altered error-related activity in patients with schizophrenia. *Neuropsychologia* 47, 2843–2849. doi: 10.1016/j.neuropsychologia.2009.06.010

Kottlow, M., Schlaepfer, A., Baenninger, A., Michels, L., Brandeis, D., and Koenig, T. (2015). Pre-stimulus BOLD-network activation modulates EEG spectral activity during working memory retention. *Front. Behav. Neurosci.* 9:111. doi: 10.3389/fnbeh.2015.00111

Kristofferson, M. W. (1972). Effects of practice on character-classification performance. *Can. J. Psychol. Can. Psychol.* 26, 54–60. doi: 10.1037/h0082415

Kuś, R., Kaminski, M., and Blinowska, K. J. (2004). Determination of EEG activity propagation: pair-wise versus multichannel estimate. *IEEE Trans. Biomed. Eng.* 51, 1501–1510. doi: 10.1109/TBME.2004.827929

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87:198701. doi: 10.1103/PhysRevLett.87.198701

Luck, S. J., and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature* 390, 279–281. doi: 10.1038/36846

Luis, E. O., Arrondo, G., Vidorreta, M., Martínez, M., Loayza, F., Fernández-Seara, M. A., et al. (2015). Successful working memory processes and cerebellum in an elderly sample: a Neuropsychological and fMRI study. *PLoS ONE* 10:e0131536. doi: 10.1371/journal.pone.0131536

MacPherson, S. E., Wagner, G. P., Murphy, P., Bozzali, M., Cipolotti, L., and Shallice, T. (2014). Bringing the cognitive estimation task into the 21st century: normative data on two new parallel forms. *PLoS ONE* 9:e92554. doi: 10.1371/journal.pone.0092554

Majerus, S., Poncelet, M., Van der Linden, M., Albouy, G., Salmon, E., Sterpenich, V., et al. (2006). The left intraparietal sulcus and verbal short-term memory: focus of attention or serial order? *Neuroimage* 32, 880–891. doi: 10.1016/j.neuroimage.2006.03.048

Marvel, C. L., and Desmond, J. E. (2010). Functional topography of the cerebellum in verbal working memory. *Neuropsychol. Rev.* 20, 271–279. doi: 10.1007/s11065-010-9137-7

MATLAB (2011). *The MathWorks, Inc*. Natick, MA.

Mathes, B., Schmiedt, J., Schmiedt-Fehr, C., Pantelis, C., and Basar-Eroglu, C. (2012). New rather than old? For working memory tasks with abstract patterns the P3 and the single-trial delta response are larger for modified than identical probe stimuli. *Psychophysiology* 49, 920–932. doi: 10.1111/j.1469-8986.2012.01372.x

Mattay, V. S., Fera, F., Tessitore, A., Hariri, A. R., Berman, K. F., Das, S., et al. (2006). Neurophysiological correlates of age-related changes in working memory capacity. *Neurosci. Lett.* 392, 32–37. doi: 10.1016/j.neulet.2005.09.025

Maurer, U., Brem, S., Liechti, M., Maurizio, S., Michels, L., and Brandeis, D. (2015). Frontal midline theta reflects individual task performance in a working memory task. *Brain Topogr.* 28, 127–134. doi: 10.1007/s10548-014-0361-y

McNab, F., and Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nat. Neurosci.* 11, 103-107. doi: 10.1038/nn2024

Milde, T., Leistritz, L., Astolfi, L., Miltner, W. H. R., Weiss, T., Babiloni, F., et al. (2010). A new Kalman filter approach for the estimation of high-dimensional time-variant multivariate AR models and its application in analysis of laser-evoked brain potentials. *Neuroimage* 50, 960–969. doi: 10.1016/j.neuroimage.2009.12.110

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., et al. (2013). A meta-analysis of executive components of working memory. *Cereb. Cortex* 23, 264–282. doi: 10.1093/cercor/bhs007

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin. Neurophysiol.* 115, 2292–2307. doi: 10.1016/j.clinph.2004.04.029

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., et al. (2008). Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* 100:234101. doi: 10.1103/PhysRevLett.100.234101

Nosofsky, R. M., Little, D. R., Donkin, C., and Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychol. Rev.* 118, 280–315. doi: 10.1037/a0022494

O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science* 314, 91–94. doi: 10.1126/science.1127242

Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E., and Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *J. Neurosci.* 26, 7523–7531. doi: 10.1523/JNEUROSCI.1948-06.2006

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59. doi: 10.1002/hbm.20131

Palva, S., and Palva, J. M. (2012). Discovering oscillatory interaction networks with M/EEG: challenges and breakthroughs. *Trends Cogn. Sci.* 16, 219–230. doi: 10.1016/j.tics.2012.02.004

Payne, L., and Kounios, J. (2009). Coherent oscillatory networks supporting short-term memory retention. *Brain Res.* 1247, 126–132. doi: 10.1016/j.brainres.2008.09.095

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience* 139, 23–38. doi: 10.1016/j.neuroscience.2005.06.005

Price, C. J. (1998). The functional anatomy of word comprehension and production. *Trends Cogn. Sci.* 2, 281–288. doi: 10.1016/S1364-6613(98)01201-7

Proskovec, A. L., Heinrichs-Graham, E., and Wilson, T. W. (2016). Aging modulates the oscillatory dynamics underlying successful working memory encoding and maintenance. *Hum. Brain Mapp.* 37, 2348–2361. doi: 10.1002/hbm.23178

Romei, V., Gross, J., and Thut, G. (2010). On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: correlation or causation? *J. Neurosci.* 30, 8692–8697. doi: 10.1523/JNEUROSCI.0160-10.2010

Roux, F., and Uhlhaas, P. J. (2014). Working memory and neural oscillations: α-γ versus θ-γ codes for distinct WM information? *Trends Cogn. Sci.* 18, 16–25. doi: 10.1016/j.tics.2013.10.010

Roux, F., Wibral, M., Mohr, H. M., Singer, W., and Uhlhaas, P. J. (2012). Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. *J. Neurosci.* 32, 12411–12420. doi: 10.1523/JNEUROSCI.0421-12.2012

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Rutman, A. M., Clapp, W. C., Chadick, J. Z., and Gazzaley, A. (2010). Early top-down control of visual processing predicts working memory performance. *J. Cogn. Neurosci.* 22, 1224–1234. doi: 10.1162/jocn.2009.21257

Rypma, B., and D'Esposito, M. (1999). The roles of prefrontal brain regions in components of working memory: effects of memory load and individual differences. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6558–6563. doi: 10.1073/pnas.96.11.6558

Rypma, B., Prabhakaran, V., Desmond, J. E., Glover, G. H., and Gabrieli, J. D. (1999). Load-dependent roles of frontal brain regions in the maintenance of working memory. *Neuroimage* 9, 216–226. doi: 10.1006/nimg.1998.0404

Scheeringa, R., Petersson, K. M., Oostenveld, R., Norris, D. G., Hagoort, P., and Bastiaansen, M. C. M. (2009). Trial-by-trial coupling between EEG and BOLD identifies networks related to alpha and theta EEG power increases during working memory maintenance. *Neuroimage* 44, 1224–1238. doi: 10.1016/j.neuroimage.2008.08.041

Schellig, D., and Schächtele, B. (2009). *Visueller und verbaler* Merkfähigkeitstest: VVM ; Manual. Pearson.

Smith, E. E., and Jonides, J. (1998). Neuroimaging analyses of human working memory. *Proc. Natl. Acad. Sci. U.S.A.* 95, 12061–12068. doi: 10.1073/pnas.95.20.12061

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends Cogn. Sci.* 8, 418–425. doi: 10.1016/j.tics.2004.07.008

Sporns, O. (2013a). Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol.* 23, 162–171. doi: 10.1016/j.conb.2012.11.015

Sporns, O. (2013b). Structure and function of complex brain networks. *Dialogues Clin. Neurosci.* 15, 247–262.

Sternberg (1966). High-speed scanning in human memory. *Science* 153, 652–654. doi: 10.1126/science.153.3736.652

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *Am. Sci.* 57, 421–457.

Sturm, W., and Willmes, K. (1999). *Verbaler Lerntest (VLT): Nonverbaler Lerntest (NVLT): Handanweisung*. Hogrefe, Verlag für Psychologie.

Tallon-Baudry, C., Bertrand, O., Peronnet, F., and Pernier, J. (1998). Induced gamma-band activity during the delay of a visual short-term memory task in humans. *J. Neurosci.* 18, 4244–4254.

Thürling, M., Hautzel, H., Küper, M., Stefanescu, M. R., Maderwald, S., Ladd, M. E., et al. (2012). Involvement of the cerebellar cortex and nuclei in verbal and visuospatial working memory: a 7 T fMRI study. *Neuroimage* 62, 1537–1550. doi: 10.1016/j.neuroimage.2012.05.037

Toppi, J., Astolfi, L., Poudel, G. R., Innes, C. R. H., Babiloni, F., and Jones, R. D. (2016a). Time-varying effective connectivity of the cortical neuroelectric activity associated with behavioural microsleeps. *Neuroimage* 124, 421–432. doi: 10.1016/j.neuroimage.2015.08.059.

Toppi, J., De Vico Fallani, F., Vecchiato, G., Maglione, A. G., Cincotti, F., Mattia, D., et al. (2012). How the statistical validation of functional connectivity patterns can prevent erroneous definition of small-world properties of a brain connectivity network. *Comput. Math. Methods Med.* 2012:130985. doi: 10.1155/2012/130985

Toppi, J., Mattia, D., Risetti, M., Formisano, R., Babiloni, F., and Astolfi, L. (2016b). Testing the significance of connectivity networks: comparison of different assessing procedures. *IEEE Trans. Biomed. Eng.* 63, 2461–2473. doi: 10.1109/TBME.2016.2621668

Tuladhar, A. M., ter Huurne, N., Schoffelen, J.-M., Maris, E., Oostenveld, R., and Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Hum. Brain Mapp.*. 28, 785–792. doi: 10.1002/hbm.20306

van den Bosch, G. E., Marroun, H. E., Schmidt, M. N., Tibboel, D., Manoach, D. S., Calhoun, V. D., et al. (2014). Brain connectivity during verbal working memory in children and adolescents. *Hum. Brain Mapp.* 35, 698–711. doi: 10.1002/hbm.22193

Vergauwe, E., Hartstra, E., Barrouillet, P., and Brass, M. (2015). Domain-general involvement of the posterior frontolateral cortex in time-based resource-sharing in working memory: An fMRI study. *Neuroimage* 115, 104–116. doi: 10.1016/j.neuroimage.2015.04.059

Vogel, E. K., and Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature* 428:748. doi: 10.1038/nature02447

Vogel, E. K., McCollough, A. W., and Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature* 438, 500–503. doi: 10.1038/nature04171

Von Aster, M., Neubauer, A., and Horn, R. (2006). *Wechsler Intelligenztest Für Erwachsene (WIE). Deutschsprachige Bearbeitung Und Adaptation Des WAIS-III von David Wechsler*. Frankfurt/Main: Harcourt Test Services.

Voytek, B., Canolty, R. T., Shestyuk, A., Crone, N. E., Parvizi, J., and Knight, R. T. (2010). Shifts in gamma phase–amplitude coupling frequency from theta to alpha over posterior cortex during visual tasks. *Front. Hum. Neurosci.* 4:191. doi: 10.3389/fnhum.2010.00191

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* 393, 440–442. doi: 10.1038/30918

Woodward, T. S., Cairo, T. A., Ruff, C. C., Takane, Y., Hunter, M. A., and Ngan, E. T. C. (2006). Functional connectivity reveals load dependent neural systems underlying encoding and maintenance in verbal working memory. *Neuroscience* 139, 317–325. doi: 10.1016/j.neuroscience.2005.05.043

Woodward, T. S., Feredoes, E., Metzak, P. D., Takane, Y., and Manoach, D. S. (2013). Epoch-specific functional networks involved in working memory. *Neuroimage* 65, 529–539. doi: 10.1016/j.neuroimage.2012.09.070

Xie, Y., Feng, Z., Xu, Y., Bian, C., and Li, M. (2016). The different oscillation patterns of alpha band in the early and later stages of working memory maintenance. *Neurosci. Lett.* 633, 220–226. doi: 10.1016/j.neulet.2016.09.047

Zakrzewska, M. Z., and Brzezicka, A. (2014). Working memory capacity as a moderator of load-related frontal midline theta variability in Sternberg task. *Front. Hum. Neurosci.* 8:399. doi: 10.3389/fnhum.2014.00399

On the Relationship Between Attention Processing and P300-Based Brain Computer Interface Control in Amyotrophic Lateral Sclerosis

Angela Riccio[1]*, Francesca Schettini[2], Luca Simione[3], Alessia Pizzimenti[4], Maurizio Inghilleri[5], Marta Olivetti-Belardinelli[6,7], Donatella Mattia[1] and Febo Cincotti[1,8]

[1]Neuroelectrical Imaging and BCI Laboratory, NeiLab, Fondazione Santa Lucia (IRCCS), Rome, Italy, [2]Servizio Ausilioteca per Riabilitazione Assistita con Tecnologia (SARA-t), Fondazione Santa Lucia (IRCCS), Rome, Italy, [3]Institute of Cognitive Sciences and Technologies, Consiglio Nazionale delle Ricerche (CNR), Rome, Italy, [4]Crossing Dialogues, Rome, Italy, [5]Department of Neurology and Psychiatry, Sapienza University of Rome, Rome, Italy, [6]Centro Interuniversitario di Ricerca sull'Elaborazione Cognitiva in Sistemi Naturali e Artificiali (ECoNA), Rome, Italy, [7]ECONA Interuniversity Centre for Research on Natural and Artificial Systems, Sapienza University of Rome, Rome, Italy, [8]Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, Rome, Italy

Our objective was to investigate the capacity to control a P3-based brain-computer interface (BCI) device for communication and its related (temporal) attention processing in a sample of amyotrophic lateral sclerosis (ALS) patients with respect to healthy subjects. The ultimate goal was to corroborate the role of cognitive mechanisms in event-related potential (ERP)-based BCI control in ALS patients. Furthermore, the possible differences in such attentional mechanisms between the two groups were investigated in order to unveil possible alterations associated with the ALS condition. Thirteen ALS patients and 13 healthy volunteers matched for age and years of education underwent a P3-speller BCI task and a rapid serial visual presentation (RSVP) task. The RSVP task was performed by participants in order to screen their temporal pattern of attentional resource allocation, namely: (i) the temporal attentional filtering capacity (scored as T1%); and (ii) the capability to adequately update the attentive filter in the temporal dynamics of the attentional selection (scored as T2%). For the P3-speller BCI task, the online accuracy and information transfer rate (ITR) were obtained. Centroid Latency and Mean Amplitude of N200 and P300 were also obtained. No significant differences emerged between ALS patients and Controls with regards to online accuracy ($p = 0.13$). Differently, the performance in controlling the P3-speller expressed as ITR values (calculated offline) were compromised in ALS patients ($p < 0.05$), with a delay in the latency of P3 when processing BCI stimuli as compared with Control group ($p < 0.01$). Furthermore, the temporal aspect of attentional filtering which was related to BCI control ($r = 0.51$; $p < 0.05$) and to the P3 wave amplitude ($r = 0.63$; $p < 0.05$) was also altered in ALS patients ($p = 0.01$). These findings ground the knowledge required to develop sensible classes of BCI specifically designed by taking into account the influence of the cognitive characteristics of the possible candidates in need of a BCI system for communication.

Keywords: brain-computer interface, amyotrophic lateral sclerosis, attention, event-related potentials, P300, BCI, ALS, EEG

INTRODUCTION

The non-invasive brain-computer interface (BCI) based on the visual event-related potential (ERP) known as P300 (P3; Farwell and Donchin, 1988) is by far the most extensively investigated BCI system to enhance or even allow communication when this latter is severely compromised due to different neurological disorders (Kleih et al., 2011; Riccio et al., 2016). Despite the large amount of studies seeking for methods to ultimately optimize P3-based BCI control accuracy, a reliable and yet flexible (to be customized for individual users) BCI system still requires some research efforts.

Within the range of users in need of a BCI for communication and control, those with amyotrophic lateral sclerosis (ALS) represent *the target* population due to a progressive muscular paralysis that leads to a loss of communication and interaction ability thus preventing persons from using conventional assistive technologies (ATs) at the later stage of the disease.

A number of studies reported that ALS patients can communicate by using a P3-based BCI (Marchetti et al., 2013) with stable performance over time (Sellers and Donchin, 2006; Nijboer et al., 2008; Silvoni et al., 2013). Communication and interaction could also be enhanced by means of a P3-based BCI combined with an AT device (Thompson et al., 2014; Schettini et al., 2015). Marchetti and Priftis (2015) reported the results of a meta-analysis (pooled studies from 2008 to 2013) indicating that the effectiveness of the P3-speller (Farwell and Donchin, 1988) in ALS patients reached an overall classification accuracy of 73.72%. Further studies aimed at investigating predictors of P3-based BCI control in ALS patients showed that both external (the stimuli exploited) and internal factors (the user's motivation) could account for the BCI performance (Nijboer et al., 2010; Townsend et al., 2010; Kaufmann et al., 2013b). Conversely, the relation between BCI performance and the clinical-functional status of ALS patients (Nijboer et al., 2008; Silvoni et al., 2013; McCane et al., 2014; Thompson et al., 2014) has not been fully investigated yet. McCane et al. (2015) reported no significant differences in BCI accuracy between ALS patients and healthy age- (but not years of education-) matched subjects. However, they found differences in the target-related ERPs characteristics: the ALS group presented a higher N2 wave peak amplitude and a latency delay in N2, P3 and late negativity (LN) with respect to the control group. However, studies comparing ALS patients with healthy participants are still scarce. Further investigations are needed on the possible impairment/alteration of brain processing in response to external inputs (such as visual stimuli) delivered within a BCI framework of stimulation to eventually unveil whether and how they could influence the BCI control.

In this regard, it is important to note that cognitive deficits have been described in ALS patients (Lomen-Hoerth et al., 2003; Ringholz et al., 2005; Christidi et al., 2012; Strutt et al., 2012; Volpato et al., 2016; Radakovic et al., 2017). Up to now, the influence of ALS patients' cognitive profile on the visual P3-based BCI control has not been fully investigated. The current studies on visual P3-based BCIs for communication often involved end-users with severe motor disabilities due to

neurological disorders of various etiology (Piccione et al., 2006; Zickler et al., 2011; Kaufmann et al., 2013b; Kübler et al., 2014; Riccio et al., 2015) as compared to those studies in which more homogenous groups of participants, such as only ALS patients, were enrolled (Sellers and Donchin, 2006; Nijboer et al., 2008; Riccio et al., 2013; Silvoni et al., 2013; Thompson et al., 2014; Schettini et al., 2015). As such, this inconsistency between studies does not allow for definitive inferences on how P3-based BCI control could vary in severely motor disabled end-users with ALS and how this variability could be related to cognitive processing. In this line of reasoning, we previously showed (Riccio et al., 2013) how some aspects of attention processing such as the stimulus *temporal filtering* (i.e., the ability to keep the attentional filter active during the selection of a target) would be a predictor of the P3-speller control accuracy in ALS patients. Since we did not include a control (healthy) group, we could only speculate that such temporal aspect of attention processing was impaired in ALS population by comparing our results with those reported in other studies which included healthy participants (Kranczioch et al., 2005; Georgiou-Karistianis et al., 2007).

In the present study, we investigated whether the accuracy in mastering a P3-based BCI by an ALS population sample would be affected by the previously identified alterations in the attention processing and whether these alterations would be exclusive of the ALS population. To this purpose, we compared a group of ALS patients with a group of healthy volunteers both controlling a P3-speller (Farwell and Donchin, 1988). Groups were matched for age and years of education since both factors are known to deeply influence performance in executing cognitive tasks (Ardila et al., 2000). As yet, the two groups underwent an identical BCI stimulation protocol (i.e., the sequence of stimuli were not customized) in order to avoid confounding factors due to the use of different stimulation protocols. Based on previous findings (Riccio et al., 2013), our present hypothesis was that the ALS patients would show an altered visual attention processing of the stimuli delivered during the P3-based BCI control, and this would, in turn, affect the ability to control the P3-based BCI (i.e., decrease in performance). We also investigated the possible relation between cognitive mechanisms and P3-speller control to further corroborate the role of cognitive dysfunctions in BCI control in ALS patients (Riccio et al., 2013).

MATERIALS AND METHODS

Participants and Baseline Assessment

Thirteen participants (8 males; mean age 62.2 ± 13; years of formal education 13.7 ± 5.1) with ALS diagnosis (ALS group) and 13 age and years of education-matched participants (9 males; mean age 55.3 ± 9; years of formal education 13.3 ± 3) with no history of neurological/psychiatric disorders (Control group) were enrolled in the study. Seven out of 13 ALS patients participated in the previous study (Riccio et al., 2013).

The ALS patients were recruited through the ALS Center of the Policlinico "Umberto I", Sapienza University, Rome.

The study was conducted at Fondazione Santa Lucia, IRCCS, Rome and approved by the Independent Ethics Committee of Fondazione Santa Lucia. All participants (or the legal representatives of ALS patients when required) provided a written informed consent.

The inclusion criterion for the ALS patients was the ability (also with the help of an AT device if required) to clearly communicate (at least) a binary response (yes/no). Patients with other concomitant neurological or psychiatric disorders, any impediment in the acquisition of electroencephalography (EEG) data from the scalp (e.g., wounds, dermatitis), severe concomitant pathologies (fever, infections, metabolic disorders, severe heart failure), or episodes of reflex epilepsy were excluded from the study.

The level of physical disability was assessed by means of the "ALS Functional Rating Scale-Revised" (ALSFRS-R; Cedarbaum et al., 1999). ALSFRS-R scores range from 0 to 48 (the higher the score, the higher the functionality). Mean ALSFRS-R score was 31.2 ± 10.4 (range from 12 to 41). The ALS patients' demographic and clinical information are reported in **Table 1**.

Participants underwent a cognitive assessment focused on attention domains, in order to have individual baseline profiles. Two clinical neuropsychological tests were applied for the cognitive screening. The computerized test for attentional performance (TAP; Zimmermann and Fimm, 1995) was used to assess selective attention (SA) and working memory (WM) whereas the executive functions (EF) were assessed by means of the perseverative response scores obtained in the Wisconsin Card Sorting Test (WCST; Berg, 1948). Between the several clinical tests, the TAP includes a go-nogo task for SA (participants had to press a key when two target items were presented, while ignoring three distracter items) and a 2-back task for WM (numbers were presented on the screen and participants had to indicate the repetition of a number within an interval of three numbers by pressing a key). The WCST test consists of a card sorting game according to either color, shape or number. The sorting rule changes over time. Participants then have to rely on the outcome or feedback after each of their choice in order to infer the new rule in effect. Eight of the 13 ALS patients completed the full

protocol (psychological session and BCI session); the remaining five performed only the BCI session (one patient participated to the earlier study—Riccio et al., 2013). All Control subjects ($n = 13$) had both the psychological session and the BCI session.

Experimental Session
The experimental design consisted of two separate sessions (performed on two different days): the BCI session and the psychological session (see below for details).

BCI Session
Scalp potentials were acquired by means of a 16-channel amplifier (g.MOBILAB, g.tec, Austria) from eight active electrodes (g.Ladybird, g.tec, Austria) placed according to 10–10 international standard (Fz, Cz, Pz, Oz, P3, P4, PO7, and PO8; right ear lobe reference, left mastoid ground). This experimental choice was dictated by a reasonable trade-off between a not exhausting experimental procedure for ALS patients and a widely accepted eight electrodes configuration to ensure a P300 based-BCI successful control. Signals were digitized at 256 Hz. Stimulus paradigm and online delivery were managed by means of the BCI2000 framework (Schalk et al., 2004). A P3-speller (Farwell and Donchin, 1988) interface (6 by 6 matrix of alphanumeric items) was displayed full screen on a 15 computer screen, placed approximately at eye level and at a distance of 100 cm from the participant.

During the calibration phase (i.e., no feedback on performance), the subjects had to focus on 15 items forming three predefined words (3 runs; 5 items for each run). The target to focus on was shown to the participants by a single flash, after which rows and columns were randomly intensified for 125 ms, with an inter stimulus interval (ISI) of 125 ms. Participants were suggested to mentally count how many times that target was flashing. Calibration data were segmented into epochs lasting 800 ms (time 0 marked the stimulus onset) that were fed into a stepwise linear discriminant analysis (SWLDA) to determine the classifier coefficients (Krusienski et al., 2006) to be applied in the online BCI session. During the online phase (i.e., provision of feedback on performance), participants had to spell four

TABLE 1 | Participants' characteristics.

	ALS group	Control group
Age (years)	62.2 ± 13 (40–80)	55.3 ± 9 (44–68)
Sex (M/F)	8/5	9/4
Years of formal education	13.7 ± 5.1 (5–18)	13.4 ± 3.4 (8–18)
EF (impaired/not impaired)	5/6	5/8
EF (perseverative responses)	88.5 ± 16.5	96.6 ± 20.2
EF (total errors)	88.5 ± 14.3	93.6 ± 15.4
SA (impaired/not impaired)	3/7	1/12
SA (errors)	2.4 ± 2.9	0.8 ± 1.9
WM (impaired/not impaired)	4/5	1/10
WM (omissions)	3.8 ± 3.8	1.6 ± 2.0
ALSFRS-R	31.2 ± 10.4 (12–41)	-
Onset (S/B)	5/8	-
Time since diagnosis (mo)	26.8 ± 22.6	-

Demographic and clinical characteristics of the ALS group and Control group (means ± standard deviations, range). Abbreviations: F, female; M, male; SA, selective attention; EF, executive functions; WM, working memory; S, spinal; B, bulbar.

predefined (copy mode) words (4 runs; 5 items for each run; 20 characters in total). The classifier coefficients extracted from calibration data were applied to epochs grouped by stimulation classes (rows and columns) and averaged over stimulation sequences. The spelled letter was identified as the intersection between the row and the column exhibiting the maximum of the sum of scored features (Krusienski et al., 2006). As mentioned in the Introduction section, the stimulation sequences were not customized for each patient and thus we performed a static data collection (Mainsah et al., 2015). Specifically, a single item (e.g., letter) was intensified 20 times (10 sequences) before the next set of stimuli would start. Such stimulation sequence was maintained fixed for each subject.

Psychological Session

The individual temporal pattern of attentional resource allocation was tested by means of a rapid serial visual presentation (RSVP) paradigm which corresponds to an attentional blink (AB) paradigm in Kranczioch et al. (2007). In brief, it consisted of two target stimuli (T1 and T2) which were embedded in a stream of 16 or 19 distracter stimuli; each stream of stimuli (the equivalent of a single trial) was presented in the center of a monitor (white background). All stimuli were capital letters (letters F, K, Q, X, Z were excluded) and were presented pseudo-randomly (with a constraint that the same letter was not presented within three sequential positions) at central fixation (1 stimulus/100 ms; presentation rate at 10 Hz). As for target stimuli, T1 was a green capital letter randomly occurring as 4th, 5th, 6th or 7th item within a single stream. T2 was a black capital "X" which followed T1 on 80% of the trials according to four conditions (each occurring with a frequency of 20%): after no intervening distracters, after one, three or five intervening distracters. In 20% of the trials, T2 was not presented (5th condition). The distracter stimuli were black capital consonants.

Upon the stimulus stream delivery, participants were asked to answer the following questions: (1) whether the green letter (T1) was a vowel (T1 was a vowel on 50% of the trials); and (2) whether the black X (T2) was contained in the stimulus stream. In the case of ALS patients, the answers to the questions were given according to their residual motor activity (e.g., verbal response, head movements, eye movements).

Twenty practice trials preceded a total of 160 experimental trials (32 trials for each of the five T2 conditions); these latter were fully randomized within two presentation blocks separated by a pause of 5 min.

DATA ANALYSIS

All acquired data were preprocessed as follows. High and low pass filters (4th order Butterworth filter) were applied with a cut off frequency of 1 Hz and 20 Hz, respectively. EEG signals with peak amplitude higher than 70 μV or lower than -70 μV were removed. Data were then segmented into epochs (time 0 denoted the stimulus onset) lasting 800 ms and 1000 ms for the BCI and ERPs analysis, respectively. Both target and non-target stimulus-related epochs were considered.

BCI Performance Analysis

The BCI online accuracy was expressed as the percentage of correct selections (i.e., the ratio between the number of correct selections and the total number of selections). Furthermore, an offline estimation of both the accuracy and the information transfer rate (ITR, Wolpaw et al., 2002) was performed in order to account for the online performance inter-subject variability that was "hidden" by the static modality of data collection.

To estimate the offline accuracy, a baseline correction was performed based on the mean amplitude of signal within the 200 ms pre-stimulus interval. The offline accuracy was then calculated for each stimulation sequence by means of a 7-fold cross-validation technique according to which six runs were used as training dataset to extract SWLDA classifier parameters and one run was used as testing dataset. A mean accuracy value for each stimulation sequence was obtained by averaging the values resulting from the seven iterations.

The ITR (bits/min) was estimated for each subject and each stimulation sequence based on the definition of bit-rate as in Wolpaw et al. (2002) and multiplying the bit-rate by the speed of selection (selections/minute). The individual highest ITR value was considered. Output metric resulting from this computation will be reported as ITR (0–800 ms).

The relative contribution of the N2 and P3 ERP components to the BCI accuracy was investigated offline as follows. Differences in the amplitudes of ERPs that were elicited by the stimulus types (target vs. non-target) were quantified using the coefficient of determination R2. We considered the epochs relative to all seven runs. The R2 values range from 0 to 1, wherein higher values correspond to larger explained variances. A *signed* R2 index was introduced to account for the different polarity of ERPs (N2 and P3) and was derived by multiplying R2 by the sign of the slope of the corresponding linear model which was positive when the amplitudes of the ERPs that were elicited by the target stimuli were higher than those elicited by non-target stimuli and vice versa (as in Aloise et al., 2012). The mean R^2 values were computed within two different temporal intervals for the N2 and P3 components that ranged from 100 ms to 400 ms for the N2 (negative values) and from 250 ms to 550 ms for the P3 (positive values).

As for the ITR, its values were computed by segmenting data into epochs between 0 ms and 550 ms after the stimulus onset—ITR (0–550 ms)—to ensure that the temporal interval would include both N2 and P3 ERP components.

ERPs Analysis

We focused the ERP analysis on the N2 as the earliest ERP that reliably correlates with visual awareness (Visual Awareness Negativity; Railo et al., 2011) and the P3 as associated with conscious access to the content of conscious vision (Raffone et al., 2014). Hence, these two ERP components can be considered a reliable reflection of attentional stimulus processing.

In this offline ERP analysis, the epochs in which a target stimulus occurred within the 500 ms preceding the stimulus onset were removed in order to reduce the contamination between consecutive epochs and the ERP overlapping (Treder and Blankertz, 2010). Target and non-target ERP waveforms

FIGURE 1 | Box plots illustrate the comparison between groups relative to the online performance **(A)** and information transfer rate (ITR) **(B)**. Experimental group = amyotrophic lateral sclerosis (ALS) patients. *Indicates significant difference ($p < 0.05$).
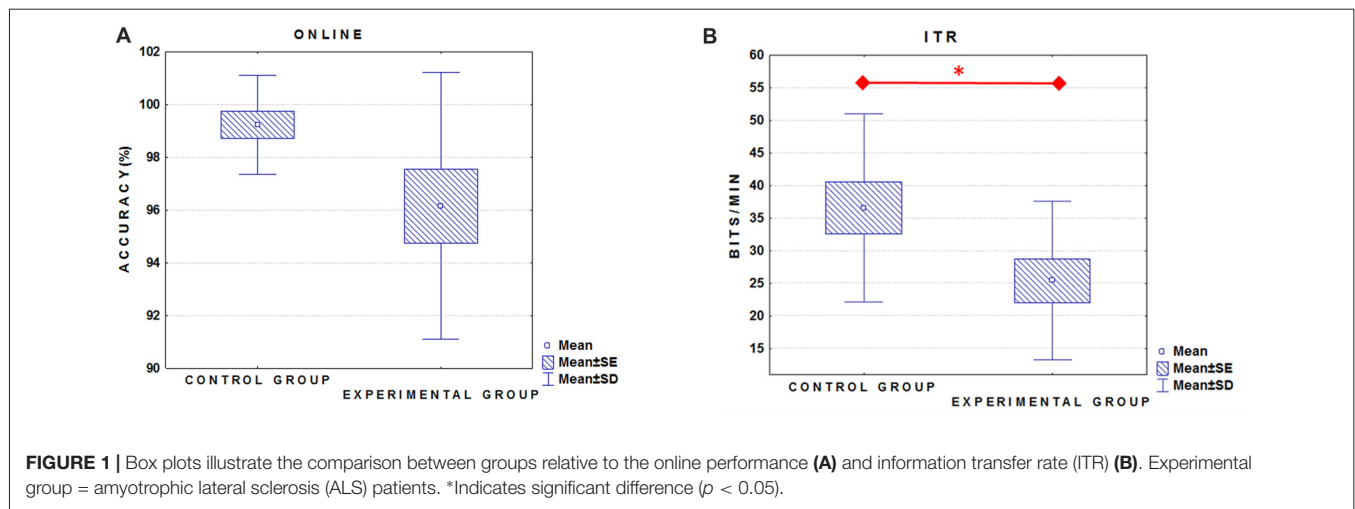
were obtained by averaging the epochs relative to each run. The ERP waveforms were obtained from a sample by sample contrast between the non-target and target ERP waveform amplitude (i.e., the difference between target and non-target).

The mean of both the N2 and P3 amplitude and the centroid latency (Luck, 2005) were obtained from all data sets (seven BCI session runs). The P3 mean amplitude was calculated by averaging the voltage of all positive points preceded or succeeded by positive values between 250 ms and 550 ms after the stimulus onset whereas for the N2 mean amplitude we averaged the voltage of all negative points preceded or succeeded by negative values between 100 ms and 400 ms after the stimulus onset (Clayson et al., 2013). The P3 and N2 centroid latencies were set as the time to which the area under the curves was divided into equal halves. Finally, the mean amplitude and the centroid latency of the P3 waves (P3-MA and P3-CL, respectively) recorded from Fz, Cz and Pz and the Mean Amplitude and the Centroid Latency of the N2 waves (N2-MA and N2-CL, respectively) recorded from PO7, PO8 and Oz were subjected to the statistical analysis.

Psychological Paradigm Data Analysis

As for the RSVP data set, the accuracy of T1 and T2 detection (T1%; T2%) was estimated (T2% was considered only in trials in which T1 had been correctly identified). T1% was considered an index of participants' temporal attentional filtering capacity and T2% was considered as an index of the capability to adequately update the attentive filter (Riccio et al., 2013) in the temporal dynamics of the attentional selection.

Statistical Analysis

Between-group (ALS and Control) differences in terms of BCI control performance were evaluated as follows. A (non-parametric) Mann-Whitney U test was applied to assess the between-group difference in BCI online accuracy (accuracy scores not normally distributed). A Student's T-test was applied to assess the between-group difference in terms of the ITR scores. The contribution of N2-R2 and P3-R2 to the BCI performance was assessed

for each group (ASL and Control) by means of two linear regression analyses with the ITR (0–550 ms) as the dependent variable and the N2-R2 and the P3-R2 as independent variables.

The (non-parametric) Spearman's rank order correlation was applied to investigate the possible correlation between ALSFRS-R scores (not normally distributed) and the ITR values.

To investigate whether ALS patients showed differences with respect to Control group in attention processing during the BCI task, we conducted two MANOVAs to determine the effect of group (independent variable) on both P3-MA and P3-CL (dependent variables). The same analysis (two MANOVAs) was performed to determine the effect of group on both N2-MA and N2-CL.

To investigate whether the temporal pattern of attentional resource allocation would be correlated with the BCI performance level (ITR), we performed a correlation analysis (Pearson's correlation coefficients) between the ITR (0–800 ms), the P3-MA in Pz and T1% and T2%. Such correlation was sought either by pooling all data from ALS and Control groups and by considering only ALS group data.

The existence of alterations in attentional resources allocation in the ALS group was assessed by means of a MANOVA to test the effect of group (independent variable) on T1% and T2% (dependent variables).

RESULTS

The ALS and Control groups did not show significant differences as regard demographic characteristics (Student's T-test; $t_{(24)} = 1.6$; $p = 0.13$ and $t_{(24)} = 0.17$; $p = 0.85$ for age and "years of formal education", respectively) and clinical assessment focused on selective attention (SA; error scores, $\chi^2 = 1.96$; $p = 0.16$), working memory (WM; errors $\chi^2 = 3.29$; $p = 0.07$) and executive functions (EF; $\chi^2 = 0.119$; $p = 0.729$).

We did not find a significant between-group difference in the online accuracy (**Figure 1A**; ALS group mean = 96.1% ± 5; Control group mean = 99.2% ± 2; $U = 55.5$; $p = 0.13$). On the contrary, ITR (0–800 ms) was significantly higher in the
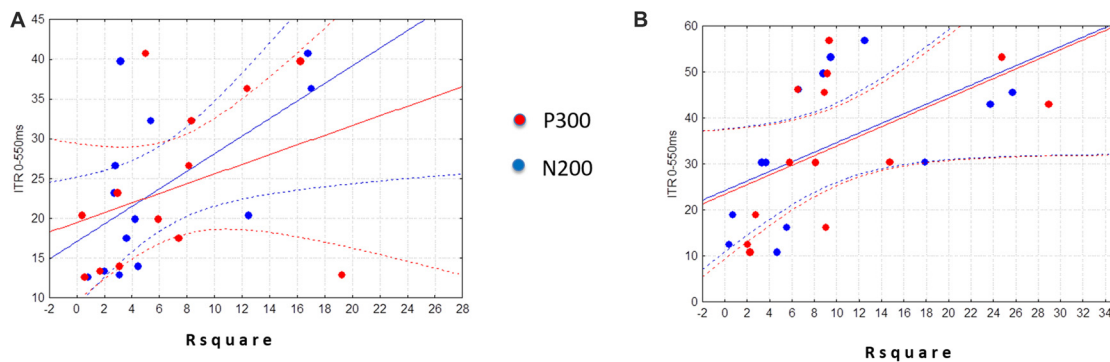
FIGURE 2 | Scatter plots illustrate the relationship of the ITR (0–550 ms) with the N2Rsquare (blue dots) and the P3Rsquare (red dots) in the ALS (experimental group) **(A)** and in the Control group **(B)**. Note that regression analysis was performed by considering only the absolute values of R2 (i.e., no *signed* R2 was considered).

Control group (36.6 ± 14.5 bits/min) as compared to ALS group (mean = 25.4 ± 12.1 bits/min; $t_{(24)} = 2.1$, $p \le 0.05$; **Figure 1B**).

The linear regression analysis ($p = 0.11$) revealed that ITR (0–550) was not significantly predicted neither by the N2-R2 values ($\beta = 0.33$; $p = 0.34$) nor the P3-R2 values ($\beta = 0.32$; $p = 0.34$) in the Control group (**Figure 2B**). Differently, in the ALS group we found that the linear regression was significant ($F_{(2,10)} = 4.3526$, $p < 0.05$). Specifically, only the N2-R2 was significantly predictive of the ITR (0–550 ms; $\beta = 0.59$, $p < 0.05$; P3-R2 = $\beta = 0.34$, $p = 0.17$; **Figure 2A**).

We found that the ITR and ALSFRS-R scores showed a high tendency to correlate which did not reach a significance ($p = 0.06$; $r = 0.52$). Such tendency, however, suggests that the degree of disability due to ALS could influence the BCI performance in a detrimental way.

No significant differences were found between ALS and Control groups (MANOVAs) in P3-MA ($\lambda = 0.71$; $F_{(3,22)} = 2.9$, $p = 0.05$), N2-MA ($\lambda = 0.88$; $F_{(3,22)} = 0.9$; $p = 0.4$) and N2-CL ($\lambda = 0.89$; $F_{(3,22)} = 1.0$; $p = 0.4$). On the contrary, the MANOVA returned a significant between-group difference in P3-CL ($\lambda = 0.59$; $F_{(3,22)} = 5.0$; $p < 0.01$) values over Fz ($p < 0.001$; Control group mean = 363.17 ± 20.14 ms; ALS group mean = 409.4 ± 38.07 ms; Bonferroni corrected) and Cz electrodes ($p < 0.05$; Control group mean = 368.60 ± 27.62 ms; ALS group mean = 396.68 ± 32.66 ms; Bonferroni corrected) with longer CL in ALS patients with respect to Controls. No significant differences were found in P3-CL values over Pz ($p = 0.3$; control group mean = 379.71 ± 37.83 ms; ALS group mean = 394.4 ± 38.82 ms; Bonferroni corrected).

As illustrated in **Figure 3**, the visual inspection of P3 topography indicates a prevalent frontal distribution of P3 in ALS whereas a parietal distribution is observed in Control group.

The analysis of the relation between cognitive substrates and BCI performance as measured by means of RSVP and BCI data returned a significant positive correlation between T1% and the ITR ($r = 0.51$; $p < 0.05$) and T1% the P3-MA in Pz ($r = 0.63$; $p < 0.05$) when considering all group data (ALS and Control data pooled). No significant correlation was found between T2%

and ITR ($r = 0.32$; $p = 0.15$) and between T2% and P3-MA ($r = 0.23$; $p = 0.31$). When considering only the ALS group, the same analysis unveiled significant positive correlation between T1% and ITR ($r = 0.71$; $p < 0.05$) and T1% and P3-MA ($r = 0.78$; $r < 0.05$) whereas T2% and ITR ($r = 0.66$; $p = 0.07$) and T2% and P3-MA ($r = 0.39$; $p = 0.31$) did not show significant correlation.

The MANOVA ($F = 4.4$; $p < 0.05$) revealed that the T1% values were significantly lower in ALS as compared to Control group (Control group mean $89.7 \pm 0.8\%$; ALS group mean = $79.4 \pm 10\%$; $p = 0.01$; Bonferroni corrected). No significant difference was found in T2% (control group mean = $63.6 \pm 2\%$; ALS group mean = $67.8 \pm 23\%$; $p = 0.6$).

Discussion

This study aimed at investigating whether ALS patients showed differences in the ability to control a P3-speller BCI system with respect to healthy subjects. We focused on the attention processing involved in the delivering of the visual BCI stimulation paradigm, in order to further (Riccio et al., 2013) elucidate if and how such cognitive abilities would be altered in ALS patients and eventually would account for patients' BCI control capacity. We hypothesized that the capacity to accomplish a P3-speller task was decreased in ALS patients and that they would have shown an alteration in the visual attention processing as elicited during the P3-based BCI control. To test our hypothesis, we compared two groups of participants (ALS patients vs. Control) in terms of performance in P3-speller control and with regard to the earliest ERP components such as N2 and P3 which are correlated with visual awareness (Railo et al., 2011; Raffone et al., 2014).

First, we found that the ALS patients showed a significantly lower ITR in the P3-speller BCI task with respect to Controls whereas the online performance was comparable between the two groups.

This finding is not in line with what reported by McCane et al. (2015). According to their study severely disabled ALS patients and age-matched healthy controls showed similar P3-based BCI performance in terms of maximum accuracy, communication rate and bit rate. Several differences between these two studies

FIGURE 3 | P300 topography and waveforms in ALS (Experimental Group) and Control group. Traces in the middle panel represent the grand average of the difference between target and non-target electroencephalography (EEG) amplitude as a function of time (interval between 0 = stimulus onset and 1000 ms) recorded for ALS (*n* = 13 patients; solid line) and Control group (*n* = 13 subjects; dotted line) during the brain-computer interface (BCI) session, over different electrode positions (left labels). The maps represent the scalp topographical distribution in Control (left) and ALS (right) group of the P3 and N2 centroid latency (grand average) over Fz, Cz and Pz and over PO7, PO8 and Oz respectively.

might account for the apparent discrepancy on the ability to master a P3-based BCI by ALS patients. First, the pattern of visual stimulation (checkerboard in McCane et al., 2015) that is well known to remarkably influence the P3-based BCI control performance (Townsend et al., 2010) and related ERPs characteristics (Kaufmann et al., 2011). Second, the ALS clinical severity that was higher in ALS population of McCane (ALSFRS-R scores = 9.4 ± 9.5 SD) with respect to our population (ALSFRS-R scores = 31.2 ± 10.4). In this regard, we found a remarkable (but not significant) correlation between the ITR and the ALSFRS-R scores that suggests a direct relation between the degree of clinical disability and the ability to use a P3-based BCI. Finally, in McCane et al. (2015) study, the ALS and healthy participants were not matched for years of formal education and this is a variable also accounting for the level of cognitive task performance (Ardila et al., 2000).

In addition to this, the overall methods (and metrics) to estimate the P3-based BCI performance are not directly comparable between the two studies. We "only" found the (offline) ITR as a distinctive metric of the ability to use a P3-based BCI in ALS with respect to Control group.

In the P3-speller task, the act of focusing attention on the target letter modulates the visual processing of the stimulus. Our ERP findings indicate that the P3 mean latency was significantly higher in ALS with respect to control group while no difference was found in the N2 parameters between the two groups. The finding of a delayed P3 associated with a "normal" N2 (i.e., physiological stimulus categorization process) in ALS can be interpreted as a *delay* occurring in the post-perceptual stage of the stimulus attentional processing (Duncan-Johnson and Kopell, 1981) that is, the variation in the attention modulation during the stimulus visual processing observed in ALS would occur when stimulus perception is complete, the target is categorized and its storing in WM has taken place. We found no significant between-group differences in P3 mean amplitude; this latter parameter can be considered as a measure of the attention allocation resources (Donchin, 1981). This finding allows us to speculate that in ALS patients the overall alteration of attention modulation during a P3-speller task is only related to the time of processing but not to the resource allocation.

The P3 wave component showed a frontal topography in the ALS group as compared to the parietal distribution observed in

the Control group (**Figure 3**). This finding is in line with previous findings reported by McCane et al. (2015). We interpret this difference in P3 topography as possibly related to the P3a and P3b components that have different generators and thus, different topography (Courchesne et al., 1975): the *frontal* P3 component observed in the ALS group would represent the P3a whereas the *parietal* P3 present in Controls would better represent the P3b component.

Fabiani and Friedman (1995) suggested that during the process of learning a task, the P3a is elicited by novel stimuli that do not require a memory template. In contrast, the P3b is elicited when the stimulus memory template is created. One can speculate that the *frontal* P3a topography in ALS might reflect an increase in the frontal activity related to a more rapid decay of the memory templates (of the stimuli processed) that would make it more difficult to create and maintain an adequate template for the target stimulus (Fabiani and Friedman, 1995; Fabiani et al., 1998).

We found that the N2-R2 and not the P3-R2 coefficient significantly predicted the BCI accuracy only in ALS group, accounting for the 59% of the ITR variance. Based on the assumption that such coefficients mostly returned the contribution of N2 and of P3 waves to the BCI classification performance, these findings indicate that in ALS patients the N2 elicited during the P3-speller task would have a major role with respect to P3 in successful target selection.

The presence of a jitter in the P3 latency has been described in healthy subjects controlling a P3-based BCI system and its magnitude would be correlated with the online performance (Thompson et al., 2013; Aricò et al., 2014). Although it remains to be demonstrated that such P3 wave jitter exists and to what extent it might influence the P3-speller BCI task performance in ALS patients, one can speculate that our observed *unbalance* in the contribution of N2 and P3 components to successful BCI performance in favor of N2 might reside in the cross-relationship between the P3 latency jitter and delayed visual processing phenomena. Hence, further investigations in this regard are of utmost relevance to address sensible design of future ERP-based BCIs for ALS user candidates.

According to our previous findings (Riccio et al., 2013), the temporal aspect of the SA investigated by means of a RSVP task and measured as T1% (i.e., the ability to maintain the attentional filter active during the selection of a target within a range of time) was found to be related to both the BCI performance and P3 amplitude in ALS patients.

In the present study, we confirmed that the temporal filter in attention processing of visual stimuli in ALS patients was altered, by directly compare the T1% and T2% values obtained from ALS and Control group. Specifically, we found that the capability to detect T1 (but not T2) was lower in the ALS group.

As such, this finding is consistent with that of a delay in the P3 latency which reflects a deficit in the temporal aspect (i.e., post-perceptual stage of the stimulus attentional processing) of the context update. Taken altogether, these findings clearly indicate the existence of an alteration in the temporal aspects of the visual stimulus processing as presented in a "conventional" P3-speller matrix in ALS population and that this time-related

alteration in the capacity to temporally process visual stimuli does influence the rate of success in BCI control.

Our findings might lay the groundwork for future clarification of some of the relevant issues in the actual deployment of ERPs-based BCI for communication to ALS users, such as the impact of end-users' cognitive profile in designing user-centered (Liberati et al., 2015; Nijboer, 2015) and reliable BCI systems (Kübler et al., 2013).

Study Limitation

Some limitations pertaining different aspects of this study deserve to be mentioned. First, our ALS population does not include ALS patients in a complete locked-in state (LIS). Although this restriction in the inclusion criterion was mandatory to allow the cognitive screening, it prevents any generalization of our findings to those patients with no means of communication (i.e., complete LIS). In this framework, our study suggests a possible role of the cognitive assessment to be performed in ALS patients before they would be in a LIS condition, including the specific cognitive abilities identified here as critical for P3-based BCI usage.

Second, our findings relative to BCI factors influencing BCI performance allow us to make inferences only regarding the control of a P3-speller in a group of ALS patients and cannot be generalized to the control of other BCIs. It is conceivable that when exploiting different features to control other BCIs, temporal aspects of attention would not have a comparable role. An example of "alternative" features to P3 would be the N400 wave, involved in the elaboration of meaningful stimuli such as face recognition (Kaufmann et al., 2013b). In addition, several single-case studies have shown significant differences in classification performance depending on the sensory modality of the ERP-based BCIs that participants controlled (Kaufmann et al., 2013a; Schreuder et al., 2013).

Third, attention is a complex domain of the cognitive functions (Posner, 1975) and its substrates can be measured in different ways; this prevents from a direct comparison of results obtained within different contexts and with different behavioral and neurophysiological approaches. For instance, attention substrates measured with different tests (i.e., Cognitrone, Schuhfried, 2007) were not found to be precursors of performance in controlling sensorimotor–based BCIs, differently from visuo-motor coordination (Hammer et al., 2012, 2014).

CONCLUSION

This study involved a group of participants with ALS and a group of healthy participants matched for age and years of formal education. Our results showed that both the capacity to accomplish the P3-speller task and the timing of the allocation of attentional resources in the post-perceptual stage of stimulus processing were altered in ALS patients. Furthermore, we confirmed that the capacity to temporally filter a target stimulus within a stream of stimuli was related to a lower capacity for ALS to control a P3-speller.

Developing AT devices that restore communication in people with severe motor disabilities is a central issue of BCI research

(Millán et al., 2010). Current BCI systems do not address the heterogeneity of the end-users often due to the lack of customizability and adaptability to their cognitive capabilities. This study contributes to the knowledge needed for developing a new class of BCI specifically designed by taking into account the influence of the cognitive characteristics of end-users on BCI usage.

AUTHOR CONTRIBUTIONS

AR was responsible for experimental design, data collection, analysis of data and manuscript writing. FS was responsible for data acquisition and analysis. LS was responsible for behavioral assessment (attention tasks). AP and MI were responsible for the patients' enrolment. MO-B, DM and FC supervised the overall experimental design implementation, data interpretation and manuscript editing.

FUNDING

REFERENCES

Aloise, F., Aricò, P., Schettini, F., Riccio, A., Salinari, S., Mattia, D., et al. (2012). A covert attention P300-based brain-computer interface: Geospell. *Ergonomics* 55, 538–551. doi: 10.1080/00140139.2012.661084

Ardila, A., Ostrosky-Solis, F., Rosselli, M., and Gómez, C. (2000). Age-related cognitive decline during normal aging: the complex effect of education. *Arch. Clin. Neuropsychol.* 15, 495–513. doi: 10.1016/s0887-6177(99)00040-2

Aricò, P., Aloise, F., Schettini, F., Salinari, S., Mattia, D., and Cincotti, F. (2014). Influence of P300 latency jitter on event related potential-based brain-computer interface performance. *J. Neural Eng.* 11:035008. doi: 10.1088/1741-2560/11/3/035008

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *J. Gen. Psychol.* 39, 15–22. doi: 10.1080/00221309.1948.9918159

Cedarbaum, J. M., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., et al. (1999). The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. *J. Neurol. Sci.* 169, 13–21. doi: 10.1016/S0022-510X(99)00210-5

Christidi, F., Zalonis, I., Smyrnis, N., and Evdokimidis, I. (2012). Selective attention and the three-process memory model for the interpretation of verbal free recall in amyotrophic lateral sclerosis. *J. Int. Neuropsychol. Soc.* 18, 809–818. doi: 10.1017/s1355617712000562

Clayson, P. E., Baldwin, S. A., and Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology* 50, 174–186. doi: 10.1111/psyp.12001

Courchesne, E., Hillyard, S. A., and Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalogr. Clin. Neurophysiol.* 39, 131–143. doi: 10.1016/0013-4694(75)90003-6

Donchin, E. (1981). Presidential address, 1980. Surprise!. . .Surprise? *Psychophysiology* 18, 493–513. doi: 10.1111/j.1469-8986.1981.tb01815.x

Duncan-Johnson, C. C., and Kopell, B. S. (1981). The stroop effect: brain potentials localize the source of interference. *Science* 214, 938–940. doi: 10.1126/science.7302571

Fabiani, M., and Friedman, D. (1995). Changes in brain activity patterns in aging: the novelty oddball. *Psychophysiology* 32, 579–594. doi: 10.1111/j.1469-8986.1995.tb01234.x

Fabiani, M., Friedman, D., and Cheng, J. C. (1998). Individual differences in P3 scalp distribution in older adults, and their relationship to frontal lobe function. *Psychophysiology* 35, 698–708. doi: 10.1111/1469-8986.3560698

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Georgiou-Karistianis, N., Tang, J., Vardy, Y., Sheppard, D., Evans, N., Wilson, M., et al. (2007). Progressive age-related changes in the attentional blink paradigm. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* 14, 213–226. doi: 10.1080/13825580500320681

Hammer, E. M., Halder, S., Blankertz, B., Sannelli, C., Dickhaus, T., Kleih, S., et al. (2012). Psychological predictors of SMR-BCI performance. *Biol. Psychol.* 89, 80–86. doi: 10.1016/j.biopsycho.2011.09.006

Hammer, E. M., Kaufmann, T., Kleih, S. C., Blankertz, B., and Kübler, A. (2014). Visuo-motor coordination ability predicts performance with brain-computer interfaces controlled by modulation of sensorimotor rhythms (SMR). *Front. Hum. Neurosci.* 8:574. doi: 10.3389/fnhum.2014.00574

Kaufmann, T., Holz, E. M., and Kübler, A. (2013a). Comparison of tactile, auditory, and visual modality for brain-computer interface use: a case study with a patient in the locked-in state. *Front. Neurosci.* 7:129. doi: 10.3389/fnins.2013.00129

Kaufmann, T., Schulz, S. M., Köblitz, A., Renner, G., Wessig, C., and Kübler, A. (2013b). Face stimuli effectively prevent brain-computer interface inefficiency in patients with neurodegenerative disease. *Clin. Neurophysiol.* 124, 893–900. doi: 10.1016/j.clinph.2012.11.006

Kaufmann, T., Schulz, S. M., Grünzinger, C., and Kübler, A. (2011). Flashing characters with famous faces improves ERP-based brain-computer interface performance. *J. Neural Eng.* 8:056016. doi: 10.1088/1741-2560/8/5/056016

Kleih, S., Kaufmann, T., Zickler, C., Halder, S., Leotta, F., Cincotti, F., et al. (2011). Out of the frying pan into the fire—the P300 based BCI faces real world challenges. *Prog. Brain Res.* 194, 27–46. doi: 10.1016/b978-0-444-53815-4.00019-4

Kranczioch, C., Debener, S., Maye, A., and Engel, A. K. (2007). Temporal dynamics of access to consciousness in the attentional blink. *Neuroimage* 37, 947–955. doi: 10.1016/j.neuroimage.2007.05.044

Kranczioch, C., Debener, S., Schwarzbach, J., Goebel, R., and Engel, A. K. (2005). Neural correlates of conscious perception in the attentional blink. *Neuroimage* 24, 704–714. doi: 10.1016/j.neuroimage.2004.09.024

Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., et al. (2006). A comparison of classification techniques for the P300 Speller. *J. Neural Eng.* 3, 299–305. doi: 10.1088/1741-2560/3/4/007

Kübler, A., Holz, E. M., Riccio, A., Zickler, C., Kaufmann, T., Kleih, S. C., et al. (2014). The user-centered design as novel perspective for evaluating the usability of BCI-controlled applications. *PLoS One* 9:e112392. doi: 10.1371/journal.pone.0112392

Kübler, A., Mattia, D., Rupp, R., and Tangermann, M. (2013). Facing the challenge: bringing brain-computer interfaces to end-users. *Artif. Intell. Med.* 59, 55–60. doi: 10.1016/j.artmed.2013.08.002

Liberati, G., Pizzimenti, A., Simione, L., Riccio, A., Schettini, F., Inghilleri, M., et al. (2015). Developing brain-computer interfaces from a user-centered perspective: assessing the needs of persons with amyotrophic lateral sclerosis, caregivers, and professionals. *Appl. Ergon.* 50, 139–146. doi: 10.1016/j.apergo.2015.03.012

Lomen-Hoerth, C., Murphy, J., Langmore, S., Kramer, J. H., Olney, R. K., and Miller, B. (2003). Are amyotrophic lateral sclerosis patients cognitively normal? *Neurology* 60, 1094–1097. doi: 10.1212/01.WNL.0000055861.95202.8D

Luck, S. (2005). An introduction to the event-related potential technique (Cognitive Neuroscience). A Bradford Book Available online at: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09–20andpath=ASIN/0262621967 [Accessed November 4, 2015].

Mainsah, B., Collins, L., Colwell, K., Sellers, E., Ryan, D., Caves, K., et al. (2015). "ALS population assessment of a dynamic stopping algorithm implementation for P300 spellers," in *Brain-Computer Interface Research*, eds C. Guger, G. Müller-Putz and B. Allison (Cham: Springer International Publishing),

79–87. Available online at: http://link.springer.com/10.1007/978-3-319-25190-5_8 [Accessed December 22, 2015].

Marchetti, M., and Priftis, K. (2015). Brain-computer interfaces in amyotrophic lateral sclerosis: a metanalysis. *Clin. Neurophysiol.* 126, 1255–1263. doi: 10.1016/j.clinph.2014.09.017

Marchetti, M., Piccione, F., Silvoni, S., Gamberini, L., and Priftis, K. (2013). Covert visuospatial attention orienting in a brain-computer interface for amyotrophic lateral sclerosis patients. *Neurorehabil. Neural Repair* 27, 430–438. doi: 10.1177/1545968312471903

McCane, L. M., Heckman, S. M., McFarland, D. J., Townsend, G., Mak, J. N., Sellers, E. W., et al. (2015). P300-based brain-computer interface (BCI) event-related potentials (ERPs): people with amyotrophic lateral sclerosis (ALS) vs. age-matched controls. *Clin. Neurophysiol.* 126, 2124–2131. doi: 10.1016/j.clinph.2015.01.013

McCane, L. M., Sellers, E. W., McFarland, D. J., Mak, J. N., Carmack, C. S., Zeitlin, D., et al. (2014). Brain-computer interface (BCI) evaluation in people with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Frontotemporal Degener.* 15, 207–215. doi: 10.3109/21678421.2013.865750

Millán, J. D. R., Rupp, R., Müller-Putz, G. R., Murray-Smith, R., Giugliemma, C., Tangermann, M., et al. (2010). Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. *Front. Neurosci.* 4:161. doi: 10.3389/fnins.2010.00161

Nijboer, F. (2015). Technology transfer of brain-computer interfaces as assistive technology: barriers and opportunities. *Ann. Phys. Rehabil. Med.* 58, 35–38. doi: 10.1016/j.rehab.2014.11.001

Nijboer, F., Birbaumer, N., and Kübler, A. (2010). The influence of psychological state and motivation on brain-computer interface performance in patients with amyotrophic lateral sclerosis—a longitudinal study. *Front. Neurosci.* 4:55. doi: 10.3389/fnins.2010.00055

Nijboer, F., Sellers, E. W., Mellinger, J., Jordan, M. A., Matuz, T., Furdea, A., et al. (2008). A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clin. Neurophysiol.* 119, 1909–1916. doi: 10.1016/j.clinph.2008.03.034

Piccione, F., Giorgi, F., Tonin, P., Priftis, K., Giove, S., Silvoni, S., et al. (2006). P300-based brain computer interface: reliability and performance in healthy and paralysed participants. *Clin. Neurophysiol.* 117, 531–537. doi: 10.1016/j.clinph.2005.07.024

Posner, M. I. (1975). "Psychobiology of attention," in *Handbook of Psychobiology*, eds M. S. Gazzaniga and C. Blakemore (New York, NY: Academic Press), 441–480.

Radakovic, R., Stephenson, L., Newton, J., Crockford, C., Swingler, R., Chandran, S., et al. (2017). Multidimensional apathy and executive dysfunction in amyotrophic lateral sclerosis. *Cortex* 94, 142–151. doi: 10.1016/j.cortex.2017.06.023

Raffone, A., Srinivasan, N., and van Leeuwen, C. (2014). Perceptual awareness and its neural basis: bridging experimental and theoretical paradigms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130203. doi: 10.1098/rstb.2013.0203

Railo, H., Koivisto, M., and Revonsuo, A. (2011). Tracking the processes behind conscious perception: a review of event-related potential correlates of visual consciousness. *Conscious. Cogn.* 20, 972–983. doi: 10.1016/j.concog.2011.03.019

Riccio, A., Holz, E. M., Aricò, P., Leotta, F., Aloise, F., Desideri, L., et al. (2015). Hybrid P300-based brain-computer interface to improve usability for people with severe motor disability: electromyographic signals for error correction during a spelling task. *Arch. Phys. Med. Rehabil.* 96, S54–S61. doi: 10.1016/j.apmr.2014.05.029

Riccio, A., Pichiorri, F., Schettini, F., Toppi, J., Risetti, M., Formisano, R., et al. (2016). Interfacing brain with computer to improve communication and rehabilitation after brain damage. *Prog. Brain Res.* 228, 357–387. doi: 10.1016/bs.pbr.2016.04.018

Riccio, A., Simione, L., Schettini, F., Pizzimenti, A., Inghilleri, M., Belardinelli, M. O., et al. (2013). Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Front. Hum. Neurosci.* 7:732. doi: 10.3389/fnhum.2013.00732

Ringholz, G. M., Appel, S. H., Bradshaw, M., Cooke, N. A., Mosnik, D. M., and Schulz, P. E. (2005). Prevalence and patterns of cognitive impairment in sporadic ALS. *Neurology* 65, 586–590. doi: 10.1212/01.wnl.0000172911.39167.b6

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., and Wolpaw, J. R. (2004). BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* 51, 1034–1043. doi: 10.1109/tbme.2004.827072

Schettini, F., Riccio, A., Simione, L., Liberati, G., Caruso, M., Frasca, V., et al. (2015). Assistive device with conventional, alternative, and brain-computer interface inputs to enhance interaction with the environment for people with amyotrophic lateral sclerosis: a feasibility and usability study. *Arch. Phys. Med. Rehabil.* 96, S46–S53. doi: 10.1016/j.apmr.2014.05.027

Schreuder, M., Riccio, A., Risetti, M., Dähne, S., Ramsay, A., Williamson, J., et al. (2013). User-centered design in brain-computer interfaces-a case study. *Artif. Intell. Med.* 59, 71–80. doi: 10.1016/j.artmed.2013.07.005

Schuhfried, G. (2007). *Cognitrone—COG.* Mödling: Schuhfried.

Sellers, E. W., and Donchin, E. (2006). A P300-based brain-computer interface: initial tests by ALS patients. *Clin. Neurophysiol.* 117, 538–548. doi: 10.1016/j.clinph.2005.06.027

Silvoni, S., Cavinato, M., Volpato, C., Ruf, C. A., Birbaumer, N., and Piccione, F. (2013). Amyotrophic lateral sclerosis progression and stability of brain-computer interface communication. *Amyotroph. Lateral Scler. Front. Degener.* 14, 390–396. doi: 10.3109/21678421.2013.770029

Strutt, A. M., Palcic, J., Wager, J. G., Titus, C., Macadam, C., Brown, J., et al. (2012). Cognition, behavior, and respiratory function in amyotrophic lateral sclerosis. *ISRN Neurol.* 2012:912123. doi: 10.5402/2012/912123

Thompson, D. E., Gruis, K. L., and Huggins, J. E. (2014). A plug-and-play brain-computer interface to operate commercial assistive technology. *Disabil. Rehabil. Assist. Technol.* 9, 144–150. doi: 10.3109/17483107.2013.785036

Thompson, D. E., Warschausky, S., and Huggins, J. E. (2013). Classifier-based latency estimation: a novel way to estimate and predict BCI accuracy. *J. Neural Eng.* 10:016006. doi: 10.1088/1741-2560/10/1/016006

Townsend, G., LaPallo, B. K., Boulay, C. B., Krusienski, D. J., Frye, G. E., Hauser, C. K., et al. (2010). A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns. *Clin. Neurophysiol.* 121, 1109–1120. doi: 10.1016/j.clinph.2010.01.030

Treder, M. S., and Blankertz, B. (2010). (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behav. Brain Funct.* 6:28. doi: 10.1186/1744-9081-6-28

Volpato, C., Prats Sedano, M. A., Silvoni, S., Segato, N., Cavinato, M., Merico, A., et al. (2016). Selective attention impairment in amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Frontotemporal Degener.* 17, 236–244. doi: 10.3109/21678421.2016.1143514

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791. doi: 10.1016/S1388-2457(02)00057-3

Zickler, C., Riccio, A., Leotta, F., Hillian-Tress, S., Halder, S., Holz, E., et al. (2011). A brain-computer interface as input channel for a standard assistive technology software. *Clin. EEG Neurosci.* 42, 236–244. doi: 10.1177/155005941104200409

Zimmermann, P., and Fimm, B. (1995). *Test for Attentional Performance (TAP).* Herzogenrath: PsyTest. pp. 76–77.

Detecting and Quantifying Mind Wandering during Simulated Driving

Carryl L. Baldwin[1]*, Daniel M. Roberts[1], Daniela Barragan[1], John D. Lee[2], Neil Lerner[3] and James S. Higgins[4]

[1]Department of Psychology, George Mason University, Fairfax, VA, United States, [2]Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, United States, [3]Center for Transportation, Technology and Safety Research, Westat, Rockville, MD, United States, [4]Office of Behavioral Safety Research, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, DC, United States

Mind wandering is a pervasive threat to transportation safety, potentially accounting for a substantial number of crashes and fatalities. In the current study, mind wandering was induced through completion of the same task for 5 days, consisting of a 20-min monotonous freeway-driving scenario, a cognitive depletion task, and a repetition of the 20-min driving scenario driven in the reverse direction. Participants were periodically probed with auditory tones to self-report whether they were mind wandering or focused on the driving task. Self-reported mind wandering frequency was high, and did not statistically change over days of participation. For measures of driving performance, participant labeled periods of mind wandering were associated with reduced speed and reduced lane variability, in comparison to periods of on task performance. For measures of electrophysiology, periods of mind wandering were associated with increased power in the alpha band of the electroencephalogram (EEG), as well as a reduction in the magnitude of the P3a component of the event related potential (ERP) in response to the auditory probe. Results support that mind wandering has an impact on driving performance and the associated change in driver's attentional state is detectable in underlying brain physiology. Further, results suggest that detecting the internal cognitive state of humans is possible in a continuous task such as automobile driving. Identifying periods of likely mind wandering could serve as a useful research tool for assessment of driver attention, and could potentially lead to future in-vehicle safety countermeasures.

Keywords: mind wandering, inattention, driving, EEG, alpha

INTRODUCTION

Driver inattention is a frequent cause of automobile crashes and fatalities. This issue has received considerable attention from the scientific community in recent years. Methods of detecting episodes of driver inattention in real-time hold promise for alleviating the human and economic costs of this safety critical issue. Drivers can be inattentive for a variety of reasons, the most obvious being distraction from mobile devices (Caird et al., 2008) or other external factors. However, many distracted driving crashes occur in the absence of an obvious visual or manual distraction. Mind wandering has been suggested as a potential source of many of these distracted driving crashes.

For most people, driving is a highly-overlearned task. Consequently, many of the tasks of everyday driving—lane and speed maintenance, stopping at signaled intersections, etc.—tend to occur relatively automatically. In addition, many trips are routinized with drivers taking the same routes back and forth to work, the grocery store, or other frequently visited locations, which

further promotes automaticity, allowing attention to be devoted to other activities. The routine nature of the driving task, particularly along familiar or monotonous routes, creates an environment ripe for internal distraction or mind wandering, as we will refer to it here. Nevertheless, to maintain safety, drivers must remain attentive to a wide variety of stimuli that may represent latent hazards (Fisher et al., 2002) and be able to swiftly and accurately respond to unexpected events.

There are many varieties of attention (Parasuraman, 1998), but a well-accepted theoretical distinction is between external and internal attention (Chun et al., 2011). External (or "bottom-up") attention is triggered reflexively by environmental events. Internal (or "top-down") attention is voluntary or involuntary application of cognitive resources away from the external environment towards internal thoughts. Most empirical research on internal attention has investigated the pursuit of goals relevant to events in the environment—e.g., such as searching for a landmark while driving in an unfamiliar neighborhood. But, by definition, internal attention can also be devoted to thoughts and memories quite unrelated to any external event. Much less research has been devoted to this topic (Forster and Lavie, 2014), but in recent years a small but growing body of work has examined an aspect of such internal attention—mind wandering (Giambra, 1995; Smallwood and Schooler, 2006).

It is well documented that driving performance is modulated by factors such as effort, fatigue and time on task (Robertson et al., 1997; Grier et al., 2003; Helton and Warm, 2008; May and Baldwin, 2009; Langner et al., 2010; Baldwin et al., 2014). Furthermore, several studies have shown that external distractions, such as talking or texting on a mobile phone, impair driving performance (Strayer et al., 2003; Strayer and Drews, 2007; Caird et al., 2008). Less appreciated and understood is the threat to safety posed by mind wandering behind the wheel, though mind wandering has been associated with an increased risk of being responsible for an automobile crash (Galéra et al., 2012). Fatigue associated with increased time on task may exacerbate both the prevalence and potential risk associated with mind wandering as it is associated with withdrawal of attention away from the driving task and can be considered a form of internal distraction (Williamson, 2007).

Mind wandering has been defined as, "a shift of attention away from a primary task toward internal information" (Smallwood and Schooler, 2006, p. 946). It often occurs without intention and may even occur without explicit awareness, making it a particular challenge to observe and measure. People may continue to move their eyes across a page of text (or the forward field of view while driving) without overtly attending to the viewed stimuli (Smallwood, 2011). Mind wandering is associated with increased activity in the default mode network (DMN) and decreased activity in the dorsal attention network (DAN). The DAN is integrally involved in controlling eye movements and directing exogenous attention to salient stimuli through top-down goal directed processing (Carretié et al., 2013). The DMN is sensitive to the presence of biologically salient non-task relevant stimuli, but this

sensitivity is generally thought to come at the cost of processing task-relevant stimuli (Smallwood, 2011; Carretié et al., 2013). Smallwood (2011) has referred to this interplay between the DAN and DMN as a decoupling process, meaning that as attention is shifted toward one system it is withdrawn from the other. This decoupling may have important implications for driving by suggesting that the more drivers are engaged in mind wandering (activation of the DMN) the less likely they are to process external perceptual cues (potential hazards), particularly if the perceptual cues involve non-biologically salient stimuli (e.g., artificial sounds such as alarms and brake lights).

Two methods have typically been used to detect and quantify mind wandering: self-caught detection and probe-caught detection. The self-caught method of detection involves participants reporting when they notice their mind wandering. In contrast to self-caught methods of detection, probe techniques allow for sampling participants cognitive states throughout a task under experimenter control, thus allowing for the detection of mind wandering episodes that are uncaught by the participants themselves. Three types of probe techniques have been previously utilized. The first and most commonly used is intermittent presentation of questions such as, "Just now, were you mind wandering?" throughout an otherwise continuous task (e.g., Broadway et al., 2015). Another variation of this technique uses probe tones, prompting participants to indicate via a button press whether they were or were not mind wandering (Smallwood and Schooler, 2006). More recently, Seli et al. (2016) used a combination of these probe techniques to determine whether performance differs when participants were aware or unaware that they were mind wandering. Participants were asked to indicate whether their thoughts were on task or they were mind wandering just before they heard the probe tone. Additionally, if participants indicated that they were mind wandering, they were then asked to indicate if they were aware or unaware that they were mind wandering prior to the probe tone. The third probe-caught method involves participants indicating the content of their thoughts at the time of probe, leaving the experimenter to classify whether their thoughts constituted mind wandering (Smallwood and Schooler, 2006). However, by necessity, both self-caught and probe-caught methods likely disrupt participants primary task performance. This may be especially true when probes require more fine-grained judgments of the degree of mind wandering. An ideal mind wandering detection methodology would not require any response on the part of the participant, but rather would provide an on-line classification using some form of machine learning. However, currently such a technique awaits further exploration of the sensitivity and robustness of various metrics using some type of self-report technique.

Results of several recent studies utilizing measures derived from electroencephalography (EEG) substantiate this claim by observed reductions in perceptual sensitivity during periods of mind wandering (Smallwood et al., 2008; Braboszcz and Delorme, 2011). In a breath counting task, Braboszcz and Delorme (2011) report that in the 10 s prior to participants'

self-detecting a state of mind wandering, there was a significant reduction of alpha band activity combined with a diffuse increase in theta band activity, relative to the 10 s period following the button press, during which participant thought had presumably returned to the breath counting task. Modulation of theta band activity may provide a means of distinguishing mind wandering from other types of internal distractions. Savage et al. (2013) found that when participants were given a riddle to solve while performing a simulated driving task there was a similar increase in theta band as that reported by a study measuring frontal sites during mind wandering (Braboszcz and Delorme, 2011). However, in contrast to mind wandering, being internally distracted by a secondary cognitive task leads to a decrease in theta band activity over occipital electrode locations.

Oscillatory activity in the alpha band of the EEG is suggested to be related to attention processes (Klimesch, 2012), specifically the degree to which attention is allocated internally vs. externally. For example, greater alpha power at parietal and/or occipital electrode locations is associated with the failure to detect (Ergenoglu et al., 2004) or discriminate (Van Dijk et al., 2008; Roberts et al., 2014) visual stimuli, while spatial attention processes similarly modulate alpha power such that alpha is suppressed contralaterally and enhanced ipsilaterally to the attended location (Worden et al., 2000; Thut et al., 2006). In contrast, alpha power has been reported to be elevated for tasks in which attention is directly internally, such as working memory retention (Jensen et al., 2002). As periods of high alpha power are associated with lapses of attention to external stimuli (O'Connell et al., 2009; Zauner et al., 2012; Borghini et al., 2014) alpha power is expected to be greater during periods of mind wandering relative to periods of on task behavior.

In terms of event related potentials (ERPs), modulations of early perceptual and attentional components during mind wandering have been observed. Broadway et al. (2015) report that in a reading task the visual N1 component, thought to index orienting and enhancement of perceptual information (Hopfinger et al., 2004), was attenuated during mind wandering. The authors also report an attenuation of the P1 component, thought to index the inhibition of irrelevant information (Hopfinger et al., 2004), over the right hemisphere when participants reported mind wandering. Attenuation of the visual P1 was reported by Baird et al. (2014) bilaterally. However, it should be noted that such early components are relatively small and short-lasting and thus may be difficult to utilize in terms of on-line detection of mind wandering, especially.

More promising as a means of future on-line detection of mind wandering is the modulation of longer latency components such as the P300, a large, long-lasting ERP component thought to reflect allocation of attentional resources (Nieuwenhuis et al., 2005). Smallwood et al. (2008) found that the P300 evoked by visual stimuli in the sustained attention to response task (SART) was significantly attenuated during periods of self-reported mind wandering relative to periods of self-reported on-task behavior. Similar P300 attenuation during mind wandering was reported by Kam et al. (2014) in

response to evaluation of more complex visual stimuli. The relatively high amplitude of the P300 relative to the background EEG allows reasonable extraction of this waveform from single trials for brain-computer interfaces, and has been an area of significant research during the past decade (Cecotti and Rivet, 2014).

Driver behavioral metrics such as speed variability, lane deviation, standard deviation of lateral position (SDLP) and steering reversal rate (SRR), have been shown to serve as an additional method to detect mind wandering (He et al., 2011, 2014; Bencich et al., 2014; Yanko and Spalek, 2014). However, previous research using these behavioral metrics have found inconsistent results that vary depending on the mind wandering detection method. For example, He et al. (2011) used the self-caught detection method and report that speed variability was decreased during mind wandering relative to on-task states. When using the probe detection method to detect instances of mind wandering, Yanko and Spalek (2014) found that mean speed was greater during mind wandering than on-task states. Conversely, using this same technique, Bencich et al. (2014) found that mean speed and speed variability were reduced during mind wandering compared to a state of alertness.

Further, He et al. (2014) found that under low cognitive load, which is thought to be similar to a mind wandering state, SDLP and SRR were reduced compared to a high cognitive load condition. He et al. (2014) interpreted these findings as suggesting that under high cognitive load, more effort and attention is needed to maintain lateral control, which could reflect an alert state. For example, during an attentive state (relative to mind wandering) participants drove at significantly greater speeds, had greater speed variability, greater SDLP and increased SRR (He et al., 2011; Bencich et al., 2014; Yanko and Spalek, 2014). However, though it is well-understood that mind wandering is associated with decreased attention to a primary task, there is currently no consensus of the effects of mind wandering on driving performance.

The purpose of this research was to investigate the frequency of mind wandering over repeated exposure to the same driving route, as well as to identify the relationship between mind wandering and both driver behavior and electrophysiology.

MATERIALS AND METHODS

Procedure

The current study assessed the relationship between mind wandering and driving across 5 days, with the time of day for participation maintained over days within each participant. While a given participant returned at the same time of day for each of their five sessions, the time of day used between participants varied between the morning and afternoon due to participant scheduling availability. The duration of each experimental session lasted from 2 h to 3 h. Each experimental task (the SART and two simulated drives) took approximately 20 min to complete. All procedures were approved by a University Institutional Review Board of George Mason University (protocol # 727867-5) and participants provided

written informed consent in accordance with the Declaration of Helsinki.

The procedures were consistent across the five experimental days with the exception of the first day, which included additional procedures. On day 1 only, participants signed an informed consent form describing the study, performed the Rosenbaum and Snellen visual acuity tests, and completed the demographics and driving history questionnaire, Simulator Sickness Screening, and six additional questionnaires described in "Questionnaires" Section. Additionally, prior to the first experimental drive on the first day of participation, participants completed a 5-min practice drive to familiarize themselves with maneuvering the driving simulator, including steering and braking. As the majority of the questionnaires were expected to be stable across days of participation, participants only completed the KSS and a short questionnaire on sleep quality on days two through five. The remaining procedures were performed across all 5 days.

On each day of participation, a 1-min resting baseline was acquired, during which participants were instructed to relax, keep their eyes open, and to look at the center monitor of the driving simulator. During the baseline, the three driving simulator displays (forward, left and right side screens) displayed visual noise, composed of screen shots of the driving scenario with all pixels randomized. Use of visual noise was used to prevent large changes in screen luminance between the resting baseline and performance of the drive. Participants were then presented with a definition of mind wandering, and were provided a demonstration as to how to respond to the probe tones. The mind wandering definition used in this study, adapted from (Singer and Antrobus, 1972), and Seli et al. (2016), was as follows: "please note that for the purposes of this experiment, the words mind wandering, daydreaming and zoning-out are all synonymous. These are popular terms for which there is no official definition. Despite the subjective nature of the mind wandering experience, we define it as thinking about something unrelated to the immediate task. For example, when driving on a highway it is not unusual for thoughts unrelated to driving to enter your mind. For example, you may think about what you ate for dinner, plans you have later with friends, or an upcoming test. These thoughts are considered mind wandering or off task for the purposes of this experiment whether they occur spontaneously or intentionally. During the experiment, you will periodically hear probe tones. As soon as you hear the tone, please indicate whether you were thinking about the immediate task (either driving or SART) or were mind wandering, meaning you were thinking about something unrelated to the immediate task".

For both experimental drives, participants were instructed to maintain a speed of 65 miles per hour (MPH), stay in the right lane of the roadway, and keep at least one hand on the steering wheel at all times. Additionally, participants were instructed to indicate on the touchscreen whether their thoughts were on-task or they were mind wandering right before they heard the probe tone by pressing the corresponding buttons. Additionally, if participants indicated that they were mind wandering, a second screen appeared on the touchscreen asking participants to indicate if they were

aware or unaware that they were mind wandering before they heard the probe tone. After the first drive was complete, participants completed the SART and then the second driving task.

While the drive was primarily composed of straight highway segments, it also included four curved roadway segments. Due to the potential for differences in attention demand between the straight and curved roadway segments, which could potentially influence attentional state or driving performance, attentional probes that occurred during or within 10 s following a roadway curve were excluded from further analyses.

Although not discussed further, in addition to EEG, participants were also outfitted with a head-mounted (eyeglass) eye-tracker, chest belt heart rate monitor, and were video recorded with a dash board video camera during the experiment.

Participants

Nine participants (5 men, 4 women) were recruited from George Mason University via an e-mail announcement, and were compensated at a rate of $15 per hour. All participants were at least 18 years of age, had normal or corrected-to-normal vision (verified with Rosenbaum card and Snellen eye chart), and possessed a valid United States driver's license. On average, participants were 24 years of age ($SD = 3.37$, range: 18–29) and had 8 years ($SD = 3.19$, range: 2–12) of driving experience.

None of the participants were taking any medications known to affect the central nervous system, none of the participants had sustained a major head injury such as a concussion, and all were right handed. None of the participants self-reported experiencing frequent motion sickness, and were additionally screened using the Simulator Sickness Screening (Hoffman et al., 2003) prior to participation. Each participant signed an informed consent document after being briefed on the procedures of the study. Procedures of the study were approved by the George Mason University Institutional Review Board.

Questionnaires

On the first day of participation, participants completed eight questionnaires: a demographics and driving history questionnaire, Simulator Sickness Screening (Hoffman et al., 2003), Mindful Attention Awareness Scale (MAAS; Brown and Ryan, 2003), Cognitive Failures Questionnaire (CFQ; Broadbent et al., 1982), Mind Wandering Scale (MWS; Singer and Antrobus, 1970; Giambra, 1980), Attention-Related Driving Errors Scale (ARDES; Barragán et al., 2016), Attention-Related Cognitive Errors Scale (ARCES; Cheyne et al., 2006) and Karolinska Sleepiness Scale (KSS; Åkerstedt and Gillberg, 1990). The MAAS, CFQ, MWS, ARDES and ARCES have been shown to reliably predict individuals with a greater propensity of experiencing lapses in attention (Broadbent et al., 1982; Cheyne et al., 2006; Barragán et al., 2016) or mind wandering (Giambra, 1980; Brown and Ryan, 2003; Burdett et al., 2016).

Driving Task

Participants completed two simulated drives on each of the 5 days of participation. The start and end points of the two drives were reversed; on the second drive, the participants began at the

destination from the first drive, and drove back to the first drive origin point (see **Figure 1**).

With the exception of the direction of travel, the driving scenarios were identical between the two drives. The 20 min drives consisted of leaving a parking lot and entering a straight highway with limited scenery and no ambient traffic, until a destination parking lot was reached. The drive did not require navigation; while the highway included curves, participants did not have to make any turns, with the exception of leaving the starting position and entering the destination position. While at the speed limit of 65 MPH, ambient road noise was presented at 60 dB A-weighted sound pressure level (SPL) from speakers integrated into the driving simulator. If participants were driving 15 MPH above or below the speed limit, a message appeared on the center monitor instructing them to slow down or speed up, respectively. Additionally, as participants approached the end of each drive, an auditory message was presented, which instructed the participants to turn into the destination parking lot.

SART

Between the first and second drives on each day of participation, participants completed the SART (Robertson et al., 1997), presented on a touchscreen display (7″ diagonal, 800 by 480 pixels resolution) mounted inside the cab of the driving simulator. The SART was included in the experiment to roughly simulate cognitively demanding office work, which could potentially influence participant performance or mind wandering frequency on the second drive of each day via a depletion of executive resources that would otherwise maintain attention towards the primary task (Thomson et al., 2015). However, the purpose of including the SART wasn't to examine the effect of the SART *per se*, but rather to ensure that enough mind wandering instances occurred throughout the course of the study for comparison of mind wandering and on task states. Within each trial of the SART, participants were presented with a single digit between 1 and 9 on the center of the display. Participants were instructed to click a response button on a Logitech Wireless Presenter remote to the digits 1, 2, 4, 5, 6, 7, 8 or 9, while withholding any response to the remaining digit 3. Participants held the remote with their right hand for the duration of the task. Participants completed two blocks of the task on each day of participation, with each block containing 450 trials and taking approximately 9 min to complete. Each block contained 50 presentations of each of the digits 1 through 9, in a randomized order. Each stimulus remained on-screen for 250 ms, following which it was removed. Response to each SART stimulus were collected for up to 1-s following stimulus onset on each trial. The inter-stimulus interval was jittered with a continuous uniform distribution between 1050 ms and 1250 ms, rounded to the next presentation frame. Stimuli were presented as white digits on a black background, in addition, black and white tracking patterns were displayed in the four corners of the display to allow forward facing camera of the eye-tracker to track the location of the display. Although performed within the cab of the driving simulator for practical purposes, the participants did not perform a drive or interact with the simulator while performing the SART. During performance of the SART the visual displays of the stimulator displayed images of visual noise generated by randomizing the pixels of screenshots of the driving scenario, as used within each day's pre drive baseline condition. The SART was implemented within the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB.

Attentional Probes

Within both the driving scenarios and SART, participants were probed to self-report their current attentional state. Probes were initiated by an auditory tone, composed of a 440 Hz sine wave, 500 ms in length with 10 ms onset and offset ramps, presented at 70 dB SPL via speakers mounted behind the seat of the simulator. Concurrent with tone presentation, a touchscreen mounted on the dashboard of the vehicle presented a response screen displaying two buttons, labeled "Mindwandering" and "On Task". If participants responded "Mindwandering", a second screen was presented displaying two buttons labeled "Aware of Mind Wandering before probe" and "NOT aware of Mind Wandering until probe". The second response was collected as the effects of mind wandering may be particularly pronounced when participants are both off-task and unaware of their inattentiveness (see Smallwood and Schooler, 2015). The time interval to the next attentional probe was relative to the submission of response to the current probe, with a response to stimulus interval jittered with a continuous uniform distribution between 30 s and 90 s. Attentional probes were presented in both the driving task and SART. In the SART, the attentional probes were always presented between SART trials. Prior to participation, participants were explained the probe procedure and response screen, and allowed to familiarize themselves with the attentional probe tone. Attentional probe presentation and response collected was implemented within the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007) for MATLAB.

Electrophysiology

EEG was recorded using a BrainVision ActiChamp 32-channel active EEG system in conjunction with BrainVision PyCorder (v1.0.8) recording software. All electrode impedances were prepared to below 25 kΩ prior to data collection, the threshold recommended by the EEG system manufacturer for active electrodes. EEG was recorded at a sampling rate of 500 Hz, an online reference of electrode TP9 (left mastoid process), and an online band-pass filter between 0.01 Hz and 100 Hz. Offline, EEG data was processed using MATLAB in conjunction with the EEGLAB toolbox (Delorme and Makeig, 2004). Data was re-referenced to the average of TP9 and TP10 (left and right mastoid process electrodes), low-pass filtered using a filter with 40 Hz cutoff and 10 Hz transition bandwidth, and high-pass filtered using a windowed sinc FIR filter with Blackmann window, 0.1 Hz cutoff, and 0.2 Hz transition bandwidth, both as implemented in the EEGLAB function pop_firws. An additional copy of the data was high-pass filtered at 1 Hz with a 2 Hz transition bandwidth, for independent

components analysis (ICA) decomposition (Debener et al., 2010; Winkler et al., 2015). Following filtering, artifactual electrodes were identified and removed via the EEGLAB pop_rejchan function, where artifactual electrodes are defined as those exceeding ±3 standard deviations probability. Artifactual electrodes were identified using the 0.1 Hz high-pass filtered data (which is used for the final analysis), however the same electrodes were subsequently also removed from the 1 Hz high-pass filtered data (which is used for ICA only). Across the 45 sessions within the experiment, an average of 1.29 electrodes were removed ($SD = 0.82$, min = 0, max = 3).

The copy of the data high-pass filtered at 1 Hz was epoched into 1 s intervals, with any 1-s epoch with activity ±500 μV on any electrode removed from further analysis. The remaining epochs were decomposed via ICA, using the Extended InfoMax algorithm as implemented in EEGLAB, following which the ICA weights were copied from the 1 Hz high-pass filtered to the 0.1 Hz high-pass filtered data. Independent components (IC) reflecting eye movements and eye blinks were identified via manual inspection of IC topography and spectra, and subtracted from the data. This ICA procedure was performed for each day of recorded data separately, as the electrodes are unlikely to be located in precisely the same location for the same participant across days.

Following rejection of artifactual ICs, EEG data was epoched into 10 1-s non-overlapping epochs preceding each auditory probe, with each epoch labeled according to the corresponding probe response. Following epoching, any epoch with activity ±100 μV on any electrode was removed from further analysis. Any electrodes previously removed due to artifact were interpolated via spherical interpolation for the purpose of topographic plotting.

For measurement of EEG spectra, the data from each of the remaining 1-s epochs were linearly detrended and converted from the time domain to the frequency domain via Welch's periodogram method, as implemented in the MATLAB function pwelch. Power values for each frequency bin within each epoch were converted into decibels (dB) power (10 * log10(power)) to more closely approximate a normal distribution. The spectral power preceding each auditory probe was then computed by averaging the dB scaled power from the remaining 1-s epochs preceding that probe. Of interest was activity in the theta and alpha bands of the EEG. *A priori*, electrode Fz was selected for investigation of theta power, while electrode Pz was selected for investigation of alpha power. Electrode Fz was selected for theta activity as it is the electrode where frontal midline theta is most prominent. Electrode Pz was selected for alpha activity as enhanced alpha activity at electrode Pz has been reported when attention is directed internally, such as to the content work working memory, and has previously been reported to be sensitive to lapses of attention to external stimuli (O'Connell et al., 2009). The frequency bins representing the peak theta and alpha frequency across subjects were identified via visual inspection of the spectrum collapsed across all participants and conditions under study at the stated electrodes. The peak

theta frequency was identified as the bin centered on 5.85 Hz, while the peak alpha frequency was identified as the bin centered on 8.78 Hz. The dB power within these frequency bins were used for statistical comparison across experimental conditions.

For measurement of time domain EEG activity (ERP), EEG was epoched from −200 ms to 800 ms relative to the onset of the attentional probe, baseline corrected to the mean of the activity within the −200 ms to 0 ms baseline period. Following baseline correction rejection, any epoch with activity ±100 μV on any electrode was removed from further analysis. Any electrodes previously removed due to artifact were interpolated via spherical interpolation for the purpose of topographic plotting. Of interest was activity in the auditory N1 and P3a. *A priori*, electrode Cz was selected for analysis of the auditory N1, a plot of N1 topography collapsed across conditions confirmed a Cz maximal negativity during the time period of the auditory N1. The P3a component classically displays a fronto-central maximum. A plot of the P3a time window collapsed across conditions under study suggested a maximal positivity between electrodes Fz and Cz (near the location of electrode FCz, although this electrode not present in the 32-channel montage used within the present study). For this reason, both electrodes Fz and Cz were selected for analysis of the P3a. Time windows of the auditory N1, and P3a components were identified inspection of the ERP collapsed across all participants and conditions under study at the stated electrodes. The time window of the auditory N1 was identified as 110–160 ms following auditory probe onset. The time window representing the P3a component of the ERP was identified as 200–400 ms following auditory probe onset.

Synchronization

Data from the equipment used within the study were synchronized via the lab streaming layer (LSL) software library[1]. LSL synchronizes the timestamps between multiple devices and computers via a network connection. Specifically, the LSL library was integrated into the driving simulator in order to synchronize the simulator with the computer responsible for the auditory probes, which utilized the LSL MATLAB library. Parallel port trigger events were additionally sent from the computer responsible for the auditory probes directly to the EEG system.

RESULTS

Data reduction was performed using MATLAB, with statistical analyses performed using R (R Core Team, 2015). Specifically, for measures of driver behavior and electrophysiology obtained during each drive, participant attentional state was labeled according to participant response following each attentional probe presentation. Further, the response to each probe was used to label the period of data that had occurred within the 10 s prior to the onset of that probe (see **Figure 2**).

---

[1] https://github.com/sccn/labstreaminglayer
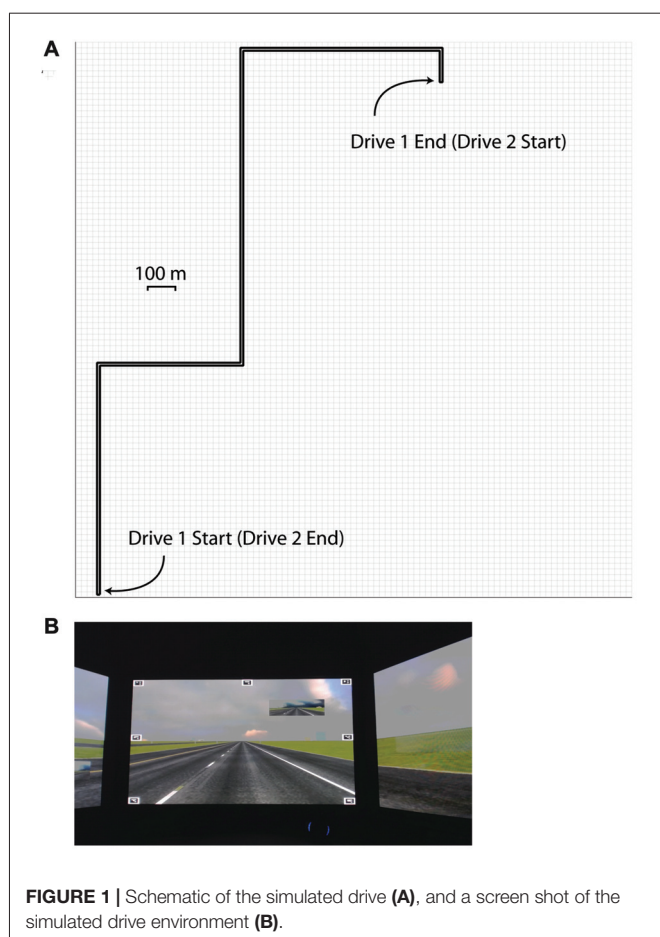
FIGURE 1 | Schematic of the simulated drive **(A)**, and a screen shot of the simulated drive environment **(B)**.
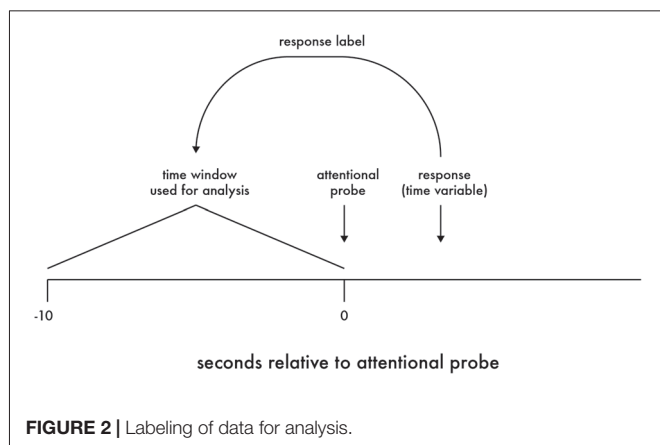
FIGURE 2 | Labeling of data for analysis.

For data from the driving simulator, measures of mean activity (e.g., speed) and variance (e.g., speed variability) were derived within each 10-s period. For data from the EEG, due to the potential of EEG artifact, each 10-s period was first split into 10, 1-s epochs, with any 1-s period containing artifact removed from further analysis and only the remaining 1-s periods averaged as representative of that 10-s period. As previously noted, attentional probes that occurred within or up to 10-s following a roadway curve were excluded from analysis.

Finally, all data were winsorized to ±3 standard deviations from the mean of the corresponding outcome measure (Dixon and Tukey, 1968).

Driver behavior and physiology were analyzed using linear mixed effects modeling. The linear mixed effects analyses were performed using the Linear Mixed-Effects Models using "Eigen" and S4 (lme4) package (Bates et al., 2015) and the Tests in Linear Mixed Effects Models (lmerTest) package (Kuznetsova et al., 2016) for R. As SRR is a count variable (i.e., number of reversals per second), a poisson generalized linear mixed model was analyzed using lme4 and lmerTest. Because participants varied in the number of mind wandering instances reported, Satterthwaite type III approximations were used to calculate the denominator degrees of freedom. The fixed effects for all models included state (mind wandering, on task) and drive (drive one, drive two) as categorical variables, entered as sum contrasts (-1, 1) and day as a continuous variable, mean centered across the data set. Model intercepts were set as a random effect, allowing participants to have varying intercepts.

Questionnaire Analyses

The descriptive statistics for the questionnaire data obtained on the first day of participation are displayed in **Table 1**. However, due to the limited sample size, further analyses were not performed.

SART Performance

The results of the SART are not reported here in depth, however participants attempted to perform the task, correctly responding to over 99% of the non "3" stimuli on average, and correctly withholding response to around 60% of the "3" stimuli, on average.

Prevalence of Mind Wandering

Responses to the attentional probes were analyzed to identify any changes in the frequency of mind wandering, or the frequency of aware relative to unaware mind wandering, over day or drive. The frequency of mind wandering, as well as the frequency of mind wandering awareness, were tested across days and drives via generalized linear mixed effects models with logit link function, using the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2016) packages for R. For mind wandering frequency, the effect of drive reached significance, such that participants were more likely to respond "mind wandering" during the second drive ($z = 2.36$, $\beta = 0.16$,

TABLE 1 | Descriptive Statistics (*M*, *SD*) for Questionnaires.

Questionnaire	*M*	*SD*	Range
ARDES	1.70	0.30	1.33–2.22
ARCES	31.44	3.13	29–38
CFQ	39.33	13.98	27–68
MAAS	3.17	0.77	2.07–4.47
MWS	2.91	0.44	2.23–3.54

Note: ARDES, Attention-Related Driving Errors Scale; ARCES, Attention-Related Cognitive Errors Scale; CFQ, Cognitive Failures Questionnaire; MAAS, Mindful Attention Awareness Scale; MWS, Mind Wandering Scale.
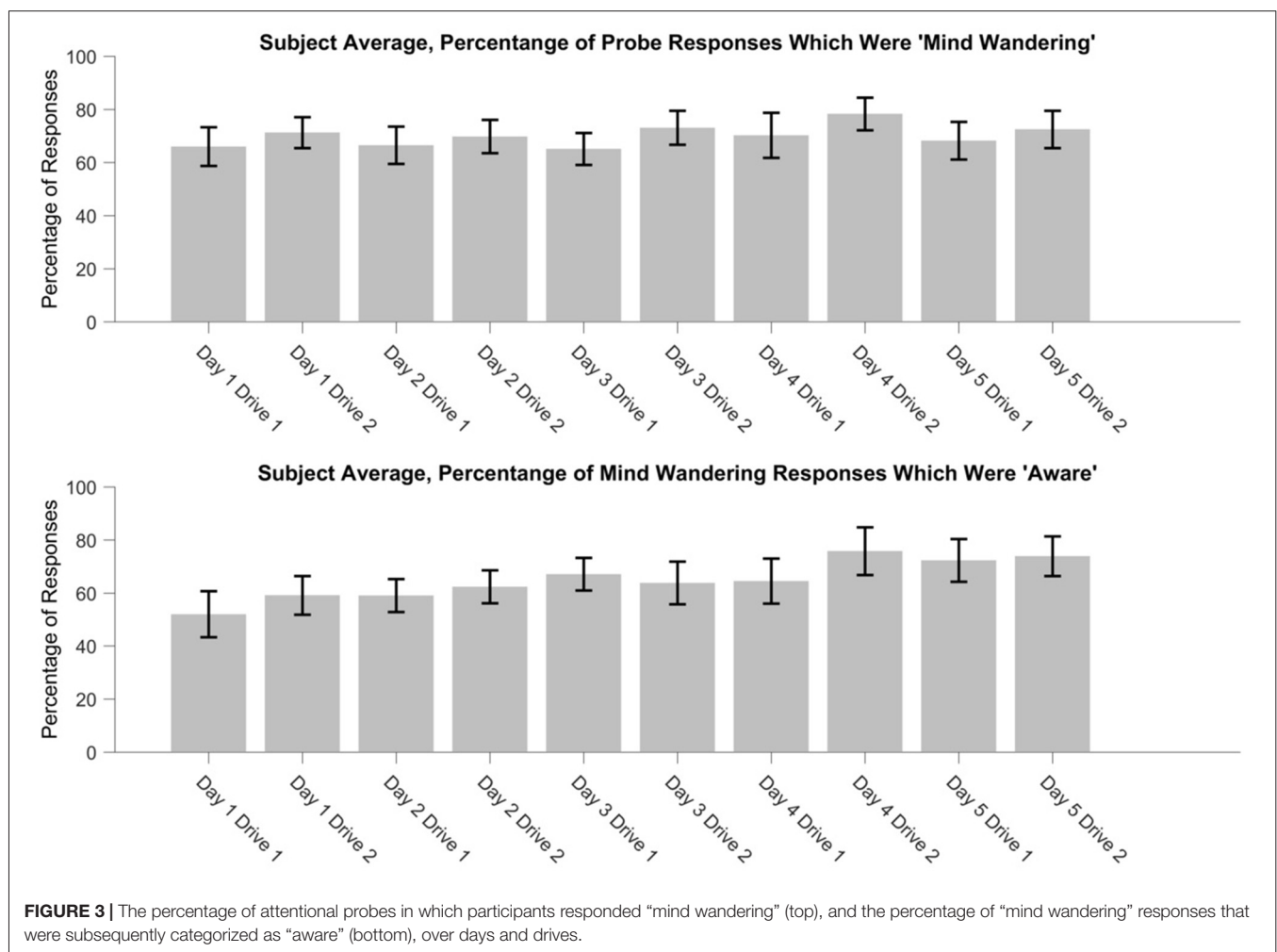
$p = 0.019$) relative to the first drive. Neither the effect of day, nor the day by drive interaction reached significance for mind wandering frequency. For mind wandering awareness frequency, the effect of day reached significance, such that participants were more likely to be aware of their mind wandering as days of the experiment progressed, $z = 4.22$, $\beta = 0.23$, $p < 0.001$. Neither the effect of drive, nor the drive by day interaction reached significance for mind wandering awareness frequency.

For illustrative purposes, the percentage of mind wandering episodes were identified by computing the percentage of probes that were responded to as "mind wandering" during each drive ("mind wandering")/("mind wandering" + "on task"). Mind wandering awareness percentage was identified by computing the percentage of "mind wandering" responses that were subsequently responded to as "aware", during each drive ("aware mind wandering")/("aware mind wandering" + "unaware mind wandering"). In general, participants often self-reported that they were mind wandering (see **Figure 3**). Collapsing across day of participation and drive within day, participants responded "mind wandering" for 70.10% of probe responses (*SD* across participants = 17.00%).

Additionally, participants often self-reported they were aware that they were mind wandering at the time of the probe. Collapsing across day of participation and drive within day, participants responded that they were aware of their mind wandering for 65.00% of mind wandering responses (*SD* across participants = 16.54%).

Driving Behavior

Data from the driving simulator, including speed, lane offset and steering wheel rotation were recorded at 30 Hz. Within each 10-s window, the raw data from the simulator was used to derive speed variability, lateral position variability, lane deviation and the SRR. Lateral position variability was defined as the vehicle's lateral position standard deviation (SDLP) in meters. Lane deviation was defined as the root mean square of the distance in meters from the center of the lane. Lastly, SRR was measured as the number of reversals per second (Hz) when steering angle passed through zero with a degree offset greater than or equal to 2 (He et al., 2014). Since SRR is a count variable requiring a Poisson analysis, it was necessary to multiply all values by 10 for this analysis. Additionally, logarithmic transformations [$\log_{10}(x) + 3$] were performed for speed variability, lateral position

FIGURE 3 | The percentage of attentional probes in which participants responded "mind wandering" (top), and the percentage of "mind wandering" responses that were subsequently categorized as "aware" (bottom), over days and drives.

and lane deviation to account for positively skewed data. The constant of 3 was added to the logarithmically transformed variables to shift the output scale back to a positive range, for convenience.

The results of the linear mixed model for speed variability showed that drive (1, 2) was not a significant predictor (drive 1: $M = 0.27$, $SE = 0.017$; drive 2: $M = 0.26$, $SE = 0.018$), $p = 0.84$. There was a significant interaction between attentional state and day such that, speed variability decreased during "on task" periods across days, $t_{(1158)} = -2.61$, $\beta = -0.03$, $p = 0.009$ (see **Figure 4**).

Additionally, lane deviation was significantly greater for drive 2 ($M = 0.36$, $SE = 0.015$) compared to drive 1 ($M = 0.34$, $SE = 0.013$), $t_{(1157)} = -2.05$, $\beta = -0.02$, $p = 0.041$. There was also a significant effect for state such that, lane deviation was greater during "on task" compared to "mind wandering" attentional state, $t_{(1164)} = 2.58$, $\beta = 0.02$, $p = 0.01$ (see **Figure 5**). However, day did not significantly predict lane deviation, and there were no significant interactions, $ps > 0.05$.
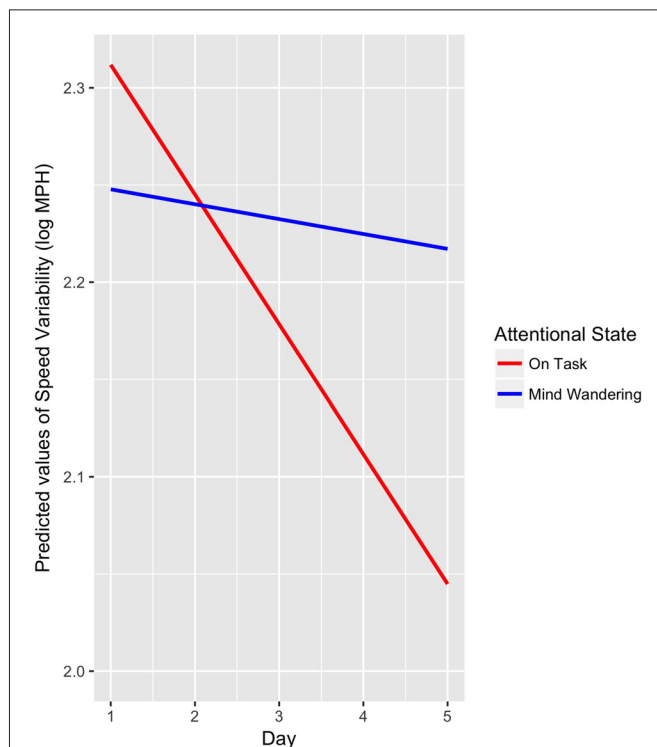
SDLP was also significantly greater when "on task" compared to "mind wandering", $t_{(1165)} = 2.07$, $\beta = -0.02$, $p = 0.04$ (see **Figure 6**). There was also a significant interaction between drive and day such that, SDLP significantly increased for the second drive across days, $t_{(1157)} = 2.36$, $\beta = 0.02$, $p = 0.02$.

A poisson mixed model showed that SRR was also significantly greater when "on task" compared to "mind wandering", $z = 6.77$, $\beta = 0.16$, $p < 0.001$ (see **Figure 7**). However,

day and drive did not significantly predict SRR, and there were no significant interactions, $ps > 0.05$.

EEG Spectra

Power in the theta frequency at frontal electrode Fz and power in the alpha frequency at parietal electrode Pz were selected *a priori* for analysis. No effects reached significance for frontal theta power. Alpha power was increased during "mind wandering" periods relative to "on task" periods, $t_{(1146.4)} = 4.41$, $\beta = 0.43$, $p < 0.001$. Additionally, alpha

FIGURE 5 | Means and standard errors of lane deviation for "mind wandering" and "on task" attentional states.
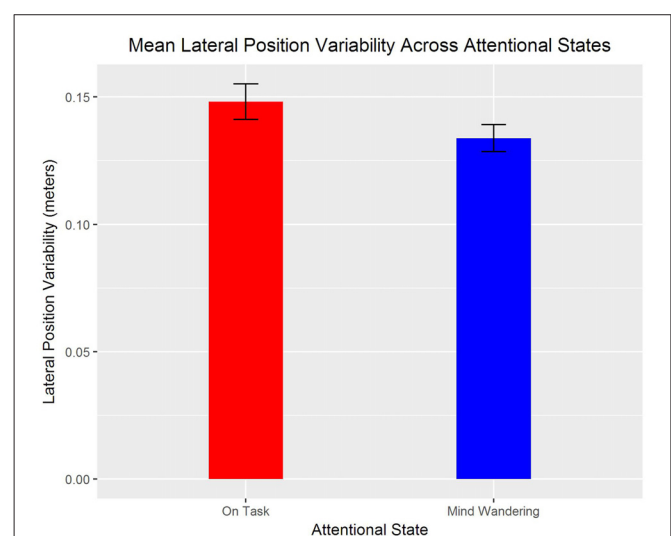
FIGURE 6 | Means and standard errors of the standard deviation of lateral position (SDLP) for "mind wandering" and "on task" attentional states.

FIGURE 4 | Linear mixed model interaction between attentional state ("mind wandering", "on task") and days (1–5) as predictors of speed variability, $p = 0.009$.

FIGURE 7 | Means and standard errors of steering reversal rate (SRR; per second) for "mind wandering" and "on task" attentional states.

FIGURE 8 | Electroencephalography (EEG) power spectrum at parietal electrode site Pz, preceding "mind wandering" vs. "on task" responses. The shaded gray region represents the width of the frequency bin defined as alpha, identified via inspection of the grand mean of both "mind wandering" and "on task" conditions.

power increased over days of participation, $t_{(1144)} = 3.36$, $\beta = 0.22$, $p < 0.001$. The main effect of alpha power at Pz is illustrated in **Figure 8**, with the topography illustrated in **Figure 9**.

ERP to Probe Tone

ERPs to the onset of the auditory probe were analyzed in order to determine whether the auditory probes were processed differently with respect to subjective attentional state. The auditory N1 component was analyzed at electrode Cz,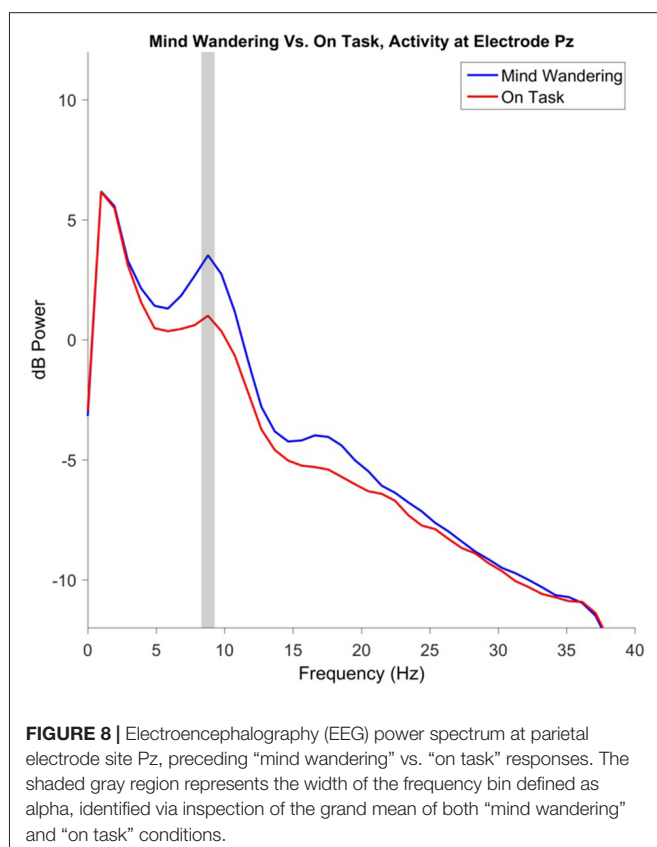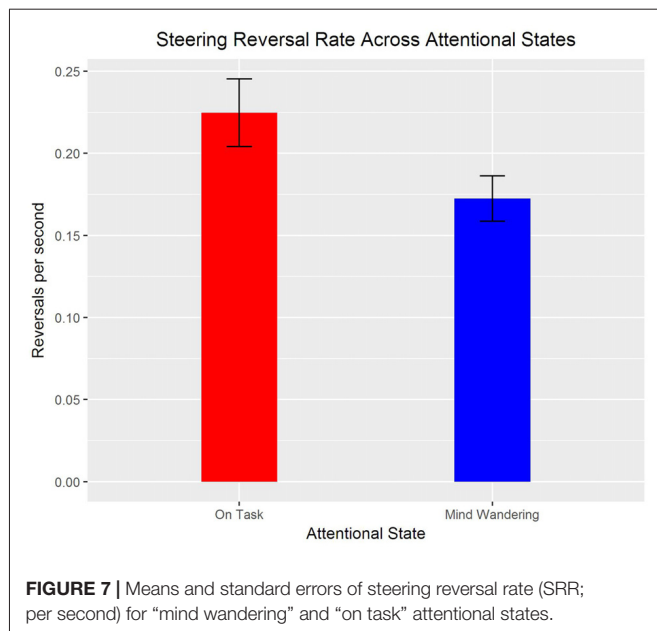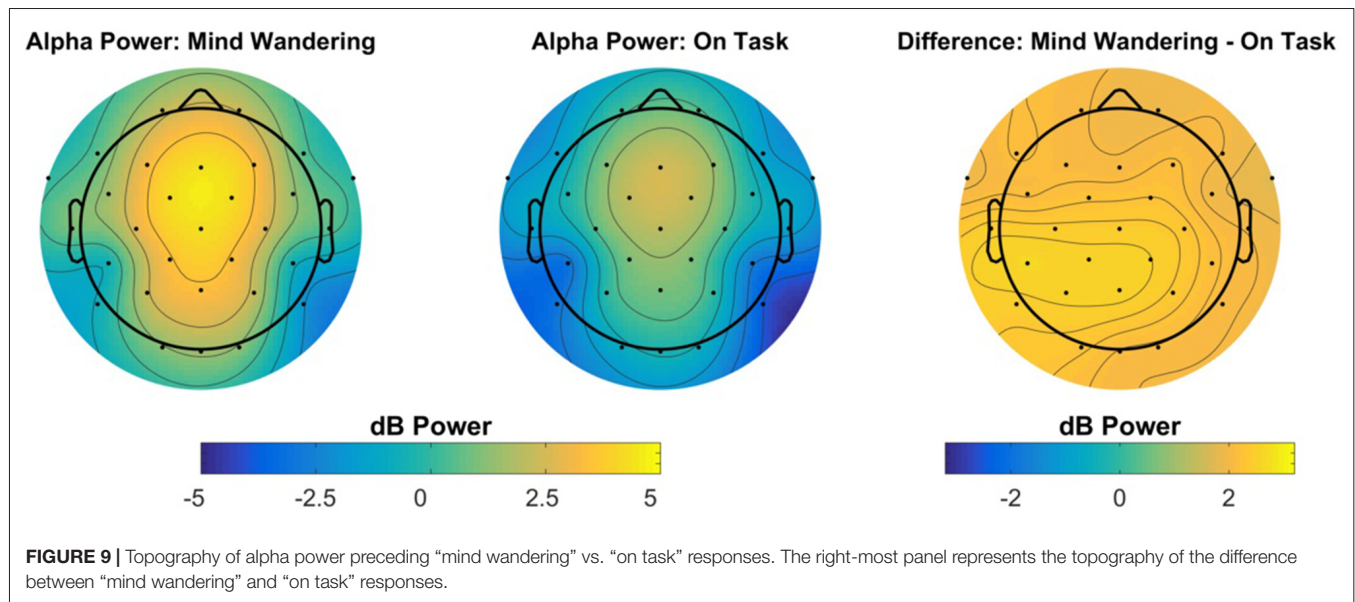 however no main effects or interactions reached significance. The P3a component was analyzed independently at electrodes Fz and Cz. For both electrode locations, the P3a component in response to the auditory probe was reduced in magnitude for probes which were subsequently responded to as "mind wandering" relative to probes which were subsequently responded to as "on task", electrode Fz: $t_{(1115)} = -3.64$, $\beta = -1.45$, $p < 0.001$, electrode Cz: $t_{(1115)} = -2.84$, $\beta = -1.09$, $p = 0.005$. In addition, for both electrode locations, there was a significant effect of drive, such that the P3a during the second drive was diminished relative to the first, electrode Fz: $t_{(1108)} = -3.27$, $\beta = -1.24$, $p = 0.001$, electrode Cz: $t_{(1108)} = -2.48$, $\beta = -0.91$, $p = 0.013$. No effects of day of participation, nor any interactions, reached significance for the P3a at either electrode location. The main effects of attentional state and drive on P3a magnitude are illustrated in **Figure 10**. The topography of the effect of attentional state is illustrated in **Figure 11**.

DISCUSSION

The experimental design was intended to roughly simulate drives to and from work, separated by a cognitively depleting work task. In this study, participants drove the same highway route twice a day for 5 days. Between the two drives on a given day, participants completed a task requiring sustained attention, the SART. Participants self-reported their current attentional state, indicating whether they were either "on-task" or "mind wandering", in response to periodic probe tones. Additionally, following mind wandering responses, participants indicated whether they were aware or not aware of their mind wandering prior to the attentional probes.

The driving scenarios were designed to be rather repetitive and monotonous in order to increase the incidence of mind wandering (Berthié et al., 2015; Thomson et al., 2015) and ensure that enough instances of mind wandering would be generated for study. On average, 70.10% of the probes in the present study were responded to as "mind wandering" by the study participants. In contrast, Killingsworth and Gilbert (2010) report the results of a study probing individuals as to the content of their thoughts throughout everyday life outside the laboratory, reporting that participants respond that they are thinking about something other than what they are currently doing approximately 47% of the time. The high frequency of mind wandering in the present experiment would likely be lessened if the driving scenarios were made to be more demanding. For example, Lin et al. (2016) report a study measuring EEG activity in a motion-base driving simulator, asking participants to detect lane departures using either visual and motion information (the simulator could move in response to road conditions such as rumble strips), or only visual information

FIGURE 9 | Topography of alpha power preceding "mind wandering" vs. "on task" responses. The right-most panel represents the topography of the difference between "mind wandering" and "on task" responses.

(the simulator's motion capabilities were deactivated). The authors report less DMN activity when participants were performing the more demanding version of the task without motion information. The addition of ambient traffic, or a more complex navigation task that also required navigation would also likely decrease the prevalence of self-reported mind wandering.
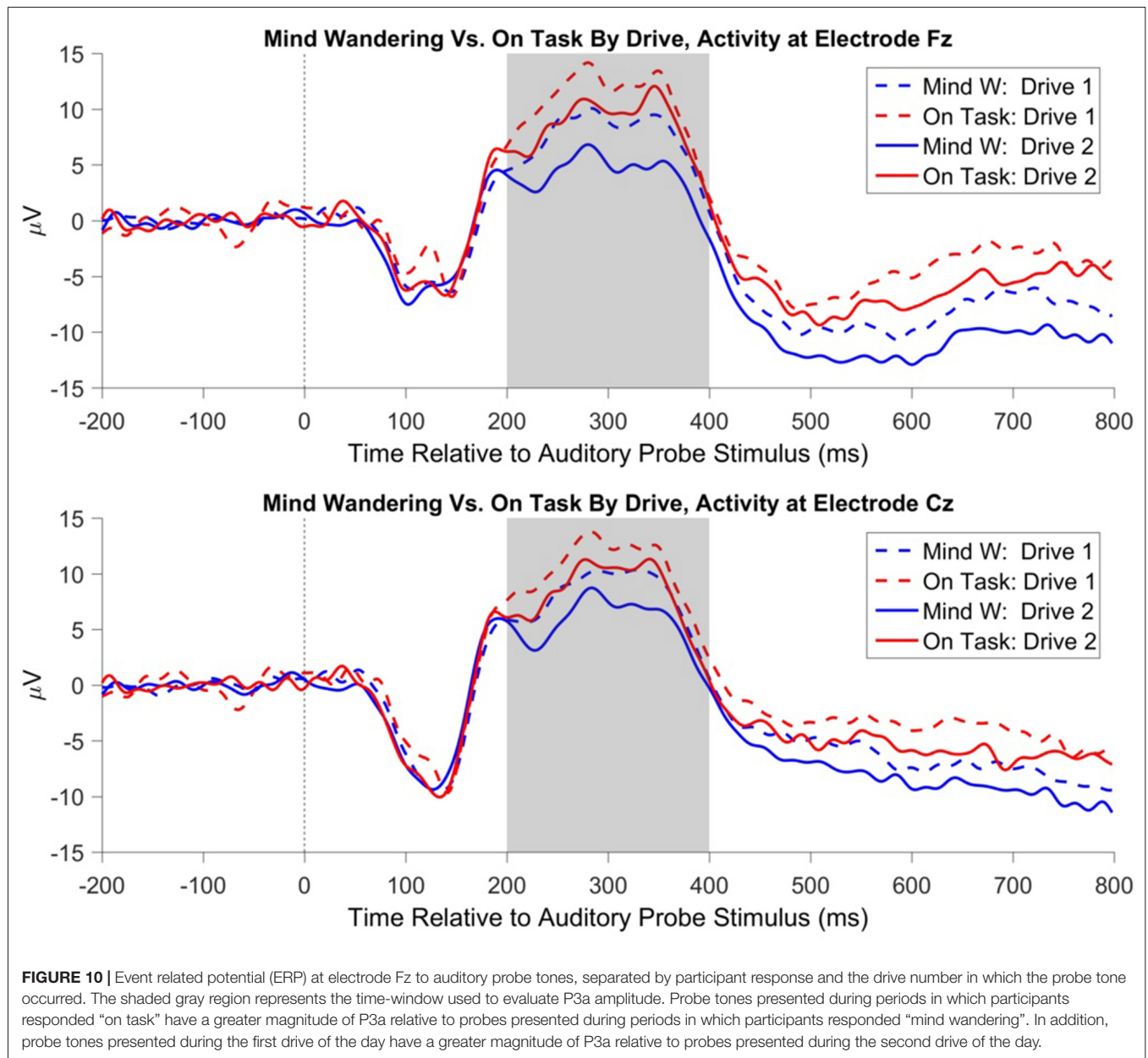
With consideration of previous work that has suggested increases in driver inattention as route familiarity increases (Yanko and Spalek, 2014), it was expected that mind wandering frequency might increase over the 5 days of participation. Instead, mind wandering frequency did not significantly differ over days of participation, though mind wandering did increase for the second drive relative to the first drive within the same day of participation. It is possible that the repetitive nature of the drives induced a maximum amount (or ceiling effect) of mind wandering while still being able to safely operate the simulated vehicle, thereby reducing the potential to see increased mind wandering frequency across days. The SART was performed between drives to roughly simulate performing a work task between commutes, and to ensure that enough mind wandering instances were available for study. As all participants performed the SART, we cannot comment on whether the increase in mind wandering frequency in the second drive of the present study is due to increased route familiarity, resource depletion from the SART, or a time on task effect.

Participants became significantly more aware that they were mind wandering across days. Specifically, for probe responses that were labeled mind wandering, there was a significant increase over days of participation in the frequency that participants reported that they were "aware" of their mind wandering at the time of the probe. It should be noted that the work of Yanko and Spalek (2014) did not query participant subjective state, and instead measured participant inattention by

assessing their ability to detect hazards in the environment. The present experiment, in contrast, did ask participants to report their subjective attentional state, but did not include roadway hazards. Thus the present results may not contradict Yanko and Spalek (2013) but instead suggest a more complex relationship between driving performance and participant awareness of their attentional state. In the present experiment, when examining the 10-s periods preceding the attentional state probes, EEG alpha power increased over days of participation, in addition to more generally being elevated during periods of mind wandering. Taken together, it is possible that participants did become somewhat less attentive to the driving environment over days of participation in the present study, despite the lack of increase in subjective mind wandering reports over days of participation.

Power in the alpha band of the EEG was of greater magnitude during periods of self-reported mind wandering, relative to periods of self-reported on-task performance. Greater alpha power during periods of task inattention has also been reported for a variety of non-driving tasks. For example, O'Connell et al. (2009) report that within the continuous temporal expectancy task (CTET), the magnitude of alpha power over parietal electrodes is elevated prior to trials in which participants missed the target of interest, relative to trials in which participants correctly detected the target. Importantly, the increase in alpha band activity for miss trials was detectable up to 20 s prior to trial onset, suggesting that a slow modulation of top-down control contributes to lapses of attention within the CTET. Alpha power has further been related to inattention within a driving context more specifically. Within a driving simulator task in which participants were provided auditory notifications of lane departures, notifications that were behaviorally successful were associated with decreased alpha power following the notification, while for ineffective notifications alpha power remained elevated (Lin et al., 2013).

FIGURE 10 | Event related potential (ERP) at electrode Fz to auditory probe tones, separated by participant response and the drive number in which the probe tone occurred. The shaded gray region represents the time-window used to evaluate P3a amplitude. Probe tones presented during periods in which participants responded "on task" have a greater magnitude of P3a relative to probes presented during periods in which participants responded "mind wandering". In addition, probe tones presented during the first drive of the day have a greater magnitude of P3a relative to probes presented during the second drive of the day.
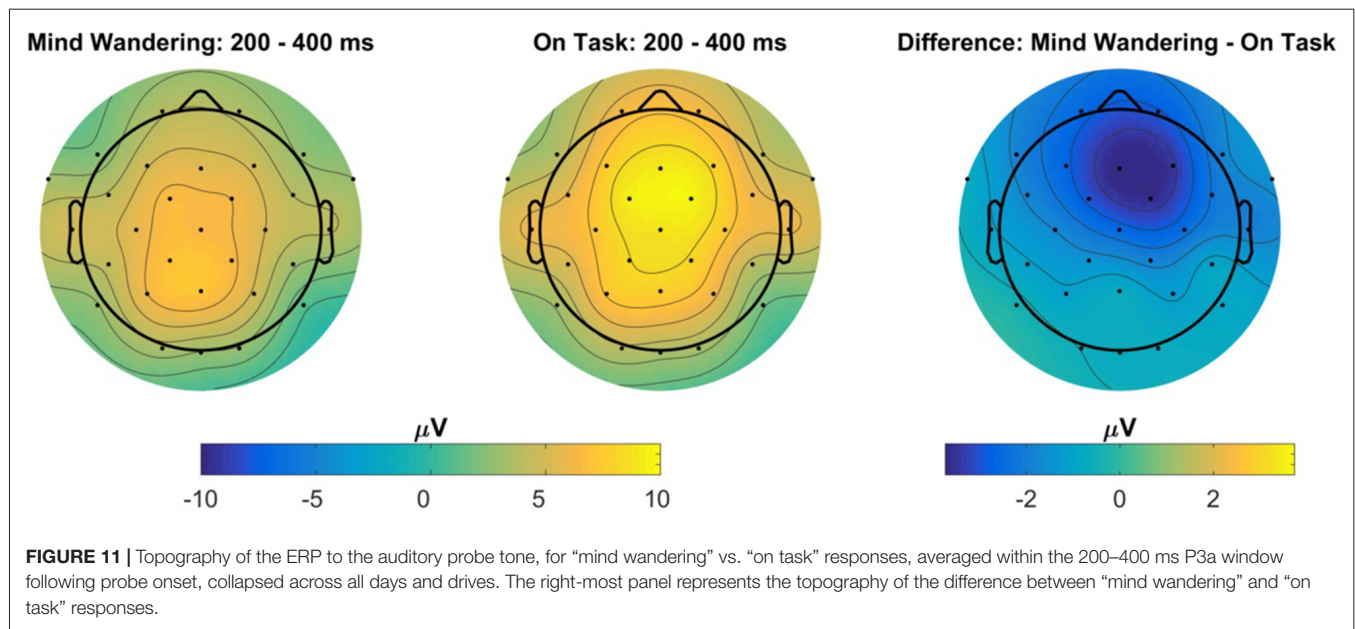
Although consistent with other work on lapses of attention, an increase in alpha power during mind wandering is in contrast to what has been reported for self-detected mind wandering in a breath-counting task, wherein alpha power is suppressed immediately prior to self-detected mind wandering (Braboszcz and Delorme, 2011). This apparent difference may be resolved by considering that the alpha oscillation may serve a different role within the primary task used within the present study, simulated driving, in comparison to the primary task used within Braboszcz and Delorme (2011), eyes closed breath-counting. Future exploration of the possibility of using alpha power as a potential predictor of the mind wandering state, and thus a greater probability of missing a potentially hazardous event warrant further research. A

supplementary method of examining a possible predictive potential for participants to miss critical events can be found in analysis of the ERPs.

Participants were periodically presented with an auditory tone, notifying them to indicate their current attentional state. Analysis of the ERP to this probe tone suggested that the P3a component was larger in response to tones that were subsequently responded to as "on task", relative to tones that were subsequently responded to as "mind wandering". As the P3a component is thought to reflect the orienting of attention towards a novel stimulus (Polich, 2007, 2011), this result supports the decoupling hypothesis of mind wandering (Smallwood, 2011) and is suggestive that participant attention towards the external environment was diminished during

FIGURE 11 | Topography of the ERP to the auditory probe tone, for "mind wandering" vs. "on task" responses, averaged within the 200–400 ms P3a window following probe onset, collapsed across all days and drives. The right-most panel represents the topography of the difference between "mind wandering" and "on task" responses.

periods that were subjectively labeled as "mind wandering" relative to "on task". Despite previous reports that mind wandering modulates the amplitude of early sensory components of the ERP (Baird et al., 2014; Broadway et al., 2015), we did not observe any statistically significant modulation of the auditory N1 component to the probe tones by attentional state. As early sensory components are known to be modulated by attention (Luck et al., 2000), the lack of an effect on the auditory N1 may reflect a lack of top-down attention towards the auditory stimulus in either the mind wandering or on task state. The auditory tone in the present experiment was supra-threshold, unpredictable with respect to onset, and did not require discrimination, only simple detection in order to provoke an attentional probe response. This is in contrast to influences of mind wandering on early sensory components in previous reports, for which the ERP is time-locked to a primary task stimulus (Baird et al., 2014; Broadway et al., 2015).

There are several limitations that should be noted. Our participants were restricted to young (18–29 years) individuals free from disease and other visual or health impairments that might compromise driving. It is not known whether mind wandering might manifest similarly in an older population or in individuals with certain health disorders. Further, none of our participants were shift workers and therefore it is unknown whether disruptions in circadian rhythms might impact performance or the underlying physiology. Additionally, the results were obtained in a driving simulator under carefully controlled conditions. More variability in both environment and behavior could be expected under naturalistic driving conditions.

In summary, both driving behavior and EEG activity demonstrated sensitivity off-line to distinguishing between periods of self-reported mind wandering vs. being on task.

These results are largely in line with previous studies on mind wandering during driving, and on attentional processes as assessed with EEG, and support that mind wandering has an impact on both driving performance and the driver's underlying physiology. Future work could extend these results by examining more closely a driver's reaction to potential hazard situations when mind wandering vs. alert. Drivers may be expected to be less likely to react appropriately to a potential hazard (particularly if it occurs in peripheral vision since during mind wandering since gaze is narrowly focused more centrally). Further, the current results suggest drivers may be less likely to detect an auditory or visual warning while mind wandering. Future work should examine the potential for advanced auditory warnings to aid hazard mitigation in differing attentional states.

AUTHOR CONTRIBUTIONS

CLB, DMR, DB, JDL, NL and JSH contributed to the design of the experiment, DMR and DB collected and analyzed the data. CLB, DMR, DB, NL and JSH contributed to the writing of the manuscript.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Åkerstedt, T., and Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *Int. J. Neurosci.* 52, 29–37. doi: 10.3109/00207459008994241

Baird, B., Smallwood, J., Lutz, A., and Schooler, J. W. (2014). The decoupled mind: mind-wandering disrupts cortical phase-locking to perceptual events. *J. Cogn. Neurosci.* 26, 2596–2607. doi: 10.1162/jocn_a_00656

Baldwin, C. L., May, J. F., and Parasuraman, R. (2014). Auditory forward collision warnings reduce crashes associated with task-induced fatigue in younger and older drivers. *Int. J. Hum. Factors Ergon.* 3, 107–121. doi: 10.1504/ijhfe.2014.067804

Barragán, D., Roberts, D. M., and Baldwin, C. L. (2016). Validation of the attention-related driving errors scale (ARDES) in an English-speaking sample. *Hum. Factors* 58, 1262–1274. doi: 10.1177/0018720816657927

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bencich, E., Gamboz, N., Coluccia, E., and Brandimonte, M. A. (2014). "When the mind "flies": effects of mind-wandering on driving," in *Proceedings of the Trieste Symposium on Perception and Cognition*, eds P. Bernardis, C. Fantoni and W. Gerbino (Trieste: EUT Edizioni Università di Trieste), 91–93. Available online at: http://hdl.handle.net/10077/10522

Berthié, G., Lemercier, C., Paubel, P.-V., Cour, M., Fort, A., Galéra, C., et al. (2015). The restless mind while driving: drivers' thoughts behind the wheel. *Accid. Anal. Prev.* 76, 159–165. doi: 10.1016/j.aap.2015.01.005

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Braboszcz, C., and Delorme, A. (2011). Lost in thoughts: neural markers of low alertness during mind wandering. *Neuroimage* 54, 3040–3047. doi: 10.1016/j.neuroimage.2010.10.008

Brainard, D. H. (1997). The psychophysics toolbox. *Spat. Vis.* 10, 433–436. doi: 10.1163/156856897x00357

Broadbent, D. E., Cooper, P. F., FitzGerald, P., and Parkes, K. R. (1982). The cognitive failures questionnaire (CFQ) and its correlates. *Br. J. Clin. Psychol.* 21, 1–16. doi: 10.1111/j.2044-8260.1982.tb01421.x

Broadway, J. M., Franklin, M. S., and Schooler, J. W. (2015). Early event-related brain potentials and hemispheric asymmetries reveal mind-wandering while reading and predict comprehension. *Biol. Psychol.* 107, 31–43. doi: 10.1016/j.biopsycho.2015.02.009

Brown, K. W., and Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *J. Pers. Soc. Psychol.* 84, 822–848. doi: 10.1037/0022-3514.84.4.822

Burdett, B. R. D., Charlton, S. G., and Starkey, N. J. (2016). Not all minds wander equally: The influence of traits, states and road environment factors on self-reported mind wandering during everyday driving. *Accid. Anal. Prev.* 95, 1–17. doi: 10.1016/j.aap.2016.06.012

Caird, J. K., Willness, C. R., Steel, P., and Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accid. Anal. Prev.* 40, 1282–1293. doi: 10.1016/j.aap.2008.01.009

Carretié, L., Albert, J., López-Martín, S., Hoyos, S., Kessel, D., and Capilla, A. (2013). Differential neural mechanisms underlying exogenous attention to peripheral and central distractors. *Neuropsychologia* 51, 1838–1847. doi: 10.1016/j.neuropsychologia.2013.06.021

Cecotti, H., and Rivet, B. (2014). "Improving single-trial detection of event-related potentials through artificial deformed signals," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (Chicago, IL), 4115–4118.

Cheyne, J. A., Carriere, J. S. A., and Smilek, D. (2006). Absent-mindedness: lapses of conscious awareness and everyday cognitive failures. *Conscious. Cogn.* 15, 578–592. doi: 10.1016/j.concog.2005.11.009

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol.* 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Debener, S., Thorne, J., Schneider, T., and Viola, F. C. (2010). "Using ICA for the analysis of multi-channel EEG data," in *Simultaneous EEG and FMRI: Recording, Analysis, and Application*, eds M. Ullsperger and S. Debener (Oxford: Oxford University Press), 121–133.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dixon, W. J., and Tukey, J. W. (1968). Approximate behavior of the distribution of winsorized t (trimming/winsorization 2). *Technometrics* 10, 83–98. doi: 10.2307/1266226

Ergenoglu, T., Demiralp, T., Bayraktaroglu, Z., Ergen, M., Beydagi, H., and Uresin, Y. (2004). Alpha rhythm of the EEG modulates visual detection performance in humans. *Brain Res. Cogn.* 20, 376–383. doi: 10.1016/j.cogbrainres.2004.03.009

Fisher, D. L., Laurie, N. E., Glaser, R., Connerney, K., Pollatsek, A., Duffy, S. A., et al. (2002). Use of a fixed-base driving simulator to evaluate the effects of experience and PC-based risk awareness training on drivers' decisions. *Hum. Factors* 44, 287–302. doi: 10.1518/0018720024497853

Forster, S., and Lavie, N. (2014). Distracted by your mind? Individual differences in distractibility predict mind wandering. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 251–260. doi: 10.1037/a0034108

Galéra, C., Orriols, L., M'Bailara, K., Laborey, M., Contrand, B., Ribéreau-Gayon, R., et al. (2012). Mind wandering and driving: responsibility case-control study. *BMJ* 345:e8105. doi: 10.1136/bmj.e8105

Giambra, L. M. (1980). Factor analysis of the items of the imaginal processes inventory. *J. Clin. Psychol.* 36, 383–409. doi: 10.1002/jclp.6120360203

Giambra, L. M. (1995). A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Conscious. Cogn.* 4, 1–21. doi: 10.1006/ccog.1995.1001

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., and Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Hum. Factors* 45, 349–359. doi: 10.1518/hfes.45.3.349.27253

He, J., Becic, E., Lee, Y.-C., and McCarley, J. S. (2011). Mind wandering behind the wheel performance and oculomotor correlates. *Hum. Factors* 53, 13–21. doi: 10.1177/0018720810391530

He, J., McCarley, J. S., and Kramer, A. F. (2014). Lane keeping under cognitive load: performance changes and mechanisms. *Hum. Factors* 56, 414–426. doi: 10.1177/0018720813485978

Helton, W. S., and Warm, J. S. (2008). Signal salience and the mindlessness theory of vigilance. *Acta Psychol.* 129, 18–25. doi: 10.1016/j.actpsy.2008.04.002

Hoffman, R. B., Molino, J. A., and Inman, V. W. (2003). "Driving simulator sickness management at turner-fairbank highway research center," in *Driving Simulation Conference* (Dearborn, MI), ISSN 1546–5071. Available online at: https://www.nads-sc.uiowa.edu/dscna/2003/papers/Hoffman_Driving%20Simulator%20Sickness%20Management....pdf

Hopfinger, J. B., Luck, S. J., and Hillyard, S. A. (2004). "Selective attention: electrophysiological and neuromagnetic studies," in *The Cognitive Neurosciences*, 3rd Edn. ed. M. S. Gazzaniga (Cambridge, MA: MIT Press), 561–574.

Jensen, O., Gelfand, J., Kounios, J., and Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cereb. Cortex* 12, 877–882. doi: 10.1093/cercor/12.8.877

Kam, J. W. Y., Xu, J., and Handy, T. C. (2014). I don't feel your pain (as much): the desensitizing effect of mind wandering on the perception of others' discomfort. *Cogn. Affect. Behav. Neurosci.* 14, 286–296. doi: 10.3758/s13415-013-0197-z

Killingsworth, M. A., and Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science* 330, 932–932. doi: 10.1126/science.1192439

Kleiner, M., Brainard, D., and Pelli, D. (2007). What's new in psychtoolbox-3? *Perception* 36, 1–16. ECVP Abstract Supplement.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends Cogn. Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2016). lmerTest: tests in linear mixed effects models. *R Package Version 2.0–32*. Available online at: https://cran.r-project.org/package=lmerTest

Langner, R., Willmes, K., Chatterjee, A., Eickhoff, S. B., and Sturm, W. (2010). Energetic effects of stimulus intensity on prolonged simple reaction-time performance. *Psychol. Res.* 74, 499–512. doi: 10.1007/s00426-010-0275-6

Lin, C.-T., Chuang, C.-H., Kerick, S., Mullen, T., Jung, T.-P., Ko, L.-W., et al. (2016). Mind-wandering tends to occur under low perceptual demands during driving. *Sci. Rep.* 6:21353. doi: 10.1038/srep21353

Lin, C.-T., Huang, K.-C., Chuang, C.-H., Ko, L.-W., and Jung, T.-P. (2013). Can arousing feedback rectify lapses in driving? Prediction from EEG power spectra. *J. Neural Eng.* 10:56024. doi: 10.1088/1741-2560/10/5/056024

Luck, S. J., Woodman, G. F., and Vogel, E. K. (2000). Event-related potential studies of attention. *Trends Cogn. Sci.* 4, 432–440. doi: 10.1016/S1364-6613(00)01545-X

May, J. F., and Baldwin, C. L. (2009). Driver fatigue: the importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transp. Res. Part F Traffic Psychol. Behav.* 12, 218–224. doi: 10.1016/j.trf.2008.11.005

Nieuwenhuis, S., Aston-Jones, G., and Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychol. Bull.* 131, 510–532. doi: 10.1037/0033-2909.131.4.510

O'Connell, R. G., Dockree, P. M., Robertson, I. H., Bellgrove, M. A., Foxe, J. J., and Kelly, S. P. (2009). Uncovering the neural signature of lapsing attention: electrophysiological signals predict errors up to 20 s before they occur. *J. Neurosci.* 29, 8604–8611. doi: 10.1523/JNEUROSCI.5967-08.2009

Parasuraman, R. (1998). "The attentive brain: issues and prospects," in *The Attentive Brain*, ed. R. Parasuraman (Cambridge, MA: MIT Press), 3–15.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* 10, 437–442. doi: 10.1163/156856897x00366

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Polich, J. (2011). "Neuropsychology of P300," in *The Oxford Handbook of Event-Related Potential Components*, 1st Edn. eds E. S. Kappenman and S. J. Luck (New York, NY: Oxford University Press), 159–188.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundations for Statistical Computing. Available online at: https://www.R-project.org/

Roberts, D. M., Fedota, J. R., Buzzell, G. A., Parasuraman, R., and McDonald, C. G. (2014). Prestimulus oscillations in the alpha band of the EEG are modulated by the difficulty of feature discrimination and predict activation of a sensory discrimination process. *J. Cogn. Neurosci.* 26, 1615–1628. doi: 10.1162/jocn_a_00569

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., and Yiend, J. (1997). 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia* 35, 747–758. doi: 10.1016/s0028-3932(97)00015-8

Savage, S. W., Potter, D. D., and Tatler, B. W. (2013). Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study. *Transp. Res. Part F Traffic Psychol. Behav.* 17, 52–62. doi: 10.1016/j.trf.2012.10.002

Seli, P., Risko, E. F., and Smilek, D. (2016). On the necessity of distinguishing between unintentional and intentional mind wandering. *Psychol. Sci.* 27, 685–691. doi: 10.1177/0956797616634068

Singer, J. L., and Antrobus, J. S. (1970). *Imaginal Processes Inventory*. Princeton, NJ: Educational Testing Service.

Singer, J. L., and Antrobus, J. S. (1972). "Daydreaming, imaginal processes, and personality: a normative study," in *The Function and Nature of Imagery*, ed. P. Sheehan (New York, NY: Academic Press), 175–202.

Smallwood, J. (2011). Mind-wandering while reading: attentional decoupling, mindless reading and the cascade model of inattention. *Lang. Linguist. Compass* 5, 63–77. doi: 10.1111/j.1749-818x.2010.00263.x

Smallwood, J., Beach, E., Schooler, J. W., and Handy, T. C. (2008). Going AWOL in the brain: mind wandering reduces cortical analysis of external events. *J. Cogn. Neurosci.* 20, 458–469. doi: 10.1162/jocn.2008.20037

Smallwood, J., and Schooler, J. W. (2006). The restless mind. *Psychol. Bull.* 132, 946–958. doi: 10.1037/0033-2909.132.6.946

Smallwood, J., and Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annu. Rev. Psychol.* 66, 487–518. doi: 10.1146/annurev-psych-010814-015331

Strayer, D. L., and Drews, F. A. (2007). Cell-phone—induced driver distraction. *Curr. Dir. Psychol. Sci.* 16, 128–131. doi: 10.1111/j.1467-8721.2007.00489.x

Strayer, D. L., Drews, F. A., and Johnston, W. A. (2003). Cell phone—induced failures of visual attention during simulated driving. *J. Exp. Psychol. Appl.* 9, 23–32. doi: 10.1037/1076-898x.9.1.23

Thomson, D. R., Besner, D., and Smilek, D. (2015). A resource-control account of sustained attention: evidence from mind-wandering and vigilance paradigms. *Perspect. Psychol. Sci.* 10, 82–92. doi: 10.1177/1745691614556681

Thut, G., Nietzel, A., Brandt, S. A., and Pascual-Leone, A. (2006). α-Band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. *J. Neurosci.* 26, 9494–9502. doi: 10.1523/JNEUROSCI.0875-06.2006

Van Dijk, H., Schoffelen, J.-M., Oostenveld, R., and Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *J. Neurosci.* 28, 1816–1823. doi: 10.1523/jneurosci.1853-07.2008

Williamson, A. (2007). "Fatigue and coping with driver distraction," in *Distracted Driving*, eds I. J. Faulks, M. A. Regan, M. Stevenson, J. Brown, A. Porter and J. D. Irwin (Sydney, NSW: Australasian College of Road Safety), 611–622.

Winkler, I., Debener, S., Müller, K.-R., and Tangermann, M. (2015). "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (Milan: IEEE), 4101–4105. doi: 10.1109/embc.2015.7319296

Worden, M. S., Foxe, J. J., Wang, N., and Simpson, G. V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific α-band electroencephalography increases over occipital cortex. *J. Neurosci.* 20:RC63.

Yanko, M. R., and Spalek, T. M. (2013). Route familiarity breeds inattention: a driving simulator study. *Accid. Anal. Prev.* 57, 80–86. doi: 10.1016/j.aap.2013.04.003

Yanko, M. R., and Spalek, T. M. (2014). Driving with the wandering mind: the effect that mind-wandering has on driving performance. *Hum. Factors* 56, 260–269. doi: 10.1177/0018720813495280

Zauner, A., Fellinger, R., Gross, J., Hanslmayr, S., Shapiro, K., Gruber, W., et al. (2012). Alpha entrainment is responsible for the attentional blink phenomenon. *Neuroimage* 63, 674–686. doi: 10.1016/j.neuroimage.2012.06.075

EEG and Eye Tracking Signatures of Target Encoding during Structured Visual Search

Anne-Marie Brouwer[1]*, Maarten A. Hogervorst[1], Bob Oudejans[1], Anthony J. Ries[2] and Jonathan Touryan[2]

[1]Perceptual and Cognitive Systems, Netherlands Organisation for Applied Scientific Research (TNO), Soesterberg, Netherlands, [2]U.S. Army Research Laboratory, Aberdeen, MD, United States

EEG and eye tracking variables are potential sources of information about the underlying processes of target detection and storage during visual search. Fixation duration, pupil size and event related potentials (ERPs) locked to the onset of fixation or saccade (saccade-related potentials, SRPs) have been reported to differ dependent on whether a target or a non-target is currently fixated. Here we focus on the question of whether these variables also differ between targets that are subsequently reported (hits) and targets that are not (misses). Observers were asked to scan 15 locations that were consecutively highlighted for 1 s in pseudo-random order. Highlighted locations displayed either a target or a non-target stimulus with two, three or four targets per trial. After scanning, participants indicated which locations had displayed a target. To induce memory encoding failures, participants concurrently performed an aurally presented math task (high load condition). In a low load condition, participants ignored the math task. As expected, more targets were missed in the high compared with the low load condition. For both conditions, eye tracking features distinguished better between hits and misses than between targets and non-targets (with larger pupil size and shorter fixations for missed compared with correctly encoded targets). In contrast, SRP features distinguished better between targets and non-targets than between hits and misses (with average SRPs showing larger P300 waveforms for targets than for non-targets). Single trial classification results were consistent with these averages. This work suggests complementary contributions of eye and EEG measures in potential applications to support search and detect tasks. SRPs may be useful to monitor what objects are relevant to an observer, and eye variables may indicate whether the observer should be reminded of them later.

Keywords: EEG, pupil size, fixation, SRP, FRP, visual search, target detection, BCI

INTRODUCTION

Visual search is a common task that is performed when looking for a singing bird in the trees, when checking a poster for graphical errors or when searching an environment for suspicious objects. We are interested in EEG and eye variables as potential sources of information about the underlying processes of target detection and encoding during visual search, i.e., a task where the eyes move. As elaborately discussed in "Application" Section, monitoring target detection and encoding on the basis of implicit variables could be useful in a number of applications.

Brain and Eye Correlates of Target Detection

The literature describes several EEG and eye indicators of target detection. Observers usually fixate longer on targets than non-targets in a search task (e.g., Brouwer et al., 2013; Jangraw et al., 2014; Wenzel et al., 2016). Also, several studies showed stronger pupil dilation responses to target compared to non-target stimuli (Nieuwenhuis et al., 2011; Hong et al., 2014). Finally, the P300 event related potential (ERP), a positive peak occurring in the EEG signal roughly 300 ms after a sensory event, indicates that an observer's attention has been drawn. It has been shown that the P300 reliably distinguishes between top-down defined "targets" and "non-targets", e.g., in cases where observers are asked to pay attention to the letter "p" presented in a sequence of successively flashed letters (e.g., Farwell and Donchin, 1988).

While in studies on pupil dilation and P300 participants usually kept their eyes still, some studies showed similar findings when participants were actively and purposefully moving their eyes. Jangraw et al. (2014) found that, for a realistic visual search scenario, pupil size locked to target fixation onset increased relative to non-target fixation. There is a growing body of research on ERPs that are not time locked to stimulus onset as determined within an experimental paradigm, but to ocular-based events such as saccades and fixations (saccade-related potentials, SRPs or fixation-related potentials, FRPs respectively) as a means to determine whether observers are looking at a target (i.e., a top-down defined relevant object). For some of the early studies (e.g., Hale et al., 2008; Luo et al., 2009), differences found between target and non-target SRPs could have been due to confounding factors such as systematic target vs. non-target differences in saccade length, low level visual features, or motor preparation to press a button. However, by now it is clear that ERPs following fixation of a target are different than ERPs following fixation of a non-target and that these differences are associated with top-down stimulus processing (e.g., Dandekar et al., 2012a,b; Kamienkowski et al., 2012; Brouwer et al., 2013).

Brain and Eye Correlates of Missed Targets

In the current study, we examine fixation duration, pupil size and SRPs in a structured visual search task. We are especially interested in encoding failures, i.e., failing to report a fixated target. Observers can fail to report fixated targets for different reasons. They may have "really" missed the targets, e.g., because the type of target was difficult to identify (perceptual identification failures). Another possible reason is that after target identification, observers forgot the target before it was time to report, for example, because they were involved in another task (memory encoding failures).

In the case that observers did not identify the target, we expect eye and SRP features for misses to be similar to non-targets—observers simply did not perceive the target as being a target. This hypothesis is supported by research by Dias et al. (2013) who studied SRPs following fixations of not-reported targets under circumstances that misses were likely due to not identifying the target. In their task, participants searched a display filled with rectangular objects where the target was defined by a combination of visual features that changed every trial. For instance, a target could be a vertical bar consisting of a red bar on the left and a yellow one on the right that was presented between non-targets that also consisted of vertical colored bar combinations. After finding the target, participants had to immediately report finding it. The average miss SRP could not be distinguished from the average non-target SRP, while the average hit SRP stood out and was consistent with a P300. Dias et al. (2013) found that misses were associated with relatively high EEG alpha activity which has also been linked to lapses of attention (e.g., Vázquez Marrufo et al., 2001).

In the case that targets are identified but not reported due to a memory encoding failure, we do not expect features for misses to be the same as for non-targets. Here, misses are detected as targets at the time of fixation. However, to the extent that differences in target vs. non-target SRPs reflect differences in allocated attention, where this difference matters for storage in memory, SRP waveforms following these missed targets could fall in between target and non-target SRPs. Evidence that is partly in line with this expectation has been found in previous P300 studies. In these studies, participants were asked to remember as many words as possible of a list of sequentially presented words. P300s following presentation of words that were later remembered were compared to those that were later forgotten. Remembered words tended to correspond with larger P300s. However, these effects were small and interacted with primacy and recency effects as well as the type of rehearsal or encoding strategy of the participants (Karis et al., 1984; Fabiani et al., 1990; Azizian and Polich, 2007; Kamp et al., 2012). For instance, Azizian and Polich found larger P300 amplitudes for recalled compared to forgotten words only for words at the beginning of the list, and Fabiani et al. (1990) only found a positive relation between P300 and recall when participants used a rote strategy (repeating the words to themselves) for remembering. As of yet, we do not know whether there is a relation between the P300 and later recall of targets in a visual search task.

Fixation duration may reflect or may enable more attention and deeper processing of the fixated object. Therefore, we not only expect fixation duration to be longer for targets than for non-targets, but also for hits compared to misses.

We discussed that SRPs and fixation duration as signatures for missed targets may be in between hit target and non-target values. The same may hold for pupil size, i.e., hit targets may be associated with larger pupil size than missed targets. However, the reverse may be found as well. In case of a search task where targets are likely to be forgotten because the observer is simultaneously attending to another task, we expect momentary high workload to be associated with misses. Since workload or memory load is strongly related to pupil size (Kahneman and Beatty, 1966; Beatty, 1982; Hogervorst et al., 2014), also specifically with memory load during visual search (Porter et al., 2007), we expect it to be larger for misses than for hits in our task.

Current Study

In the current study we examine the general association between SRP and eye features on the one hand and whether a fixated object is a target and is going to be missed on the other hand. In addition, we examine how well we can distinguish between targets and non-targets, and between hits and misses on a single fixation basis and which combination of these sources of information gives us the best result. A surge of recent studies show that it is possible to distinguish between target and non-target SRPs above chance on a single SRP basis, also in rather challenging circumstances. For instance, Ušćumlić and Blankertz (2016) show a single trial distinction when moving stimuli are involved; single trial classification has been shown for a mixture of foveally and parafoveally identified stimuli (Brouwer et al., 2014; Wenzel et al., 2016); and it has been shown when using more natural stimuli such as looking for a face in a crowd (Kaunitz et al., 2014) or when viewing signs during navigating a virtual environment (Jangraw et al., 2014). Some previous studies have shown that combining SRP features and eye related features increased classification performance for targets and non-targets (Jangraw et al., 2014; Wenzel et al., 2016). Thus, for target and non-target distinction, eye and brain signals can potentially add complimentary information. It remains to be seen how this works out for the distinction between hits and misses.

In our task, participants performed a structured visual search task consisting of scanning 15 locations on a screen. Target locations were reported after scanning all locations. An auditory math task was performed in the high load condition but ignored in the low load condition. In pilot experiments we verified that performing such a double task results in failures to report targets. With respect to the SRP, we expect a larger P300 for targets compared to non-targets, and possibly a higher P300 for hit compared to missed targets. We expect longer fixation duration for targets compared to non-targets, and longer fixation duration for hits than for misses. Pupil size may be larger for targets than for non-targets, and—through a general association between high workload and pupil size—larger for misses than for hits.

MATERIALS AND METHODS

Participants

Twenty-one participants (nine males, 12 females) between the age of 19 and 30 years (average age: 23) were recruited through the participant pool of the Netherlands Organization for Applied Scientific Research (TNO). None of the participants wore glasses. Each participant received a monetary reward for his or her time and travel costs. This study was carried out in accordance with the recommendations of the Human Research Protections Official (HRPO) and the TNO Institutional Review Board (TCPE) with written informed consent from all subjects. All participants signed an informed consent form in accordance with the Declaration of Helsinki. This study was approved by the HRPO and TCPE and conducted in accordance with the Army Research Laboratory's IRB requirements (32 CFR 219 and DoDI 3216.02).

Materials

The task was presented on a 19″ flat-screen monitor (Dell 1907FP Flatpanel 19″, display size 37.5 × 30 cm). The screen resolution was 1280 × 1024 and the refresh rate was set at 60 Hz. Participants' eyes were located approximately 40 cm from the screen. Audio output was coming from a dual speaker set (TEAC PowerMax 60/2) placed left and right of the screen.

Gaze and pupil size were recorded at 60 frames per second using SmartEyePro V6.1.6 (Smart Eye AB, Göteburg, Sweden). This system consists of two cameras (Basler acA640-120 gm, HR 8.0 mm lens) placed at the left and right side of the screen.

EEG and EOG signals were recorded using an ActiveTwoMK II system (BioSemi, Amsterdam, Netherlands) with a sampling frequency of 512 Hz. For EEG, 32 active silver-chloride EEG electrodes were placed according to the 10–20 system and were referenced to the Common Mode Sense (CMS) active electrode and Driven Right Leg (DRL) passive electrode. Four EOG electrodes (BioSemi Flat Active electrodes, Amsterdam, Netherlands) were used to record eye movement. Two EOG electrodes were placed at the approximately 0.5 cm off the lateral canthi of both eyes, and were used to record horizontal eye movement. Another two EOG electrodes were placed above and below the left eye to record vertical eye movement and blinks. The electrode offset of all electrodes was below 25.

Task and Design

The experiment featured two tasks: a monitoring task and an auditory math task. In the high load condition, participants performed both tasks. In the low load condition, they only needed to perform the monitoring task, even though the math task was still played to keep auditory stimulation constant across conditions.

Monitoring Task

Participants were asked to monitor 15 "systems", represented by strings of symbols on a screen and placed in three rows of five columns. There were three different system conditions: hidden ("####"), working as intended ("#OK#") or system failure ("#FA#"). At the start of a trial, all system conditions were hidden. Then, each of the systems was successively highlighted for 1 s (1027 ms) by displaying a square around it while its condition changed from "####" into either "#OK#" or "#FA#" (**Figure 1** shows an example of the stimulus display, and also presents the dimensions of the different stimulus elements). Highlighting the systems happened in random order, except for that two subsequently presented systems were never further apart than two steps in horizontal direction and one in vertical direction, or two vertical and one horizontal. The next highlighted system was in peripheral vision, such that we could not distinguish between "#OK#" or "#FA#" without making a saccade. After all system conditions had been shown, empty boxes appeared at the system locations and the participant had to indicate which systems failed during the trial by clicking the appropriate boxes with the left mouse button. When finished, the participant pressed an OK button at the top left of the screen. Every trial, two, three or four "#FA#"s were presented.

FIGURE 1 | Depiction of the stimulus screen at a moment where the lower left system is highlighted. Dimensions are indicated in cm. The distance between the screen and the eyes was approximately 40 cm, i.e., in degrees of visual angle, the "####" system was about 2.6 × 0.7° of visual angle.
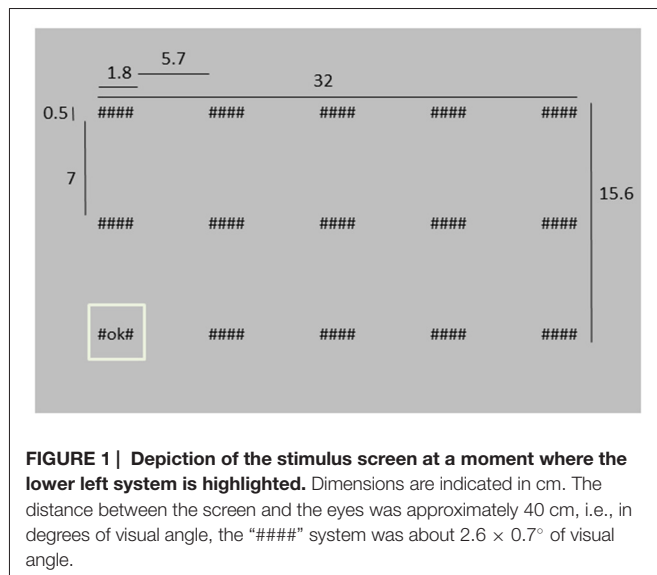
The amount (two, three or four) and the "#FA#" locations were chosen randomly.

Math Task

The math task was an aurally presented sum consisting of six numbers between 6 and 12. Only addition (+) and subtraction (−) operations were used. An example is "8−6 + 10−12 + 11 + 7". The first number was presented 1 s after the start of the monitor task, and every 2660 ms another number was presented. Thus, the last number was presented after 14.3 s. Performing this task involves attention and working memory. When participants had to perform the math task (i.e., in the high load condition), they were required to give the answer of the sum after having indicated where the "#FA#"s were located. This was done by typing the answer and pressing enter. In order to motivate participants to perform the math task, they received feedback on their answer. If the answer was incorrect, the correct answer was shown.

For each of the load conditions, participants performed eight blocks of 11 trials. High and low load conditions were presented alternately, starting with the high load condition.

Procedure

After a general explanation and signing the informed consent, the EEG and EOG electrodes were attached. During this time, the participant had time to read detailed instructions. Participants were asked to take a comfortable position in front of the screen. Even though they were able to move freely, they were instructed to minimize their head and body movements. Before the task began, a four point-calibration was performed to calibrate the SmartEye system. There was a few minutes break after eight blocks of trials, i.e., half-way through the experiment. The participants indicated when they were ready to continue.

Analysis

Electrophysiological Data

EEG and EOG data was resampled to 256 Hz. Bad EEG channels were identified as channels with standard deviations exceeding five times the median standard deviation over all channels, after bandpass filtering between 0.5 and 32 Hz. This affected 1–3 channels for five participants. The unfiltered data from bad channels were replaced by the weighted average of unfiltered data from the surrounding channels. Next, the EEG-data was re-referenced to the mean of all unfiltered data excluding the bad channels. The resulting signals were submitted to bandpass filtering between 0.5 and 32 Hz.

We extracted saccades to divide the data in saccade-locked segments. This was done as follows (see also **Figure 2**). First, horizontal and vertical EOG were cleaned from noise. This was done by detecting values that exceeded five times the standard deviation. The signal around these peaks was cut out and interpolated. Next, blinks were detected and removed in vertical EOG: after band pass filtering between 2 Hz and 100 Hz, peaks exceeding a threshold of three times the standard deviation were considered to be blinks, removed and interpolated. Derivatives of the vertical and horizontal cleaned EOG signals were calculated using a derivative Gaussian filter with a standard deviation (sigma) of eight samples (about 31 ms). Values exceeding four times the standard deviation were associated with potential stimulus-to-stimulus
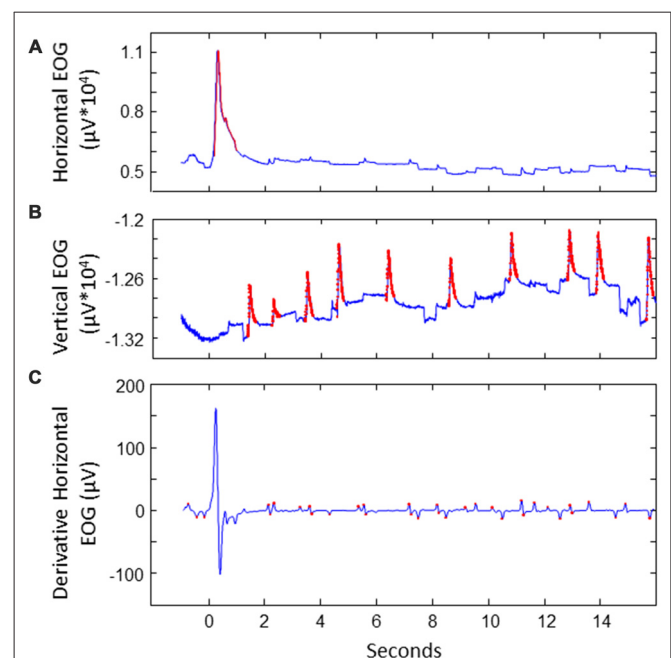
FIGURE 2 | Illustration of extracting saccades of interest for a sample trial. (A) Detecting noise (indicated in red) in EOG. This example shows the horizontal EOG. Noise was cut out and interpolated. **(B)** Detecting blinks in vertical EOG. Blinks were cut out and interpolated. **(C)** Detecting potential saccades of interest in the derivative of the cleaned EOG. This example shows the horizontal EOG.

saccades. We then looked at candidate saccades occurring between 100 ms and 800 ms after a next location was highlighted on the screen. The first saccade where the sign of the HEOG signal matched the direction of the stimulus-to-stimulus transition was selected as the saccade of interest. If no match was found in the HEOG signal, we looked for saccades in the VEOG with a sign that matched the vertical saccade jump. In 10% of the data no matching saccade was found. The EEG and EOG data was split into segments starting from the point of the highest saccade speed to 1 s after, and baselined using the first 100 ms of the epoch. EEG epochs with extremely high variance (standard deviation exceeding 50 times the standard deviation) were discarded as outliers.

Eye Tracking Data

One participant was excluded from the eye data analyses, because of technical difficulties with the eye tracking hardware during this measurement. After the measurement, the fixation locations were recalibrated using the 15 displayed stimulus positions to obtain higher gaze localization accuracy. Fixations were considered to be on the stimulus when the fixation position was within a radius of 150 pixels (4.4 cm or 6.4°) from the center of the current stimulus location. Fixation duration was determined as the time that eye fixation was on the stimulus. Only valid samples in a window starting at stimulus onset until 2 s after were taken into account. Pupil size was determined as the mean of the pupil size values over these same samples.

Classification

For classification, we used linear SVM models that were trained to distinguish between either targets vs. non-targets (for each of the load conditions) or hits vs. misses (in the high load condition) using 5-fold cross validation. Classification was performed using the Donders machine learning toolbox developed by van Gerven et al. (2013) and implemented in the FieldTrip open source Matlab toolbox (Oostenveld et al., 2011). The features were standardized to have mean 0 and standard deviation 1 on the basis of data from the training set. Included features were EEG voltages over a time interval of 250–1000 ms after peak velocity of the stimulus saccade, in which all EEG electrodes were included. In order to examine potential information from EOG leaking into EEG, we also used EOG voltages over the same time interval as features. Different models were trained using different combinations of EEG (i.e., SRP) features, EOG features, fixation duration and pupil size. Classification was performed separately for each participant and each load condition. Random selections of non-targets and hits were used in the training sets in order to match the numbers of available target and miss epochs to ensure balanced training of the model. For each participant, each load condition, each type of distinction (target vs. non targets and hits vs. misses) and each model we determined whether classification was above chance using a binomial test. An alpha level of 5% was used.
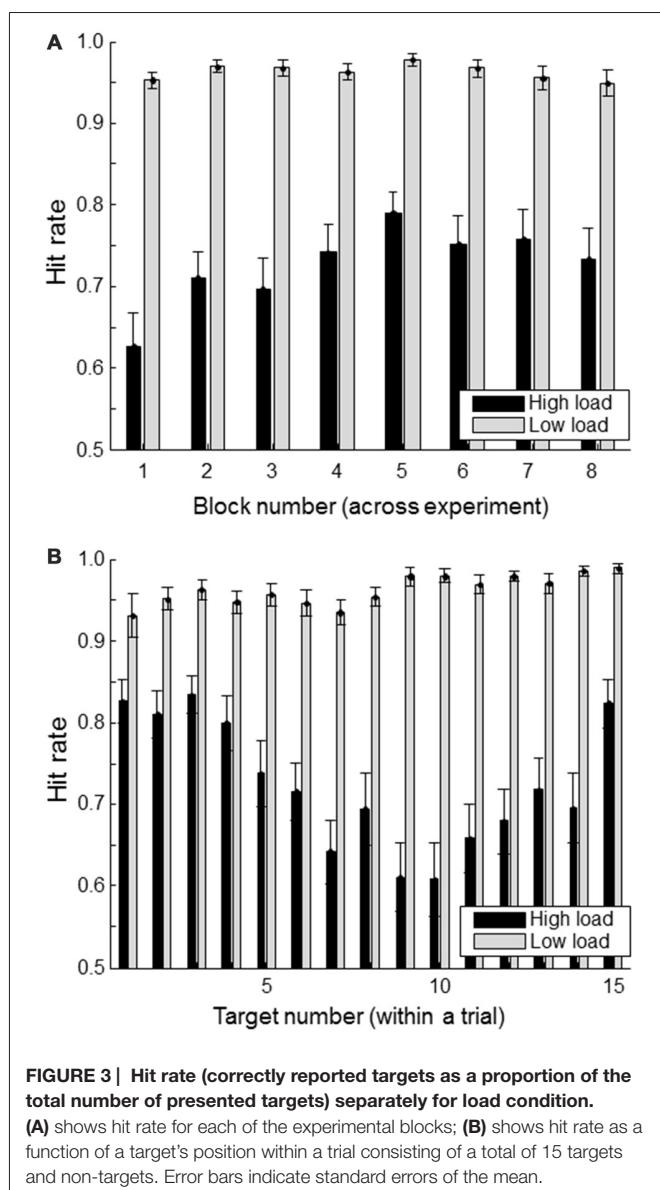
RESULTS

Behavioral Performance

Performance on the secondary (math) task was on average 62% correct (SD 17%) indicating that the secondary task was quite difficult and performance varied strongly between subjects. High workload data from both trials with correct and incorrect responses to the math task were included in the analyses. Note that performance on the math task is no direct indicator of workload. Performance could be high because a participant tried hard (high load) or, for that participant, the sum was easy (low load). Conversely, low performance could be caused by lack of trying (low load) or because the participant simply did not manage, despite trying hard (high load). There was no evidence for participants choosing to focus on the one rather than the other as indicated by the lack of (negative) correlation between participants' performance on the math task and performance on the monitoring task (Pearson correlation: $r = 0.25$, $p = 0.27$).

Figure 3A shows the hit rate of the primary task (defined as the proportion of "#FA#" targets whose location was correctly indicated) for successive blocks (of 11 trials each) in the high and low workload conditions. There seems to be some indication of a learning effect in the high load condition with increasing performance up to block 5. The **Figure 3B** shows the hit rate as a function of when the target was presented within a trial. For the high load condition, it is clear that targets presented at the beginning or the end of a trial are remembered better than the ones in between. This is consistent with primacy and recency effects.

As expected and intended, the hit rate in the high load condition was much lower (average hit rate of 0.73 (SD 0.13) than in the low load condition (average hit rate of 0.96, SD 0.04). Except for one participant who only missed nine targets, all participants missed at least 22 targets in the high load condition, with an average of 72 missed targets (range: [9, 133]). In the low load condition, the average number of missed targets was eight (with a range of [0, 32]). We consider the average number of eight missed targets in the low load condition as too little to do meaningful hit vs. miss comparisons in this condition.

Missed targets can be accompanied by wrongly identified targets (i.e., false alarms), or not (resulting in less targets being reported than being presented). We found that in general, the latter is the case. In the high load condition, the number of false alarms was on average 42 (range of [5, 132]). Given the number of missed targets, this means that participants reported on average 30 targets less than the number of targets that was actually shown (with a range of [−103, 10]). While brain processes will be different when a target is not identified as a target (definitely leading to not report the target at all) compared to when a target has been identified but not properly encoded (which, depending on the reporting strategy of the participant, could lead to not reporting it at all or to an accompanying false alarm), we do not distinguish between not reporting a target at all and indicating a wrong location in this study. Given the design of our experiment, and the results described in the following,

FIGURE 3 | Hit rate (correctly reported targets as a proportion of the total number of presented targets) separately for load condition.
(A) shows hit rate for each of the experimental blocks; **(B)** shows hit rate as a function of a target's position within a trial consisting of a total of 15 targets and non-targets. Error bars indicate standard errors of the mean.

TABLE 1 | Means and standard deviations of saccade latency (time of peak velocity of the stimulus saccade relative to stimulus onset), separately for each low and high load condition, targets, non-targets, hits and misses.

	Low load		High load	
	mean	**SD**	**mean**	**SD**
Target	157	40	221	52
Non-target	165	45	229	53
Hit	154	38	199	43
Miss	267	245	284	97

The gray font of the hits and misses in the low load condition signifies that because of few misses in the low load condition, the hit trials are about the same as the target trials and the misses are represented by few data points.

for targets than non-targets (see "Eye—Fixation Duration" Section). The object fixated prior to a target is more likely to have been a non-target (that does not keep the eyes linger relatively long and results in a short saccade latency for the next object) compared to the object fixated prior to non-target.
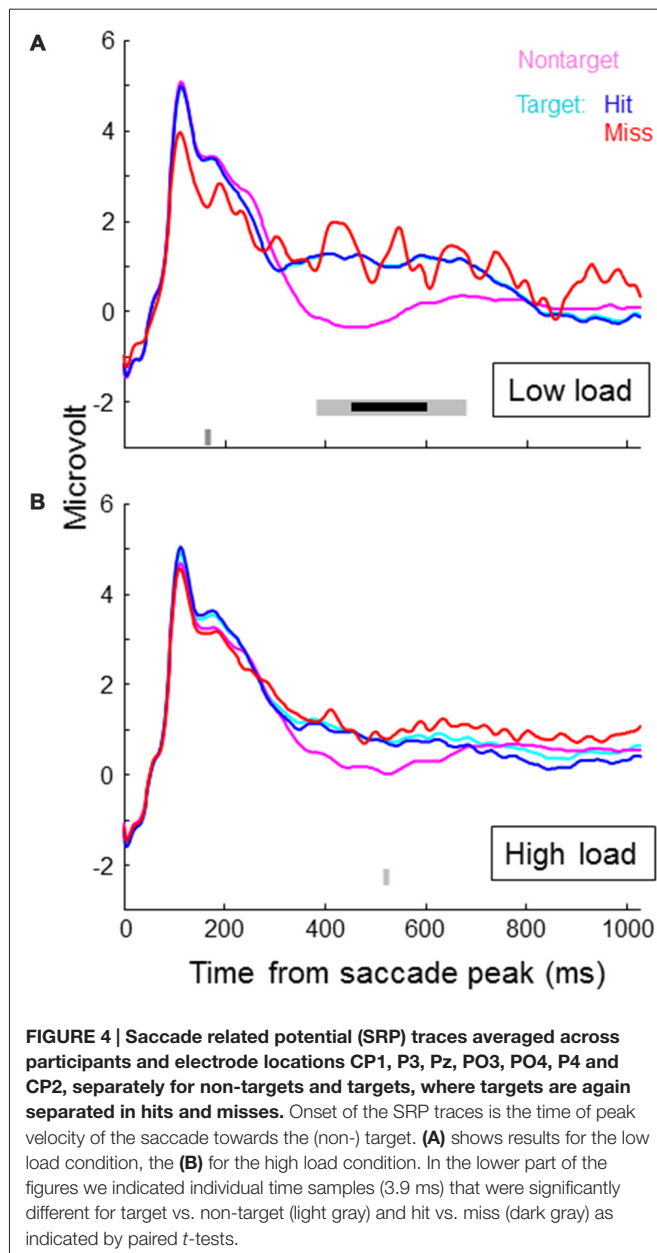
EEG—SRP

Figure 4 shows SRP traces associated with targets, non-targets, hits and misses for each of the two load conditions, averaged across participants and electrode locations around Pz (CP1, P3, Pz, PO3, PO4, P4, CP2). In the lower part of the figures we indicated individual time samples (3.9 ms) that were significantly different for target vs. non-target (light gray) and hit vs. miss (dark gray) as indicated by paired t-tests (alpha level of 5%). Around 500 ms, target SRPs are larger than non-target SRPs, which is consistent with a stronger P300 for targets than non-targets. The difference appears stronger in the low load condition—only in this condition, the difference between target and non-target traces reached significance for an uninterrupted interval of almost 300 ms. After correcting for multiple testing (Benjamini and Hochberg, 1995), this interval is reduced to around 150 ms (indicated by the bold, black line in **Figure 4**). When examining traces for load condition separately, hit and miss traces did not differ for a substantial period of time. When collapsing across load conditions, the higher values for miss traces towards the end of the epoch becomes significant. However, it is clear that at least up to 450 ms, miss traces overlap with hit traces. Miss traces certainly do not lie in between hit and non-target traces.

Exploring average traces for all individual channels revealed no significant differences between target and non-target, and hits and misses in the math condition. For the non-math condition, we found significant effects as reflected in **Figure 4** (higher voltages for targets compared to non-targets around 500 ms) for channel CP1, Pz and P4. P7 showed a lower voltage for targets compared to non-targets around 250 ms, and F8 a lower voltage around 450 ms.

Eye—Fixation Duration

Figure 5 shows average fixation duration. As expected, fixation duration was longer for targets than for non-targets, both

we take both to mean that the target has not been properly encoded. However, it is important to keep in mind that treating these results as the same may not be appropriate under other circumstances.

Saccade Latencies

Table 1 shows the saccade latencies, defined as the peak velocity of the stimulus saccade relative to stimulus onset. Relative to this point SRP onsets were determined. Latencies toward targets that are missed are longer than toward hit targets ($t_{(19)} = 2.18$, $p = 0.04$ for the low load condition; $t_{(19)} = 4.85$, $p < 0.01$ for high load). Saccade latencies toward targets are shorter than toward non-targets. Although the difference is small (8 ms in both low and high load conditions), it is statistically significant (respectively ($t_{(19)} = 2.83$, $p = 0.01$ for low load; $t_{(19)} = 2.12$, $p = 0.05$ for high load). We think that this effect is mediated by a longer fixation duration

FIGURE 4 | Saccade related potential (SRP) traces averaged across participants and electrode locations CP1, P3, Pz, PO3, PO4, P4 and CP2, separately for non-targets and targets, where targets are again separated in hits and misses. Onset of the SRP traces is the time of peak velocity of the saccade towards the (non-) target. **(A)** shows results for the low load condition, the **(B)** for the high load condition. In the lower part of the figures we indicated individual time samples (3.9 ms) that were significantly different for target vs. non-target (light gray) and hit vs. miss (dark gray) as indicated by paired t-tests.

FIGURE 5 | Average fixation duration for the high load and low load condition, separately for targets and non-targets (A) and, for the targets, separately for hits and misses **(B)**. The light color of the low load condition bars in the **(B)** signify that these represent little data since there were few misses in the low load condition. Error bars indicate standard errors of the mean. Stars indicate significant differences (**representing $p < 0.01$, *representing $p < 0.05$).
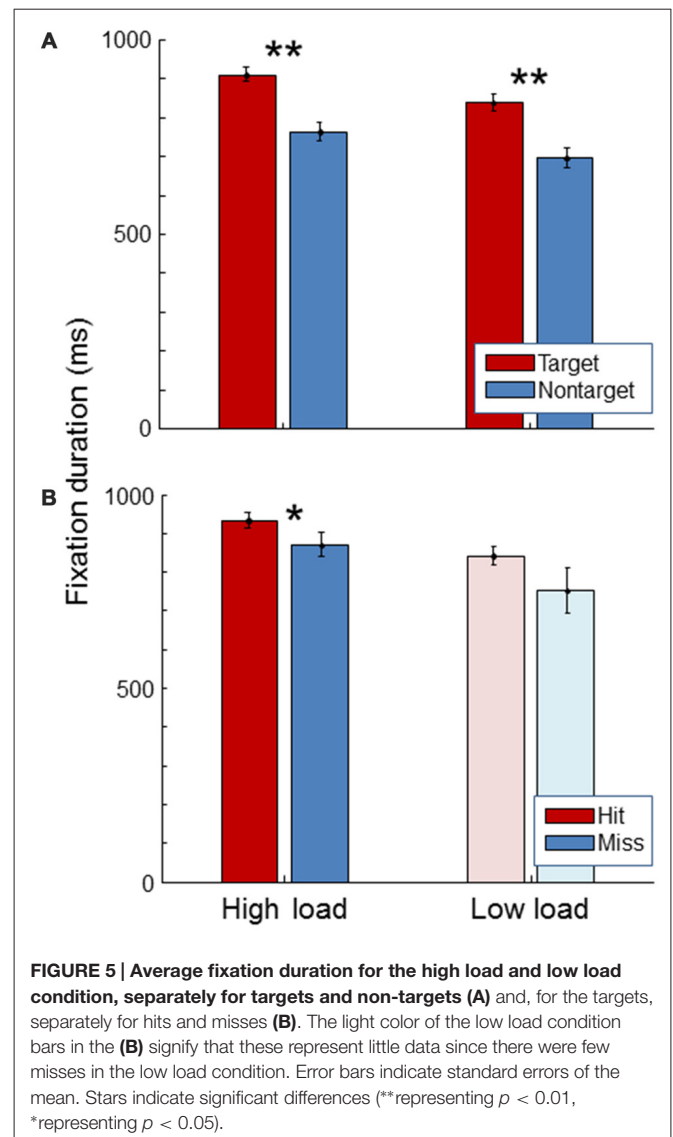
in the high and in the low load condition (paired t-tests, $t_{(19)} = 11.47$, $p < 0.01$; $t_{(19)} = 11.43$, $p < 0.01$). Fixation duration was longer for hits than for misses in the high load condition (paired t-test, $t_{(19)} = 2.38$, $p = 0.03$). The same trend was seen in the low load condition. No significant difference was found in fixation duration between high and low workload conditions (paired t-test, $t_{(19)} = 1.40$, $p = 0.18$).

Eye—Pupil Size

Figure 6 shows pupil size. For both high and low load conditions, pupil size was the same between targets and non-targets (paired t-test, $t_{(19)} = 0.54$, $p = 0.59$; $t_{(19)} = 0.12$, $p = 0.91$). As expected, pupil size was significantly larger
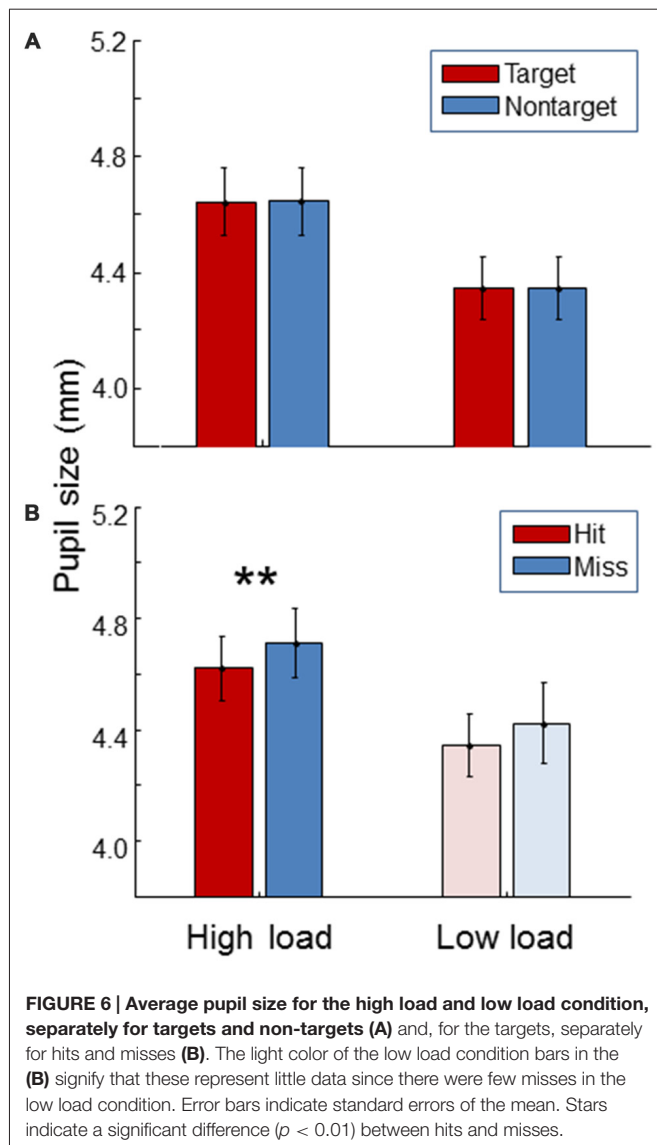
during the high load condition compared to the low load condition (paired t-test, $t_{(19)} = 12.5$, $p < 0.01$). Additionally, pupil size was found to be larger for misses than for hits in the high workload condition (paired t-test, $t_{(19)} = 4.25$, $p < 0.01$). The same trend was found in the low load condition.

Single Trial Analysis

Figure 7 shows the classification accuracy between targets and non-targets (blue bars) separately for the low and high load conditions, and hits and misses (red bars) for the high load condition. As expected from the average SRPs as presented above and from previous studies, single fixation classification was possible for target vs. non-targets based on SRP in the low load condition (on average 65% correct, with performance significantly above chance for 13 out of 21 participants according to binomial

FIGURE 6 | Average pupil size for the high load and low load condition, separately for targets and non-targets (A) and, for the targets, separately for hits and misses **(B)**. The light color of the low load condition bars in the **(B)** signify that these represent little data since there were few misses in the low load condition. Error bars indicate standard errors of the mean. Stars indicate a significant difference (*p* < 0.01) between hits and misses.

FIGURE 7 | Classification results of models averaged over individual participants that distinguish between target and non-target fixations (blue bars) and hits and misses (red bars), separately for data from the low load condition (upper half of the figure—no results for hit and miss due to few misses) and the high load condition (lower half of the figure). On the left the features that the models were based on are indicated: SRP, fixation duration, pupil size, combined eye features, and a combination of all features. Error bars indicate standard errors of the mean.

FIGURE 8 | Classification results of models averaged over individual participants that distinguish between target and non-target fixations (blue bars) and hits and misses (red bars), separately for data from the low load condition (upper half of the figure—no results for hit and miss due to few misses) and the high load condition (lower half of the figure). On the left the features that the models were based on are indicated: SRP, EOG, and a combination of SRP and EOG features. Error bars indicate standard errors of the mean.

tests). For the high load condition average classification performance on the basis of SRP reaches 59% correct (7 out of 21 participants above chance). For both load conditions, classification accuracy between target and non-targets is highest when SRP features are used. Adding fixation duration and pupil size does not (substantially) improve performance.

Another picture emerges for the distinction between hits and misses—which we could only meaningfully examine for the high load condition due to the very small number of misses in the low load condition. Classification based on SRPs is around chance level (52% correct), while classification based on pupil size results in an average single fixation classification accuracy of 58%. Adding fixation duration does not (substantially) improve performance; adding SRP rather decreases performance. Distinguishing between hits and misses did not reach above chance performance for single subjects (note that the power of the binomial tests is much weaker for hits vs. misses
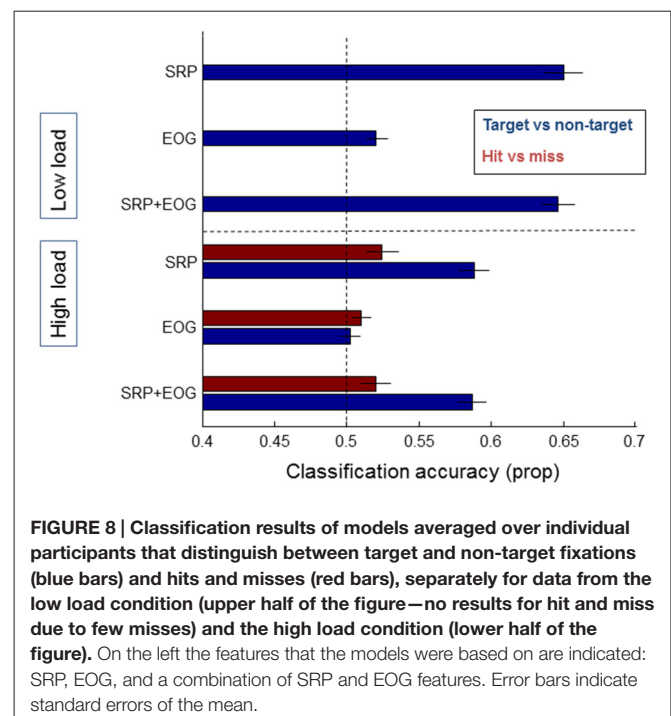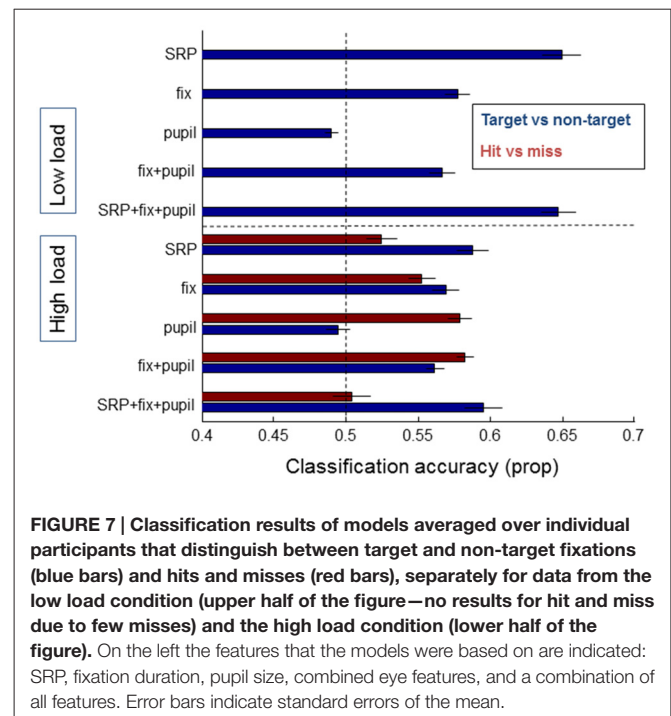
compared to targets vs. non-targets where more data was available).

Figure 8 shows results for classification based on EOG features, and EOG and SRP features combined, with the result of the SRP based models as a comparison. When using EOG, classification accuracy between either targets and non-targets or

hits and misses does not rise over 52% and adding EOG to SRP does not improve classification.

DISCUSSION

We investigated a structured search task where fixated targets that observers can easily identify are relatively likely to remain unreported due to a concurrent secondary task that is expected to interfere with working memory. We found that in such a case, SRPs differ between targets and non-targets (consistent with a target eliciting a P300), but not between hit targets and missed targets. Fixation duration was longer for targets than non-targets as expected, and also for hits compared to misses. Pupil size did not differ between targets and non-targets, but was larger for misses than for hits. In sum, EEG features appeared more suitable to distinguish targets from non-targets while eye features (especially pupil size) were suitable to distinguish fixated targets that were subsequently reported from those that were not. These results were also reflected in single trial classification analyses.

We interpret our findings as reflecting distinct underlying cognitive processes as discussed further below.

SRP P300

Differences between target and non-target SRPs are as expected with a higher amplitude P300 for targets compared to non-targets. This difference is smaller between the average target and non-target SRP traces in the high compared to the low load condition. Previous studies also found the P300 to be less pronounced under high load conditions (Allison and Polich, 2008; Gherri and Eimer, 2011; Pratt et al., 2011). Also for ERPs following fixations, smaller target P300s during high workload compared to low workload have been reported before Ries et al. (2016). This reducing effect of workload on the P300 is consistent with less attentional resources being available for the target detection task.

While the P300 reflects attentional processes, this did not translate to larger P300s for hit compared to missed targets. Clearly, SRPs associated with missed targets more closely resembled hit target rather than non-target SRPs. This suggests that, as we intended, there is no problem in target identification and that the amount of attention allocated to the fixated object around the time of fixation is not critical for its encoding in memory.

Fixation Duration

As expected, fixation duration is longer for targets than for non-targets. We also found it to be longer for hits than for misses. A short duration of a fixation being associated with misses could in principle be caused by participants moving the eyes from the target to the next stimulus too quickly, or because participants lingered relatively long on the stimulus that was fixated before the target, which causes a late arrival on the target and leaves less time for target fixation. Comparing the differences in hit-miss saccade latency to the difference in hit-miss fixation duration indicates that misses are mostly due to arriving late rather than leaving too early.

Note that in the present structured search task, potentially relevant information appeared every second, encouraging rather long, and relatively invariable fixation durations. Thus, in different types of search tasks fixation duration effects can be expected to be stronger.

Also note that the rather strong association between fixation duration and hits and misses could potentially have led to confounding effects in the SRPs, i.e., differences between hit and miss SRPs that cannot be attributed to memory or attentional processes but are e.g., due to visual processes caused by systematic timing differences in the change of stimulus appearance. In this study, we did not find any significant difference between hit and miss SRPs, but this should be kept in mind for similar future studies.

Pupil Size

Pupil size is larger in the high load compared to the low load condition. An increased pupil size with an increase in memory load or workload is a robust finding in the literature (Kahneman and Beatty, 1966; Beatty, 1982; Hogervorst et al., 2014). We also observed the hypothesized effect of larger pupil size during missed targets, consistent with a momentary high workload being the cause of the miss. Such momentary increase in cognitive workload could have been caused by a difficult (part in the) math trial, or by having to store yet another target. Also, momentary fluctuations could be caused by participants sometimes giving up on the math task half way during a trial. Our data do not reflect differences in phasic pupil size responses—there was no significant difference in pupil size between targets and non-targets as determined by the median size over the fixation interval. While our dependent measure of pupil size was not optimal to capture differences in pupil dilation that start to diverge at longer latencies and reach their maximum difference at about 1.5 s post fixation (Hong et al., 2014; Jangraw et al., 2014), averaged pupil size traces in our data did not show the expected differential pattern for targets and non-targets. At present, we do not know why this difference was not observed. The average pupil size was quite large (around 4.5 mm diameter) which may have played a role in attenuating the phasic response.

In sum, SRP and eye features provide complementary information in the search task under study. Saccade-related P300s signify that a target has been detected but not that it has been encoded for successful recall. For that, fixation duration and pupil size are informative, likely because they respectively reflect time taken to store the target and variations in workload associated with the secondary task.

Application

Determining whether an observer is looking at a target (i.e., at an object that is of interest) using SRPs and eye related features may be useful in two general areas. If it is not known what visual information is important (in a particular situation, or for a particular person) such features provide a means to examine what is of interest to an observer without requiring conscious behavioral responses. If it is known what information

is important, brain and eye signals can be used to judge whether this important information is properly recognized as being important. If we know whether an observer is gazing at relevant information (as judged from the brain and eye signals, or as *a priori* knowledge of the task), determining whether this information is going to be remembered could be used for deciding whether or not to present the information again or in another way. Target recognition and encoding indicators may also be used as indicators of (momentarily) suboptimal performance so that the system can advise the human observer or operator to take break. The interplay between user state detection and responses to individual targets may be used in higher level classification strategies. If observers are detected to be in a high memory load situation (because of a large pupil size and EEG features), a model that classifies fixated targets (as detected through SRP) into hits and misses may be activated.

In addition to physiological or eye-based sources of information, task- and context-related information can be used directly in models for state monitoring and behavior prediction. For instance, knowledge of the presence of an additional or difficult task will add to the likelihood of misses, and indicate that a classifier distinguishing between hits and misses should be put to work. In some cases, knowledge of the task may strongly influence the interpretation of eye and brain signals that occur while searching and storing visual information. As mentioned before, low workload as may be indicated by a small pupil size, can be associated with misses in cases that a target is difficult to identify and there are no other tasks (the situation as in Dias et al., 2013) or it can be associated with hits in cases that a target is easy to identify and there is a concurrent task (current study). Also, the type of search task influences which physiological or eye based features can be expected to be informative. When there is no time pressure, fixation duration is expected to be more informative compared to when there is.

It has to be realized that the certainty with which it is possible to determine whether an observer is looking at a target or whether an observer is going to miss a target at the individual fixation level can never achieve perfect accuracy. The highest fixation classification accuracy as found in the current study is 65% and associated with distinguishing targets from non-targets in the low load condition. While it is obviously difficult to compare classification performance across studies, a classification accuracy of around 65% (with a 50% chance level) has also been found in other studies distinguishing targets from non-targets using SRPs (Brouwer et al., 2013: 62%; Wenzel et al.,

2016: 67%; Kaunitz et al., 2014: 63%). SRP studies that report a relatively good performance (Touryan et al., 2016: up to around 82%; Ušćumlić and Blankertz, 2016: a mean AUC of almost 80%) employed methods of subtracting out eye movement artifacts. Brouwer et al. (2014) found a decrease of the Equal Error Rate for distinguishing between target and non-target SRPs from 40% to 31% when removing eye artifacts. Amongst a range of possible techniques to boost classification performance, removing eye movement components from SRP traces seems to be an important one.

For real applications it is also important to consider the context and circumstances where uncertain information about "targetness" or memory encoding, based on eye and brain signals, are expected to add to other available or easy-to-retrieve information such as behavioral data (Brouwer et al., 2015). Preferably, this would be in an environment with minimal noise affecting the brain and eye signals, but also situations where the signal is strong—e.g., low workload scenarios are beneficial when exploiting the P300 signal to distinguish between targets and non-targets (current study, Thurlings et al., 2013; Ries et al., 2016). An example that may capture these elements is robust image triage. If a subset of rapidly presented images is identified for careful review and inspection (e.g., x-ray images of luggage), it would be prudent to include images with a relatively high risk of (missed) targets.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Allison, B. Z., and Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biol. Psychol.* 77, 277–283. doi: 10.1016/j.biopsycho.2007.10.014

Azizian, A., and Polich, J. (2007). Evidence for attentional gradient in the serial position memory curve from event-related potentials. *J. Cogn. Neurosci.* 19, 2071–2081. doi: 10.1162/jocn.2007.19.12.2071

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* 91, 276–292. doi: 10.1037/0033-2909.91.2.276

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 289–300.

Brouwer, A.-M., Brinkhuis, M. A. B., Reuderink, B., Hogervorst, M. A. J., and van Erp, J. B. F. (2014). "Fixation-related potentials: foveal versus parafoveal target identification," in *Proceedings of the 6th International Brain-Computer Interface Conference 2014* (Graz, Switzerland).

Brouwer, A.-M., Reuderink, B., Vincent, J., van Gerven, M. A. J., and van Erp, J. B. F. (2013). Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. *J. Vis.* 13:17. doi: 10.1167/13.3.17

Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., and Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Front. Neurosci.* 9:136. doi: 10.3389/fnins.2015.00136

Dandekar, S., Ding, J., Privitera, C., Carney, T., and Klein, S. A. (2012a). The fixation and saccade P3. *PLoS One* 7:e48761. doi: 10.1371/journal.pone.0048761

Dandekar, S., Privitera, C., Carney, T., and Klein, S. A. (2012b). Neural saccadic response estimation during natural viewing. *J. Neurophysiol.* 107, 1776–1790. doi: 10.1152/jn.00237.2011

Dias, J. C., Sajda, P., Dmochowski, J. P., and Parra, L. C. (2013). EEG precursors of detected and missed targets during free-viewing search. *J. Vis.* 13:13. doi: 10.1167/13.13.13

Fabiani, M., Karis, D., and Donchin, E. (1990). Effects of mnemonic strategy manipulation in a Von Restorff paradigm. *Electroencephalogr. Clin. Neurophysiol.* 75, 22–35. doi: 10.1016/0013-4694(90)90149-e

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Gherri, E., and Eimer, M. (2011). Active listening impairs visual perception and selectivity: an ERP study of auditory dual task costs on visual attention. *J. Cogn. Neurosci.* 23, 832–844. doi: 10.1162/jocn.2010.21468

Hale, K. S., Fuchs, S., and Berka, C. (2008). "Driving EEG cognitive assessment using eye fixations," in *Paper Presented at the Applied Human Factors and Ergonomics Conference* (Las Vegas, NV).

Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* 8:322. doi: 10.3389/fnins.2014.00322

Hong, L., Walz, J. M., and Sajda, P. (2014). Your eyes give you away: prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. *PLoS One* 9:e91321. doi: 10.1371/journal.pone.0091321

Jangraw, D. C., Wang, J., Lance, B. J., Chang, S. F., and Sajda, P. (2014). Neurally and ocularly informed graph-based models for searching 3D environments. *J. Neural Eng.* 11:046003. doi: 10.1088/1741-2560/11/4/046003

Kahneman, D., and Beatty, J. (1966). Pupil diameter and load on memory. *Science* 154, 1583–1585. doi: 10.1126/science.154.3756.1583

Kamienkowski, J. E., Ison, M. J., Quiroga, R. Q., and Sigman, M. (2012). Fixation-related potentials in visual search: a combined EEG and eye tracking study. *J. Vis.* 12:4. doi: 10.1167/12.7.4

Kamp, S., Forester, G. R., Murphy, A. R., Brumback, T., and Donchin, E. (2012). "Testing a distinctiveness explanation of the primacy effect in free recall using event-related potentials," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, eds N. Miyake, D. Peebles and R. P. Cooper (Austin, TX: Cognitive Science Society), 539–544.

Karis, D., Fabiani, M., and Donchin, E. (1984). "P300" and memory: individual differences in the von Restorff effect. *Cogn. Psychol.* 16, 177–216. doi: 10.1016/0010-0285(84)90007-0

Kaunitz, L. N., Kamienkowski, J. E., Varatharajah, A., Sigman, M., Quiroga, R. Q., and Ison, M. J. (2014). Looking for a face in the crowd: fixation-related potentials in an eye-movement visual search task. *Neuroimage* 89, 297–305. doi: 10.1016/j.neuroimage.2013.12.006

Luo, A., Parra, L. C., and Sajda, P. (2009). "We find before we look: neural signatures of target detection preceding saccades during visual search," in *Poster Presented at Vision Sciences Society 9th Annual Meeting* (Naples, FL).

Nieuwenhuis, S., De Geus, E. J., and Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology* 48, 162–175. doi: 10.1111/j.1469-8986.2010.01057.x

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Porter, G., Troscianko, T., and Gilchrist, I. D. (2007). Effort during visual search and counting: insights from pupillometry. *Q. J. Exp. Psychol.* 60, 211–229. doi: 10.1080/17470210600673818

Pratt, N., Willoughby, A., and Swick, D. (2011). Effects of working memory load on visual selective attention: behavioral and electrophysiological evidence. *Front. Hum. Neurosci.* 5:57. doi: 10.3389/fnhum.2011.00057

Ries, A. J., Touryan, J., Ahrens, B., and Connolly, P. (2016). The impact of task demands on fixation-related brain potentials during guided search. *PLoS One* 11:e0157260. doi: 10.1371/journal.pone.0157260

Thurlings, M. E., van Erp, J. B. F., Brouwer, A.-M., and Werkhoven, P. J. (2013). "Controlling a tactile ERP-BCI in a dual-task," in *IEEE Transactions on Computational Intelligence and AI in Games.*

Touryan, J., Ries, A. J., Lawhern, V., and Connolly, P. (2016). "Classification of visual target detection during guided search using EEG source localization," in *Proceedings of the Sixth International Brain-Computer Interface Meeting*, eds G. Müller-Putz, J. Huggins and D. Steyrl (Austria: Verlag der Technischen Universitat Graz), 146. Available online at: http://doi.org/10.3217/978-3-85125-467-9.

Ušćumlić, M., and Blankertz, B. (2016). Active visual search in non-stationary scenes: coping with temporal variability and uncertainty. *J. Neural Eng.* 13:016015. doi: 10.1088/1741-2560/13/1/016015

van Gerven, M., Bahramisharif, A., Farquhar, J., and Heskes, T. (2013). Donders Machine Learning Toolbox (DMLT). Available online at: https://github.com/distrep/DMLT, version 26/06/2013.

Vázquez Marrufo, M. M., Vaquero, E., Cardoso, M. J., and Gomez, C. M. (2001). Temporal evolution of α and β bands during visual spatial attention. *Cogn. Brain Res.* 12, 315–320. doi: 10.1016/s0926-6410(01)00025-8

Wenzel, M. A., Golenia, J., and Blankertz, B. (2016). Classification of eye fixation related potentials for variable stimulus saliency. *Front. Neurosci.* 10:23. doi: 10.3389/fnins.2016.00023

Detecting Pilot's Engagement Using fNIRS Connectivity Features in an Automated vs. Manual Landing Scenario

Kevin J. Verdière*, Raphaëlle N. Roy and Frédéric Dehais

ISAE-SUPAERO, Institut Supérieur de l'Aéronautique et de l'Espace, Université Fédérale de Midi-Pyrénées, Toulouse , France

Monitoring pilot's mental states is a relevant approach to mitigate human error and enhance human machine interaction. A promising brain imaging technique to perform such a continuous measure of human mental state under ecological settings is Functional Near-InfraRed Spectroscopy (fNIRS). However, to our knowledge no study has yet assessed the potential of fNIRS connectivity metrics as long as passive Brain Computer Interfaces (BCI) are concerned. Therefore, we designed an experimental scenario in a realistic simulator in which 12 pilots had to perform landings under two contrasted levels of engagement (manual vs. automated). The collected data were used to benchmark the performance of classical oxygenation features (i.e., Average, Peak, Variance, Skewness, Kurtosis, Area Under the Curve, and Slope) and connectivity features (i.e., Covariance, Pearson's, and Spearman's Correlation, Spectral Coherence, and Wavelet Coherence) to discriminate these two landing conditions. Classification performance was obtained by using a shrinkage Linear Discriminant Analysis (sLDA) and a stratified cross validation using each feature alone or by combining them. Our findings disclosed that the connectivity features performed significantly better than the classical concentration metrics with a higher accuracy for the wavelet coherence (average: 65.3/59.9 %, min: 45.3/45.0, max: 80.5/74.7 computed for HbO/HbR signals respectively). A maximum classification performance was obtained by combining the area under the curve with the wavelet coherence (average: 66.9/61.6 %, min: 57.3/44.8, max: 80.0/81.3 computed for HbO/HbR signals respectively). In a general manner all connectivity measures allowed an efficient classification when computed over HbO signals. Those promising results provide methodological cues for further implementation of fNIRS-based passive BCIs.

Keywords: fNIRS, passive brain-computer-interface, classification, functional connectivity, wavelet coherence, engagement

1. INTRODUCTION

It is largely admitted that pilot error represents a major cause of aircraft crashes (Li et al., 2001), being more frequently cited than mechanical failure. Safety statistics show that the progressive introduction of automation in the cockpit since the 1960's has improved safety, with modern "computerized" cockpits taking pride in an accident rate half that of the previous generation of aircraft. However, it appears that such technologies have created a new category of potentially

deadly incidents whereby crews are unable to comprehend the situation presented before them, and persevere in erroneous decision-making (Dehais et al., 2010, 2015). This is especially true for the final approach and landing phases that represent almost half the on-board accidents and fatal accidents (Myers and Arnold, 2016).

Without question the development of automation has dramatically changed the role of the crew from "direct (manual) controllers" to "system supervisors/decision makers." Both increased trust in automation and complexity of these computerized systems (Sarter et al., 1997; Dehais et al., 2012; Tessier and Dehais, 2012) reduce crew's basic flying abilities, and leave them ill equipped to cope with emergency situations when automation fails (Mumaw et al., 2001). Another drawback of automation is that it imposes long periods of inactivity and thus dramatically decreases pilot's vigilance (Wright and McGown, 2001). For instance, some recent surveys disclosed that 56% of British Airways pilots experienced sleep while on duty (Steptoe and Bostock, 2012; Reis et al., 2013). These operational situations show that automation can vary pilot's engagement from a very low engagement state (disengagement) that induces states of low vigilance and mind wandering, to a very high engagement state (over-engagement) yielding to perseveration and attentional tunneling (Wickens and Alexander, 2009). These extreme cognitive states may jeopardize safety and advocate for the introduction of monitoring solutions.

The idea of introducing physiological data into Human Machine Interface, called "Physiological Computing" (Fairclough, 2008) could allow the system to take the operator's states into account. Brain monitoring techniques such as passive Brain Computer Interfaces (pBCI) have shown their ability to detect and characterize several operator's mental state such as workload, fatigue or more generally engagement (Zander et al., 2010; Zander and Kothe, 2011; Khan and Hong, 2015; Roy and Frey, 2016). Building a system capable of doing a continuous monitoring or detecting some operator's degraded states would potentially permit it to adapt to this change to optimize both safety and performance. Such kinds of closed-looped systems are the ultimate goal for neuroadaptative technology.

While the real-time identification of these degraded mental states still remains a challenge, a first reasonable step is to characterize the brain activity when flying with and without the use of automation. One possible solution to meet this goal is to consider the use of functional near infra-red spectroscopy (fNIRS). Less popular in the BCI community than electro-encephalography (Cutini and Brigadoi, 2014), mainly due to its low temporal resolution, this brain imaging technique presents several advantages for ecological settings as its signal is less affected by electrical and motion artifacts. Moreover, its high spatial resolution allows to give a direct access to specific brain structures without additional computational costs as long as cortical areas are concerned. Thus, several studies have shown the potential of fNIRS to infer several mental states under laboratory settings or ecological settings such as flight simulators (Ayaz et al., 2012; Gateau et al., 2015).

Classically, the authors used the relative variation of local HbO and HbR concentration and related features (e.g., slope,

area under the curve, skewness) to relate cerebral activation to specific cognitive tasks (Tai and Chau, 2009; Durantin et al., 2015; Gateau et al., 2015). Yet the goal is always to improve the estimation, especially in critical settings. A solution proposed by some authors is to use connectivity measures (Borghini et al., 2014) to account for brain dynamics (for a review on functional connectivity see Bastos and Schoffelen, 2015). Indeed, cognition cannot be reduced to activation of specialized brain areas but should rather been seen as the cooperation among large scale distributed neural networks (Siegel et al., 2012; Hutchison et al., 2013; van den Heuvel and Sporns, 2013). In other words, examining spontaneous hemodynamic fluctuations can provide us a great picture of the functional architecture of the brain (Fox and Raichle, 2007)

Moreover, connectivity features have been used with success to estimate various mental states based on EEG data (Roy and Frey, 2016) in laboratory settings as well as in ecological settings. For instance, a recent study combined EEG connectivity analysis and crew monitoring in simulator and showed differences in connectivity patterns during different flight phases (Toppi et al., 2016). A few studies combined optical brain imaging like fNIRS with connectivity analysis (Lu et al., 2010; Funane et al., 2011; Cui et al., 2012; Molavi et al., 2012; İşbilir et al., 2016) either to identify brain dynamics or brain-to-brain relationship yet they did not perform mental state estimation. Hence, the contribution of connectivity measures for fNIRS based on mental state estimation is yet to be assessed.

Classical correlation/covariance measures were successfully used in EEG (Gevins et al., 1987), however some spontaneous oscillation observed in blood-related imaging (fNIRS and fMRI) seems to be frequency specific, especially Low Frequency Oscillation (LFO) around 0.1 Hz (Obrig et al., 2000; Tong and Frederick, 2010). Knowing this, frequency specific connectivity metrics such as coherence and also wavelet coherence were used, which has gained some momentum in fNIRS signal analysis (Rowley et al., 2006; Cui et al., 2012; Holper et al., 2012; Mirelman et al., 2014).

The objectives of the present study are : (i) to evaluate the feasibility to estimate the pilot's engagement using fNIRS connectivity measures in an ecological setting such as a flight simulator. Secondly : (ii) to assess the potential of connectivity measures to better characterize engagement than classical measures.

To meet these goals, a simplified task was designed whereby pilots had to perform different manual and automated landings. Parieto-occipital areas were targeted as they play a key role for visual attention, particularly involved while flying (Dehais et al., 2016). Prefrontal cortex activity was also measured as its activation reflects mental demands (Gateau et al., 2015; Moro et al., 2016) and top down regulation. Off-line classification was performed over different classical metrics (average, peak, variance, skewness, kurtosis, area under curve and slope) and connectivity metrics to identify the most predictive ones. Regarding connectivity features, classical dependency measures such as : Covariance, Pearson's correlation (Greenblatt et al., 2012), Spearman's correlation (Spearman, 1904) and some spectral measures : magnitude squared

coherence (Mandel and Wolf, 1976) and the wavelet coherence (Torrence and Compo, 1998; Lachaux et al., 2002; Grinsted et al., 2004) were compared (for review on connectivity metrics see Lachaux et al., 2002; Greenblatt et al., 2012).

2. MATERIALS AND METHODS

2.1. Participants

Twelve visual flight rules (VFR) pilots (11 males, mean group age 24 ± 3) completed the experiment. Pilots had normal or corrected-to-normal vision, normal hearing, and no psychiatric disorders. They all had medical clearance to fly. After providing written informed consent, they were instructed to complete a 5-min task training. Typical total duration of a subject's session (informed consent approval, practice task and real task) was about 1 h. This work was approved by the local ISAE-SUPAERO committee (Approval Number: CERNI-Université fédérale de Toulouse-2017-057).

2.2. Experimental Design

The protocol consisted in 8 scenarios in a flight simulator : 4 in manual landing and 4 in automated landing. The Airbus A320 full motion simulator at ISAE-SUPAERO (French Aeronautical Engineering School in Toulouse) was used to conduct the experiment in ecological conditions. It simulates a twin-engine aircraft flight mode. The user interface is composed of a Primary Flight Display (PFD), a Navigation Display and an upper Electronic Central Aircraft Monitoring Display page. The pilot also had a Flight Control Unit (FCU) to interact with the autopilot (**Figure 1**).

The scenarios were divided into 3 phases: a rest phase, a cruise phase and lastly a landing phase, which were performed either in manual mode (i.e., hard condition) in which they control the aircraft speed and trajectory, or with the autopilot engaged (easy condition; **Figure 2**). Landing conditions (Auto vs. Manual) were pseudo-randomly distributed.

During the cruise phase, the autopilot was engaged and the pilots were asked to relax. This phase was mostly set to serve as a baseline. When approaching the ILS (Instrument Landing System) range (approximately 2 min) they were asked either to let the autoflight system perform the landing or to disengage the automation to manually land the aircraft. Autopilot and auto throttle deactivation was done by pushing a red button on the flight stick and the throttle respectively. Participants did not know in advance whether the landing would be automated or manually executed. Considering the whole spectrum of the landing task, our experimental conditions were designed to be contrasted in terms of mental demands. The landing phase ended 10 s after the pilots touched down on the landing ground. Before starting the experiment, the participants performed a 30-min training session to familiarize themselves with the simulator environment.

2.3. Data Acquisition
2.3.1. Subjective Workload Assessment

After the end of the experiment, the pilots were asked to complete a commonly used subjective workload level questionnaire, the

NASA-TLX (Hart and Staveland, 1988) in order to compare the two conditions. This questionnaire combines 6 factors, i.e., mental demand, physical demand, temporal demand, overall performance, frustration level, and effort.

2.3.2. fNIRS Recording

Two NIRSport acquisition devices (NIRx Medical Technologies) were used in tandem mode to increase the number of sensors. Each system has 8 sources and 8 detectors receiving wavelength at 760 and 850 nm recorded at 7.8125 Hz. By using 2 systems, Frontal and Occipital areas were both covered with 8 sources and 8 detectors constrained mechanically by a plastic spacer at the appropriate distance (3 cm maximum), resulting in 42 optodes or channels. The probabilistic path of photon through cortex were estimated using the Monte-Carlo transport software tMCimg via the Atlas Viewer from Homer2 (Boas et al., 2002; Aasted et al., 2015). The optodes placement and the results of the simulation are shown **Figure 3**. Before starting the experiment a calibration was performed in order to check each optode's signal quality.

2.4. Data Analysis
2.4.1. Pre-processing

FNIRS data were analyzed using Matlab R2015b with several functions from the Homer2 software package (Dubb and Boas, 2016). The overall analysis pipeline is described in **Figure 4**. The landing phase was divided into epochs of 200 samples ($\sim$25 s) overlapping by 60 samples ($\sim$7.5 s). As the landing duration (152 ± 22 s) could slightly differ among participants depending on their performance, the fixed number of extracted epochs was based on the shortest landing in duration, resulting in 12 epochs per landing and per subject.

Each epoch was processed independently in order to potentially extend our method to online processing. Raw data were converted to optical densities; an artifact removal algorithm and a band pass filter were applied on each epoch separately. A wavelet interpolation method was used for the artifact correction (Molavi and Dumont, 2012). This method has been shown to have the greatest signal to noise ratio among the current artifact removal methods available (Brigadoi et al., 2014). A butterworth high pass filter (cutoff: 0.01 Hz - order 3) and a low pass filter (cutoff: 0.5 Hz - order 5) were applied for the band pass filtering step.

The filtered and artifact free data were then converted to oxy-hemoglobin [HbO] and deoxy-hemoglobin [HbR] concentration variations.

For further analysis, only the 80 centered samples ($\sim$10 s) of each epoch were kept by applying a boxcar function. This window was applied to avoid spectral leakage, specifically from the wavelet transform, and to obtain a 10 s window without overlap. At the end of this processing stage, for each landing (trial) we had 12 non-overlapping, filtered and artifact free epochs of 80 samples.
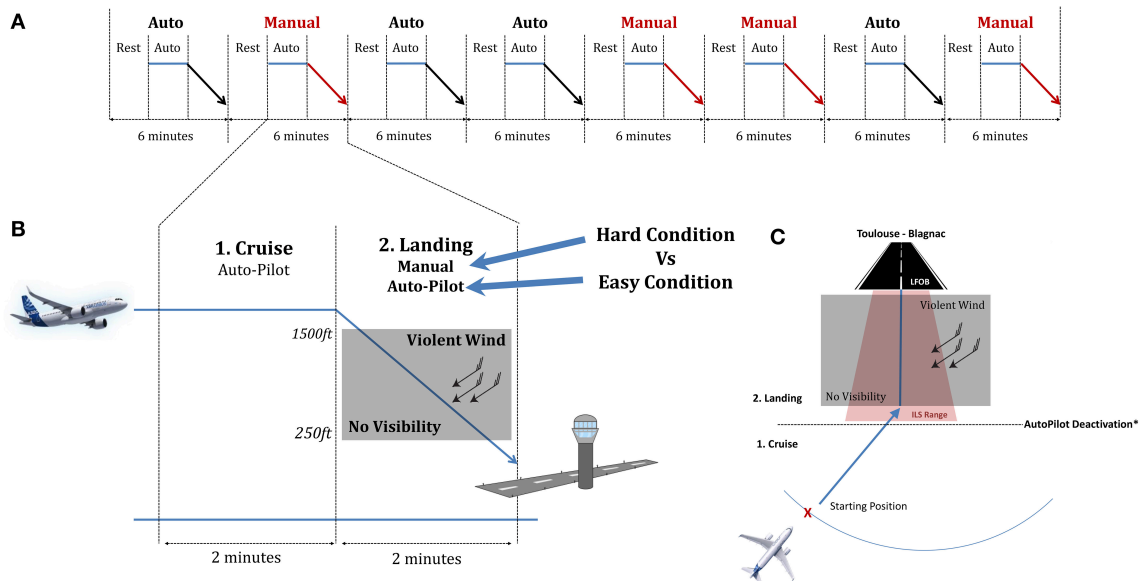
2.4.2. Oxygenation Measures

Oxygenation measures were computed using both the [HbO] and [HbR] signals on each epoch separately, where x represents either the [HbO] and [HbR] signal for one epoch (80 samples) and one optode. Seven oxygenation measures were computed (Peak,

FIGURE 1 | Airbus A320 twin-engine simulator at ISAE-SUPAERO. Pictures used with written consent.

FIGURE 2 | Structure of the experimental session. **(A)** Overall Flow of the experiment with the 8 scenarios. **(B)** Detailed trial for the cruise and landing phases. **(C)** Upper view of the plane trajectory. The starting position was pseudo randomly placed on the blue arc. The blackline delimited the cruise from the landing phase where the pilot deactivate or not the autopilot regarding the condition.

Mean, Variance, Kurtosis, Skewness, Area Under the Curve, and Slope).

The peak (maximum) and the 4th moment (average, variance, skewness, and kurtosis) were computed as follows:

$$Average(x) = E(x) \quad Var(x) = E[(x - E(x))^2] \quad (1)$$

$$Skew(x) = \frac{E[(x - E(x))^3]}{(E[(x - E(x))^2]^{3/2})} \quad Kurt(x) = \frac{E[(x - E(x))^4]}{(E[(x - E(x))^2]^2)} \quad (2)$$

The Area Under the Curve (AUC) was calculated by summing the absolute values of the signal.

$$AUC = \sum |x| \quad (3)$$

The slope was computed using the least-squared linear regression with the polyfit matlab function.

2.4.3. Connectivity Measures

Connectivity measures were computed, as previously, using both the [HbO] and [HbR] signals on each epoch separately, where x and y represents two signals from two different channels. Five oxygenation measures were computed (Covariance, Pearson's correlation, Spearman's correlation, Coherence, and Wavelet Coherence).

Covariance of two signals x and y can be described as a "measure of joint variability":

$$COV(x, y) = E(x - E(x)) \times E(y - E(y)) \quad (4)$$

Where E represents the expected value. Intuitively, covariance characterizes the simultaneous variations of two signals. Covariance will be positive when the differences between the signals and their averages tend to be of the same sign and tend to be negative in the opposite case.
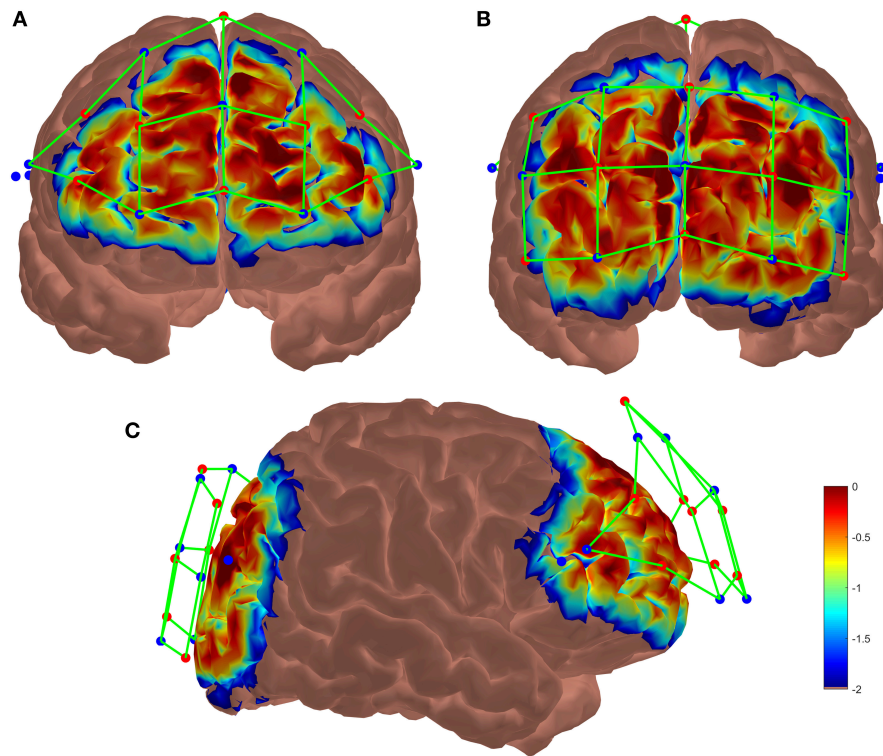
FIGURE 3 | Results of the Monte-Carlo simulation over **(A)** the frontal cortex, **(B)** the occipital cortex and **(C)** both cortices from a lateral view. Red dots represent the LED emitters, blue dots the photoreceptors and green lines the channels. The colorbar unit represents the spatial sensitivity of the fNIRS measurements. It is expressed in mm^{-1} and values range from 0.01 to 1 in log10 units : −2 to 0.

Pearson's correlation coefficient is the covariance of two signals normalized by the product of their standard deviation (std). It represents the linear correlation between two signals, its values ranges from −1 to +1 meaning respectively a linear negative and positive correlation and 0 corresponding to no correlation at all.

$$Pearson(x, y) = \frac{COV(x, y)}{std(x) \times std(y)} \qquad (5)$$

Spearman rank correlation coefficient is "defined as the Pearson correlation coefficient between the ranked variable" (Myers and Arnold, 2003).

$$Spearman(x, y) = \frac{COV(rg_x, rg_Y)}{std(rg_x) \times std(rg_y)} \qquad (6)$$

Where rg_x and rg_y are the ranked variable (of x and y respectively). Using the rank instead of the values allows describing monotonic non-linear relationship between signals where the Pearson's coefficient only characterizes linear relationship.

Spectral Coherence $C_{xy}(f)$ or Magnitude squared Coherence is defined as the absolute squared value of the cross-spectral density of two signals (x and y) for a frequency f, normalized by the

product of their auto-spectral density:

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f) \times G_{yy}(f)} \qquad (7)$$

Where $G_{xy}(f)$ represents the cross spectral density (being the spectrum of the cross correlation function) for a frequency f. $G_{xx}(f)$ and $G_{yy}(f)$ being the auto spectral density (i.e., the spectrum of the auto correlation function) respectively for x and y. Spectral coherence can be seen as a correlation coefficient in the frequency domain.
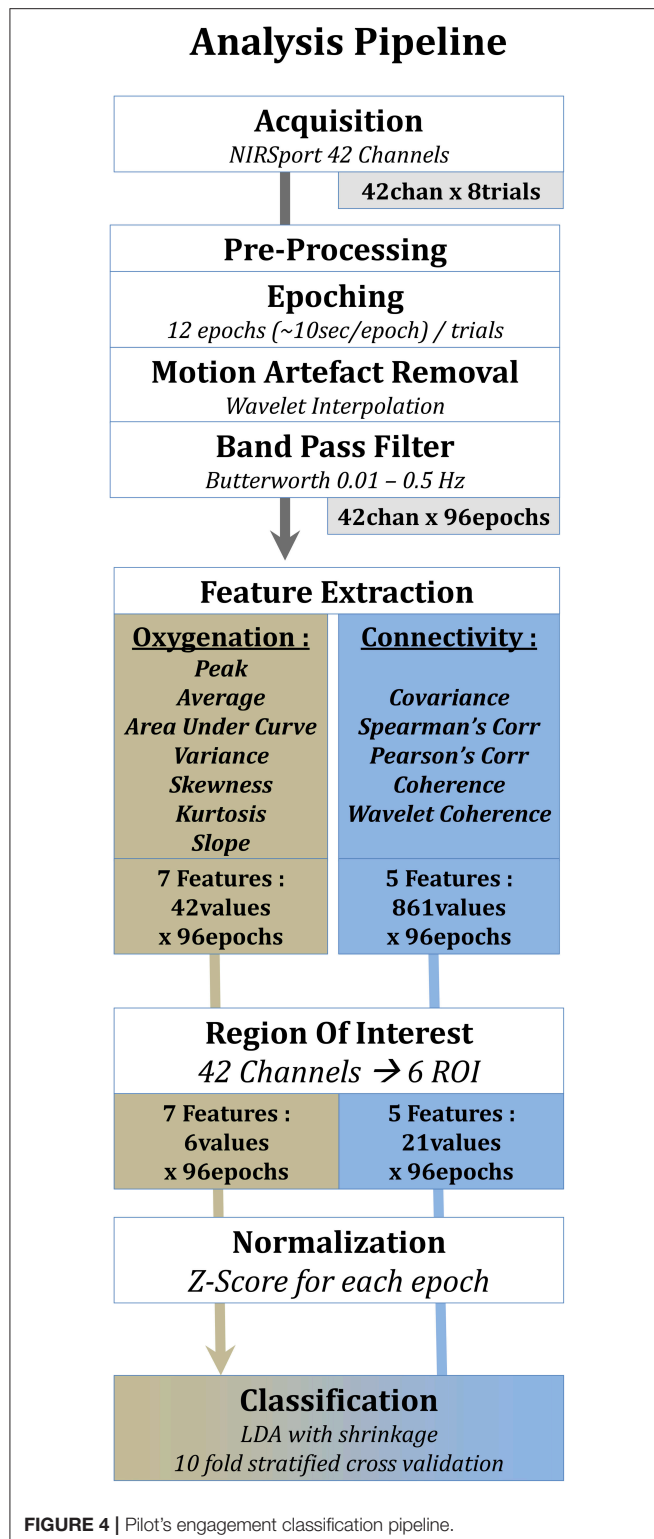
For the last one, a coherence measure based on the wavelet transform was used (Torrence and Compo, 1998): the wavelet coherence. The wavelet coherence power $R_n^2(s)$ can be defined as:

$$R_n^2(s) = \frac{|S(s^{-1} W_n^{xy}(s)|^2}{S\left(s^{-1}|W_n^x(s)|^2\right) S\left(s^{-1}|W_n^y(s)|^2\right)} \qquad (8)$$

Where $W_n^x(s)$ and $W_n^y(s)$ represent respectively the wavelet transform of x and y at the n time point for a wavelet scale s. $W_n^{xy}(s)$ is the cross wavelet transform of x and y (being the wavelet transform of the cross correlation function). S is a smoothing operator (for more detail see Torrence and Compo, 1998).

This measure can be seen as "a localized correlation coefficient in time frequency space" (Grinsted et al., 2004). Coherence values

Analysis Pipeline

Acquisition
NIRSport 42 Channels

42chan x 8trials

Pre-Processing

Epoching
12 epochs (~10sec/epoch) / trials

Motion Artefact Removal
Wavelet Interpolation

Band Pass Filter
Butterworth 0.01 – 0.5 Hz

42chan x 96epochs

Feature Extraction

Oxygenation :	Connectivity :
Peak	
Average	*Covariance*
Area Under Curve	*Spearman's Corr*
Variance	*Pearson's Corr*
Skewness	*Coherence*
Kurtosis	*Wavelet Coherence*
Slope	
7 Features : 42values x 96epochs	**5 Features : 861values x 96epochs**

Region Of Interest
42 Channels → 6 ROI

7 Features : 6values x 96epochs	**5 Features : 21values x 96epochs**

Normalization
Z-Score for each epoch

Classification
LDA with shrinkage
10 fold stratified cross validation

FIGURE 4 | Pilot's engagement classification pipeline.

range from 0 to 1, 1 meaning there is a perfectly phase-locked oscillations at a given frequency for the two analyzed signals.

Connectivity measures were computed on each epoch separately and for each couple of channels namely $C\binom{n}{k} = 861$

couples ($k = 2$, $n = 42$). At this step, we had 42 measures for each oxygenation feature and 861 measures for each connectivity feature per epoch and per subject.

2.5. Data Classification

2.5.1. Feature Extraction

2.5.1.1. Region of interest
In order to reduce the amount of data and the dimensionality, the 42 different channels were combined into 6 regions of interest (ROI): Frontal-Left and Right, Fronto-Central; Occipital-Left and Right and Occipito-Central.

For the oxygenation features, it was done by averaging all the features from channels included in these 6 differents regions. For the connectivity features, 15 possible connections were possible across the 6 ROI. Values were firstly evaluated for each pair (861 couples) and then averaged across couples connecting the same regions. Couples of channels included inside one ROI were also kept, which gave $15 + 6 = 21$ connectivity measures.

At this step, we had 6 measures for each oxygenation feature and 21 measures for each connectivity feature per epoch and per subject.

2.5.1.2. Frequency specific measures
For the two coherence measures (Magnitude Squared Coherence and Wavelet Coherence), the obtained coherence values were averaged for a frequency range between 0.3125 Hz (1/3.2 s) and 0.08 Hz (1/12.8 s) accordingly to the fNIRS literature (Cui et al., 2012).

2.5.1.3. Normalization: z-score
Regarding all features, they were normalized by z-scoring (i.e., transform it to have 0 mean and 1 standard deviation; Toronov et al., 2001; Tsunashima and Yanagisawa, 2009; Sasai et al., 2011).

2.5.2. Classification and Cross-Validation
A Linear Discriminant Analysis (LDA) with regularization of the empirical covariance matrix by shrinkage, also known as "shrinkage method", was used (Friedman, 1989; Blankertz et al., 2011). This method has proved its robustness for BCI and passive BCI (pBCI) application (Roy et al., 2016a,b) but also with fNIRS (Herff et al., 2013; Bauernfeind et al., 2014; Hennrich et al., 2015).

Our paradigm was an intra-subject binary classification. Each subject performed 8 landings (4 of each of the 2 conditions). Data were processed to obtain 12 10 s epochs for each landing which gives $12 \times 8 = 96$ epochs (examples) for each subject. Our model prediction performance was assessed by using a stratified cross validation, which is a good tradeoff between bias and variance estimation (Kohavi and Sommerfield, 1995; Friedman et al., 2001). The classifier was trained with examples that originated from 6 different landings (3 of each of the 2 conditions, i.e., 6*12 = 72 examples) and tried to predict examples from the last 2 landings (1 of each condition, i.e., 2*12 = 24 examples). This method was applied for every combination (16) of landings left out of the training set and the averaged performance was kept.

Regarding the features, 2 types of comparisons were done. Firstly, a single feature comparison where each feature classification performance is assessed separately was performed.

Secondly, features were merged together to evaluate their potential. They were combined 2 by 2 and the classification performance obtained with each couple was assessed.

2.5.3. Statistical Assessment

2.5.3.1. Subjective workload comparison

A paired-sample t-test was performed in order to compare the average overall workload obtained for the 2 conditions among subjects.

2.5.3.2. Classification performance significance

For a 2-class problem like ours, the theoretical chance level for classification is 100/2 = 50 %, but this is only right for an infinite sample number. To assess the significance of our classifier (decoding accuracy) the classification error was modelized by a binomial cumulative distribution (see Combrisson and Jerbi, 2015 for more details):

$$P(Z) = \sum_{i=z}^{n} \binom{n}{k} \times \left(\frac{1}{c}\right)^{i} \times \left(\frac{c-1}{c}\right)^{n-1} \tag{9}$$

Where - P is the probability to predict the correct class at least Z times - n the number of samples - c the number of classes.

The performance of our classification pipeline was assessed by repeating the stratified cross validation 16 times and averaging it. As stated earlier, our classifier was trained with 72 samples and tested on the last 24 samples. By using the cumulative binomial distribution, it sets the 5% significance classification threshold at 58.3%.

2.5.3.3. Classification performance comparison

In order to compare the classification performance for each feature, a repeated measure ANOVA was used considering FEATURES (or FEATURES COUPLE) and CHROMOPHORE (HbO/HbR) as within factors. A *post-hoc* Tukey's Honestly Significant Difference (HSD) procedure was applied to perform multiple comparisons.

3. RESULTS

3.1. Subjective Workload Assessment

Participants rated their workload significantly higher for the manual landing condition ($M = 66.6 \pm 9$) than the automatic landing condition ($M = 18.7 \pm 7$; $t(11) = -17.43$, $p < 10^{-8}$).

3.2. Classification with Individual Features

Figure 5 illustrates the classification performances for each feature computed over the HbR and HbO signals. In order to compare classification performance among features, a repeated measure ANOVA was done.

The statistical analysis showed that there was a significant effect of feature type on classification performance [$F_{(11,121)} = 5.66, p < 10^{-3}$] and it also revealed a significant effect of the chromophore used [$F_{(1,11)} = 8.73, p < 0.05$]. *Post-hoc* comparisons revealed significant differences among features mainly for HbO. In particular, Wavelet Coherence had a significantly better performance than the Average, Skewness,

Kurtosis, and Slope. Also, every connectivity feature gave a significantly greater performance than the Skewness. Moreover, regarding HbR, the Wavelet Coherence and the Covariance gave a significantly greater performance than the Kurtosis. All the connectivity features did not exhibit significant differences between one another. *Post-hoc* comparisons did not show any significant effect of the chromophore on the classification performance regardless of the feature used. In other words, every feature gave non-significant different results when using either the HbO or HbR signals for the classification.

Moreover, every connectivity feature computed over the HbO signals led to an average classification performance above chance level (>58.3 %). Furthermore, Pearson's, Spearman's correlation, and the Wavelet Coherence exceeded the chance level for both HbO and HbR. Concerning classical oxygenation features, the AUC and Variance were the only features to reach a classification performance above chance level but only when computed over HbO signals.

Regarding the best features, Wavelet Coherence benefited of the best classification performance among subjects with an average 65.34 and 59.94% of good classification respectively for HbO and HbR. The second was the Covariance (62.93 and 56.03 %) followed by the Area Under the Curve (61.76 and 57.83%) for HbO and HbR respectively.

3.3. Classification with Combined Features

Figures 6, **7** show the averaged classification performance for all the possible combinations of 2 oxygenation or 2 connectivity features respectively.
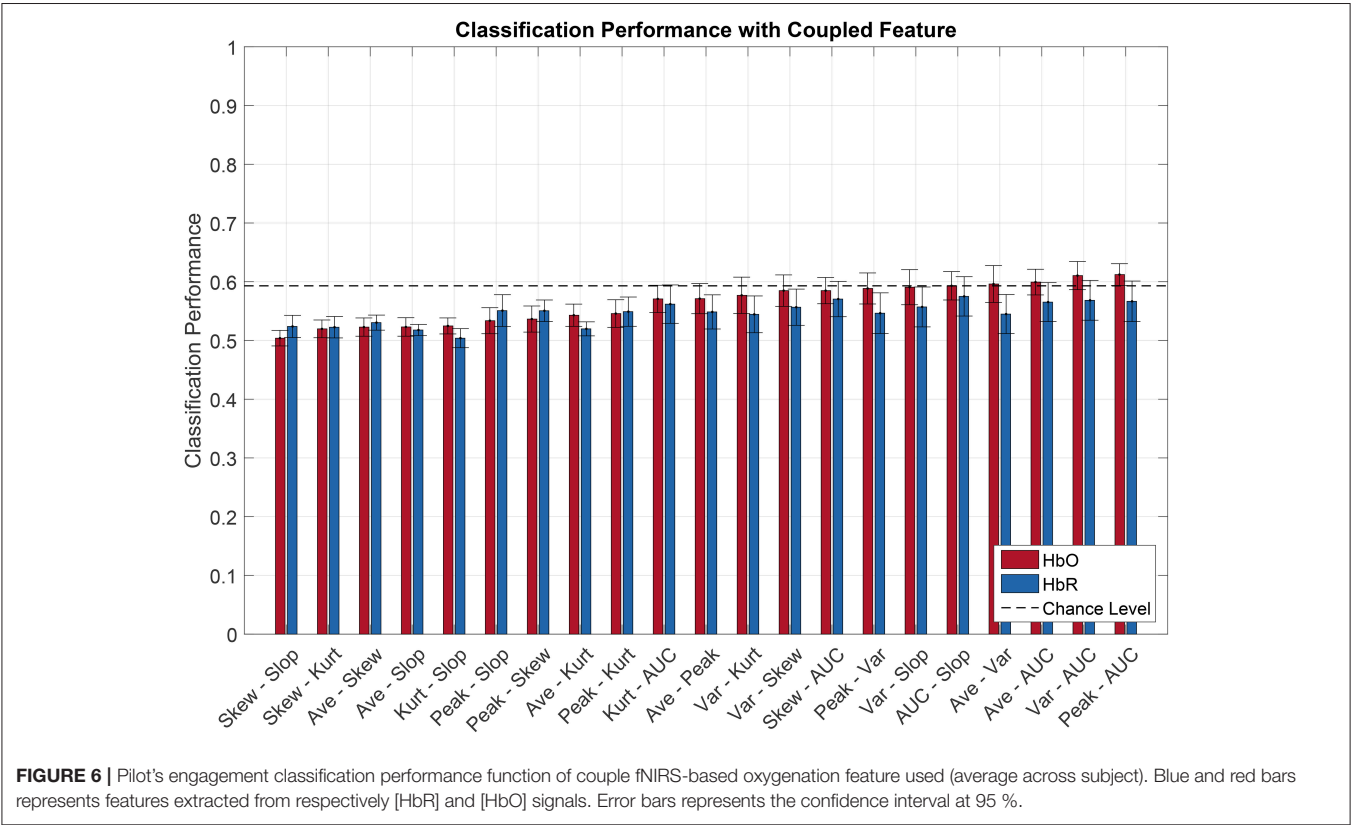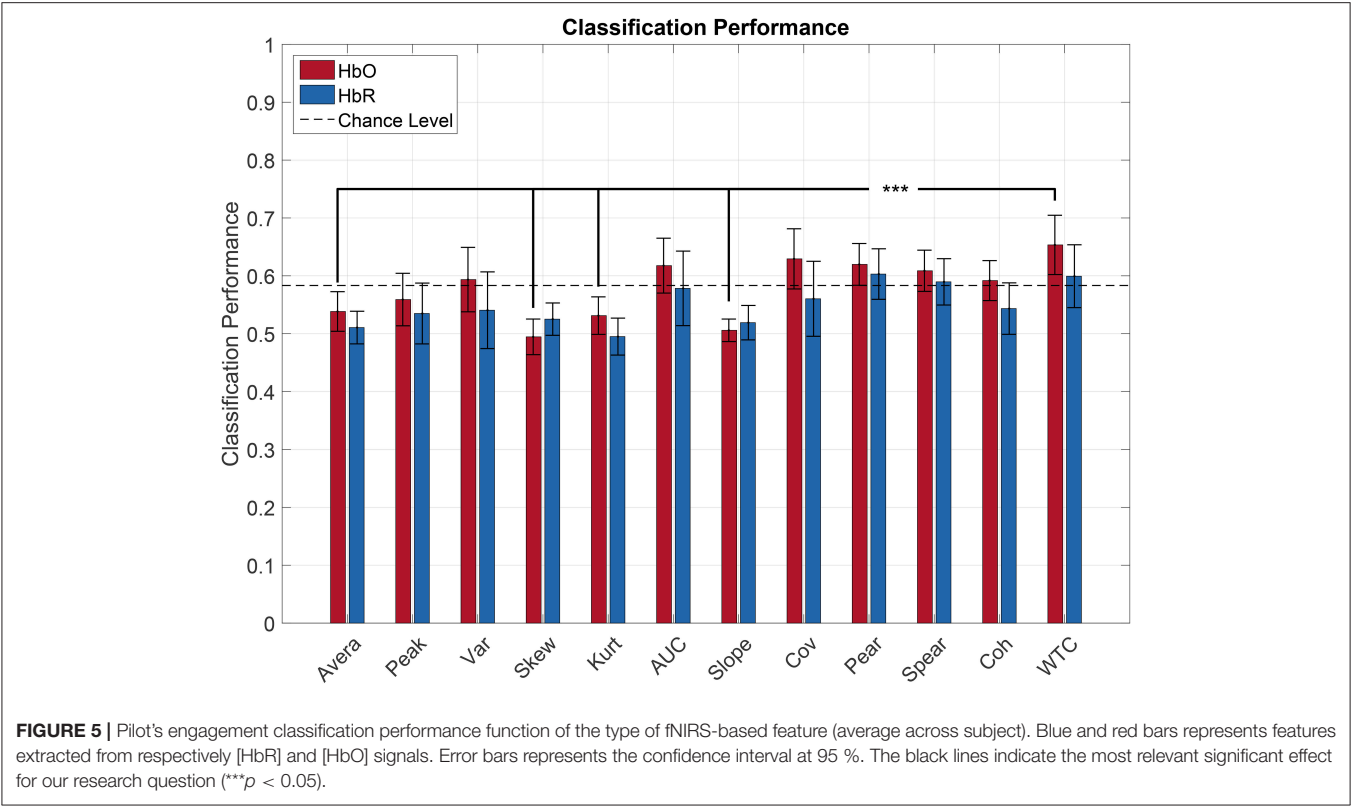
Following the same procedure as before, a repeated measure ANOVA was done with the data showed **Figures 6**, **7**. It revealed that there was a significant effect of the feature couple [$F_{(30, 330)} = 5.42, p < 10^{-3}$] but not of the chromophore [$F_{(1, 11)} = 2.47, p = 0.14$] on the classification performance.

When evaluating multiple comparisons for HbO, the main observation is that the 7 best connectivity couples gave a significantly greater classification performance than the 7 worse oxygenation couples. Besides that, it can also be noted here that connectivity couples did not exhibit significant differences between one another.
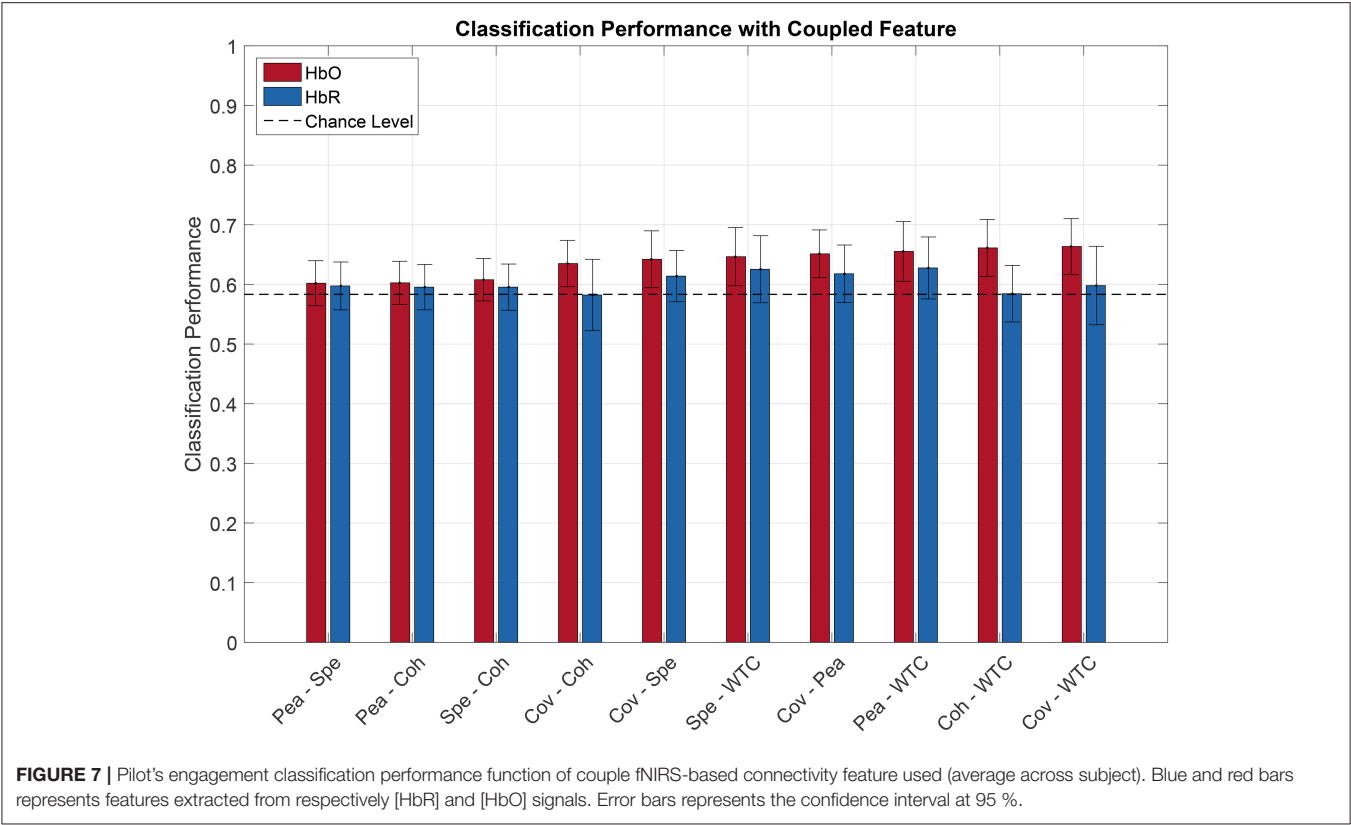
For oxygenation features, 9 out of 21 couples of features led to a classification performance above chance level, namely AUC-Peak, AUC-Variance, AUC-Average, Average-Variance, AUC-Slope, Variance-Slope, Peak-Variance, AUC-Skewness, and Variance-Skew. The AUC-Peak couple reached a classification performance of 61.2 and 56.7% for HbO and HbR respectively. Moreover AUC is in 5 of these 9 best couples. Regarding combined connectivity features, every connectivity couple reached a classification performance above chance except the couple Covariance-Coherence when computed over HbR. The best couple (Covariance-WaveletCoherence) led to a classification performance of 66.4 and 59.8% (for HbO and HbR respectively).

Results for every feature couple, including couples mixing oxygenation and connectivity features, for every subject are given in **Tables 1**, **2**.

FIGURE 5 | Pilot's engagement classification performance function of the type of fNIRS-based feature (average across subject). Blue and red bars represents features extracted from respectively [HbR] and [HbO] signals. Error bars represents the confidence interval at 95 %. The black lines indicate the most relevant significant effect for our research question (***$p < 0.05$).

FIGURE 6 | Pilot's engagement classification performance function of couple fNIRS-based oxygenation feature used (average across subject). Blue and red bars represents features extracted from respectively [HbR] and [HbO] signals. Error bars represents the confidence interval at 95 %.

FIGURE 7 | Pilot's engagement classification performance function of couple fNIRS-based connectivity feature used (average across subject). Blue and red bars represents features extracted from respectively [HbR] and [HbO] signals. Error bars represents the confidence interval at 95 %.

TABLE 1 | Classification performance (HbO/HbR) for every possible combination of 2 features with an average performance across subjects **under chance level** (<58.3 %) computed over HbO.

Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
Skew - Slo	40/55	46/53	46/47	57/56	50/58	58/64	50/57	51/49	51/45	52/51	55/39	49/56	50/52
Skew - Kurt	57/57	45/54	53/48	54/57	48/65	55/51	61/51	55/55	50/48	47/47	55/39	43/56	52/52
Ave - Skew	54/53	58/51	48/51	64/51	48/62	52/63	52/51	51/52	56/49	43/51	56/49	47/54	52/53
Ave - Slo	61/57	59/51	48/47	57/51	52/51	56/59	41/49	52/51	51/53	47/52	51/49	53/52	52/52
Kurt - Slo	60/53	49/49	53/47	47/55	52/55	57/53	60/51	57/53	49/53	46/44	49/36	50/56	52/50
Peak - Slo	56/55	60/52	42/39	55/57	54/47	58/73	43/55	62/70	63/53	47/50	59/63	41/47	53/55
Peak - Skew	51/56	56/53	45/46	62/54	52/56	53/69	53/54	59/62	65/51	47/49	63/62	38/49	54/55
Ave - Kurt	67/50	57/53	54/49	59/54	53/51	52/57	60/47	61/55	50/59	47/43	49/53	43/53	54/52
Peak - Kurt	61/54	61/54	50/45	48/55	55/46	57/65	55/55	65/71	64/64	46/42	57/63	35/47	55/55
Kurt - AUC	69/61	57/63	60/42	52/55	60/45	63/66	60/60	65/81	61/55	53/40	45/62	40/44	57/56
Ave - Peak	65/52	64/59	50/43	62/42	60/49	59/74	49/56	65/71	68/57	46/49	60/62	38/43	57/55
Skew - Coh	66/48	51/50	50/53	65/54	51/64	46/56	69/52	62/47	58/54	54/52	59/69	56/66	57/55
Var - Kurt	67/55	64/55	62/43	41/53	64/46	57/66	64/60	74/82	64/53	52/44	45/57	38/41	58/54
Skew - Spe	58/60	49/57	49/53	68/66	64/77	57/59	66/73	62/46	61/54	48/50	49/48	66/62	58/59

Results are rounded to the closest integers and ordered by their average value (the last couple (row) is the best performing on average across subjects). Columns refer to subjects (S1–S12) and rows to each feature couple (Ave, Average; Pea, Peak; Var, Variance; Skew, Skewness; Kurt, Kurtosis; AUC, Area Under the Curve; Cov, Covariance; Pear, Pearson's; Spear, Spearman's; Coh, Coherence; WTC, Wavelet Coherence).

4. DISCUSSION

The main motivation of the present study was to assess the potential of connectivity measures to classify two different levels of task engagement with fNIRS under relatively ecological settings. We therefore designed a protocol whereby pilots had to perform several manual and automated landings. Our subjective measures confirmed that these two situations were contrasted as manual landing led to significantly higher subjective NASA-TLX scores than automated landing. Our overall classification

TABLE 2 | Classification performance (HbO/HbR) for every possible combination of 2 features with an average performance across subjects **above chance level** (>58.3 %) computed over HbO.

Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
Var - Skew	64/56	60/52	57/46	57/51	60/60	60/70	62/59	78/82	66/45	50/51	51/54	38/42	58/56
Skew - AUC	68/60	54/57	58/46	61/53	59/56	66/71	62/59	69/80	63/50	51/48	49/63	42/41	58/57
Kurt - Coh	66/53	52/45	57/48	58/57	53/59	52/50	69/49	69/53	55/59	55/54	57/68	60/67	59/55
Slo - Coh	66/52	58/41	52/45	61/59	51/61	51/64	71/52	69/52	55/58	54/57	59/68	59/70	59/56
Peak - Var	64/50	65/52	57/40	52/50	64/41	61/72	56/56	73/81	70/55	55/55	53/63	37/41	59/55
Var - Slo	67/58	69/49	57/42	48/56	66/51	69/72	56/59	75/84	63/47	50/54	44/57	43/40	59/56
AUC - Slo	72/60	59/59	60/40	59/64	57/51	67/76	60/62	71/78	64/48	53/46	44/62	46/43	59/58
Skew - Pea	58/60	50/56	48/53	63/66	64/75	58/63	69/76	66/48	68/56	48/52	55/45	67/61	59/59
Ave - Var	66/51	67/53	59/50	59/40	64/48	64/72	57/58	80/82	69/54	47/51	46/56	39/40	60/54
Ave - AUC	70/61	58/60	61/45	61/47	57/44	67/72	60/61	70/81	66/54	53/45	49/63	46/45	60/57
Peak - Coh	65/48	58/49	53/50	66/55	54/57	56/67	70/52	65/58	64/52	54/55	60/71	54/61	60/56
Kurt - Cov	66/57	72/51	64/46	52/53	63/63	63/67	65/53	74/81	64/49	52/48	46/63	39/42	60/56
Pea - Spe	64/58	52/59	55/58	65/60	63/73	59/64	66/72	67/52	65/58	50/54	49/48	66/62	60/60
Pea - Coh	63/57	52/51	51/51	64/64	64/72	54/60	70/68	65/49	64/58	54/59	55/61	68/64	60/60
Ave - Coh	71/45	60/43	58/51	67/53	56/58	53/60	70/49	65/50	59/60	55/57	56/68	58/65	61/55
Spe - Coh	67/57	53/53	52/53	64/65	66/71	57/56	69/68	64/46	64/58	53/60	54/63	67/66	61/60
Skew - Cov	63/61	65/49	60/53	61/47	66/69	65/68	63/52	75/80	68/44	49/50	54/63	42/45	61/57
Kurt - Spe	65/57	51/58	59/56	63/67	66/70	58/56	68/72	72/51	63/56	48/48	54/49	65/60	61/58
Var - AUC	71/62	63/61	66/42	61/54	60/44	67/69	58/60	72/83	66/53	54/47	50/64	43/43	61/57
Peak - AUC	69/60	61/55	60/39	62/57	59/43	65/76	61/57	70/79	69/54	57/49	56/67	46/44	61/57
Peak - Cov	65/52	70/52	63/50	59/50	65/61	69/70	61/47	75/82	68/46	53/56	52/66	37/44	61/56
Kurt - Pea	64/57	55/55	59/56	61/65	68/71	57/59	71/74	73/53	65/55	48/50	54/47	65/60	62/59
Slo - Spe	60/58	58/53	55/51	68/61	68/70	62/66	67/72	67/49	65/55	49/55	51/47	67/61	62/58
Slo - Cov	65/58	71/41	59/52	59/53	70/63	68/71	61/53	76/84	65/45	49/53	50/65	47/43	62/57
Slo - Pea	59/59	60/51	54/49	62/59	69/70	62/70	69/74	68/52	67/55	50/53	54/47	69/61	62/58
Peak - Spe	61/62	58/56	53/55	72/62	66/71	62/66	65/72	71/64	70/55	49/52	56/57	63/55	62/60
Var - Cov	64/55	70/52	63/53	59/45	67/66	70/68	60/52	75/82	71/51	52/52	52/67	44/43	62/57
AUC - Coh	71/53	60/48	57/45	63/56	57/54	62/71	73/54	75/80	63/60	62/57	50/68	58/59	63/59
Ave - Spe	66/62	61/55	52/60	73/61	69/71	63/65	64/71	66/51	66/58	49/52	55/52	65/61	63/60
Ave - Cov	68/51	73/46	64/55	64/39	65/63	67/70	58/50	78/81	68/49	49/52	50/62	46/42	63/55
Var - Coh	70/50	65/44	59/49	61/56	58/60	58/68	72/53	76/80	66/57	58/52	56/66	53/59	63/58
Peak - Pea	63/62	60/55	52/54	67/59	66/70	60/69	68/74	74/65	71/55	51/52	59/56	63/57	63/61
AUC - Cov	69/62	67/54	66/52	64/54	67/63	68/68	60/52	76/82	69/49	54/51	52/70	43/47	63/59
Ave - Pea	66/62	59/52	52/59	70/58	69/71	61/65	69/76	68/52	68/60	51/51	55/51	69/63	63/60
Var - Spe	64/64	62/56	57/53	67/64	66/71	61/66	66/72	83/77	67/55	54/53	49/58	61/53	63/62
Skew - WTC	42/58	53/65	57/65	70/54	81/70	64/74	78/71	63/46	67/49	54/48	60/57	68/64	63/60
AUC - Spe	64/66	57/65	52/53	70/64	67/71	66/67	70/72	78/78	68/56	52/55	50/66	64/54	63/64
Cov - Coh	68/55	72/40	59/53	64/52	62/69	64/66	69/48	76/80	66/55	57/54	53/69	53/55	63/58
AUC - Pea	65/65	59/64	52/53	66/60	65/69	67/70	70/74	79/80	71/57	55/54	52/61	67/56	64/64
Cov - Spe	64/64	69/55	61/55	70/62	67/70	62/67	66/67	85/78	64/52	53/56	50/58	60/53	64/61
Kurt - WTC	52/54	58/65	60/65	67/55	80/64	66/73	76/66	67/46	65/48	56/43	61/56	65/64	64/58
Peak - WTC	53/54	64/67	52/60	66/53	82/65	66/77	76/70	65/62	68/46	55/47	60/59	65/60	64/60
Var - Pea	66/64	65/55	58/53	63/59	66/71	61/68	71/75	83/79	69/58	55/53	52/57	66/55	65/62
Slo - WTC	44/57	62/61	64/63	67/58	83/70	65/76	76/70	65/46	65/52	54/49	63/60	68/66	65/61
Spe - WTC	55/60	54/66	60/62	70/65	85/75	65/74	74/74	66/45	66/56	57/45	57/59	65/71	65/63
Cov - Pea	64/64	68/54	59/54	65/57	66/71	64/70	71/71	82/80	67/54	56/55	53/57	67/54	65/62
Ave - WTC	54/53	64/64	62/64	66/50	84/70	64/74	77/68	63/46	67/51	55/49	62/57	66/64	65/59
Pea - WTC	54/59	57/62	59/61	71/63	85/74	67/77	78/74	66/49	68/57	57/48	59/59	67/71	66/63

(Continued)

Features	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
Coh - WTC	59/50	61/55	60/55	72/56	84/69	65/74	78/63	68/46	67/52	56/50	55/63	68/68	66/58
Cov - WTC	57/55	67/57	61/63	66/50	81/70	67/77	79/67	77/81	68/45	58/43	57/59	59/51	66/60
Var - WTC	59/53	63/63	59/63	67/51	81/66	68/80	78/70	75/83	70/48	59/43	59/57	62/53	67/61
AUC - WTC	64/59	61/64	60/61	70/53	80/63	69/81	77/70	68/78	72/48	59/45	57/64	65/55	67/62

Results are rounded to the closest integers and ordered by their average value (the last couple (row) is the best performing on average across subjects). Columns refer to subjects (S1–S12) and rows to each feature couple (Ave, Average; Pea, Peak; Var, Variance; Skew, Skewness; Kurt, Kurtosis; AUC, Area Under the Curve; Cov, Covariance; Pear, Pearson's; Spear, Spearman's; Coh, Coherence; WTC, Wavelet Coherence).

results confirmed that the two different engagement levels could be discriminated in a flight simulator. This is in line with previous neuroergonomics studies showing that this brain optical imaging technique is well suited for mental state monitoring in ecological situations (Herff et al., 2013; Durantin et al., 2015; Gateau et al., 2015; Foy et al., 2016).

The best classification accuracy reached 66.9 %, a result that does not compare favorably with recent studies at first hand. For instance, Hong et al. (2015) obtained a classification performance of 75.6 % on 10 subjects with a mental motor imagery and mental arithmetic paradigm using average and slope features over chromophore concentration. Holper and Wolf (2011) did a complex vs. simple imaginary movement paradigm with 12 subjects. By combining different features such as the average, variance, skewness and kurtosis computed over HbO and HbR, they reached a performance of 81.3 %. Naseer et al. (2016) obtained a 93 % classification performance with almost similar features to classify mental arithmetic vs. rest on 7 subjects. However, these studies did not consider a continuous but rather an event locked assessment of a specific cognitive activity contrarily to our flying task involved different executive and attentional skills. Interestingly enough and contrary to our results, Khan and Hong (2015) showed that classical oxygenation metrics could yield to a high accuracy (84.9 %) when continuously monitoring drowsiness under ecological settings such as driving in simulated conditions. The comparison with our study remains challenging as the construct of engagement is probably more subtle to be captured. Eventually, the limited number of trials did not allow us to optimize the training of our model to guarantee high classification accuracy.

Interestingly, the connectivity measures led to better classification performance than the classical oxygenation metrics (i.e., chromophore concentration variation). The better performance of the connectivity metrics over classical ones could rely on two main explanations. Firstly, one has to consider that the analysis of task-related concentrations (i.e., hemodynamic response) is time-locked to the event. It has been proposed that these task-related responses induce a small increase (<5%) in neural energy consumption compared to the overall brain energy consumption (Raichle and Mintun, 2006). Thus by focusing only on a localized hemodynamic response, the majority of the brain activity is dismissed. It is now well admitted that cognition relies on the activation of several distributed brain areas rather than single dedicated processing units (Siegel et al.,

2012; Hutchison et al., 2013; van den Heuvel and Sporns, 2013). Thus, the analysis of the interaction between neural networks provides more information on the brain dynamics, especially when concerned with the understanding of complex real-life task (Cui et al., 2012; Leff et al., 2015; İşbilir et al., 2016). Secondly, some relevant studies disclosed that frequency or amplitude correlations among spontaneous LFOs (around 0.1 Hz) are tightly linked to cortical processes (Lowe et al., 1998; Xiong et al., 1999; Obrig et al., 2000, see Siegel et al., 2012 for a review). As a matter of fact, when considering continuous monitoring of the brain activity, where no specific events are expected, connectivity features based on frequency or amplitude coupling can give an insight on the ongoing cognitive processes.

The comparison of the connectivity metrics classification performance revealed that covariance, correlation (Pearson's or Spearman's) and wavelet coherence led to significantly higher classification accuracies than respectively 3, 2, and 4 classical oxygenation metrics. It is interesting to note that the formers present complementary advantages. On one hand, correlation and covariance are straightforward and low cost computational measure to implement. This is of great advantage as long as pBCIs are concerned. On the other hand, the wavelet coherence takes into account both time and phased locked oscillations. While being used for some years by the fMRI community, the wavelet coherence metrics has only recently been applied to fNIRS signal (Cui et al., 2012; İşbilir et al., 2016). Wavelet coherence also allows to target specific and relevant frequency bands such as LFOs as discussed previously. However, the implementation of wavelet coherence based pBCIs remains challenging as this metric requires a high number of wavelet convolutions and the calculation costs could be critical in an online paradigm. One possible promising approach to overcome this issue is to consider dimensionality reduction (Guyon and Elisseeff, 2003). Taken together our findings provide some methodological guidance for the implementation of fNIRS based BCI metrics. To the best of our knowledge, this study is one of the rare to benchmark different fNIRS connectivity metrics and to use them for classification purposes in ecological settings. It paves the way toward online mental state estimation in ecological aeronautical settings, but some challenges still remain.

Despite its potential interests, our paper has several limitations. Firstly, this experiment involved 12 subjects that only performed four trials of each conditions. This limited number of trials relied on a compromise as the participants

would experience fatigue and discomfort if wearing the cap for a long period (around 40 min). Secondly, the choice of the two contrasted conditions (automatic vs. manual) can be discussed since potential confounds such as motor responses could influence our measures. Yet we did not target motor areas therefore the risk is low. However, our motivation is to monitor pilots' brain activity when facing realistic flying conditions. The designing of well contrasted and controlled conditions in ecological environments such as flying remains challenging. This first experiment was meant to set the path to more refined protocols to characterize different tasks with a view to perform crew monitoring as achieved by Toppi et al. (2016). The third limitation is regarding the fNIRS signal analysis. Indeed, fNIRS signal is the result of a global component influenced by skin blood flow and a local neuronal component. Some algorithms based on spatial filtering and principal component analysis such as the one proposed by Zhang et al. (2016) could have been used if the analysis was not done on each epoch separately. Moreover fNIRS signals can also be influenced by other physiological activities such as heartbeats, respiration or changes in blood pressure. It would have been interesting to also record those activities to evaluate how they can correlate with the engagement level. Regarding the paradigm settings and because of these limitations, despite the fact that our classification performance were very high and satisfying, it is not possible to make any claim regarding the underlying neurophysiological processes.

Finally, the performance of the classification pipeline needs to be improved before its implementation in the cockpit as such rate of false negative detection cannot be afforded as it is in such critical systems, even though using multisensory fusion this accuracy level is still usable. A promising way to increase classification performance could be to use a bimodal EEG-fNIRS pBCI (Fazli et al., 2012).

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., et al. (2015). Anatomical guidance for functional near-infrared spectroscopy: atlasviewer tutorial. *Neurophotonics* 2, 020801–020801. doi: 10.1117/1.NPh.2.2.020801

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Bastos, A. M., and Schoffelen, J.-M. (2015). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* 9:175. doi: 10.3389/fnsys.2015.00175

Bauernfeind, G., Steyrl, D., Brunner, C., and Müller-Putz, G. R. (2014). "Single trial classification of fnirs-based brain-computer interface mental arithmetic data: a comparison between different classifiers," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE* (Chicago, IL: IEEE), 2004–2007.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. (2011). Single-trial analysis and classification of erp componentsa tutorial. *Neuroimage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Boas, D. A., Culver, J., Stott, J., and Dunn, A. (2002). Three dimensional monte carlo code for photon migration through complex heterogeneous media including the adult human head. *Opt. Express* 10, 159–170. doi: 10.1364/OE.10.000159

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Brigadoi, S., Ceccherini, L., Cutini, S., Scarpa, F., Scatturin, P., Selb, J., et al. (2014). Motion artifacts in functional near-infrared spectroscopy: a comparison of motion correction techniques applied to real cognitive data. *Neuroimage* 85, 181–191. doi: 10.1016/j.neuroimage.2013.04.082

Combrisson, E., and Jerbi, K. (2015). Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J. Neurosci. Methods* 250, 126–136. doi: 10.1016/j.jneumeth.2015.01.010

Cui, X., Bryant, D. M., and Reiss, A. L. (2012). Nirs-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *Neuroimage* 59, 2430–2437. doi: 10.1016/j.neuroimage.2011.09.003

Cutini, S., and Brigadoi, S. (2014). Unleashing the future potential of functional near-infrared spectroscopy in brain sciences. *J. Neurosci. Methods* 232, 152–156. doi: 10.1016/j.jneumeth.2014.05.024

Dehais, F., Causse, M., Vachon, F., and Tremblay, S. (2012). Cognitive conflict in human–automation interactions: a psychophysiological study. *Appl. Ergon.* 43, 588–595. doi: 10.1016/j.apergo.2011.09.004

Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., and Gateau, T. (2015). "Automation surprise in aviation: real-time solutions," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul: ACM), 2525–2534.

Dehais, F., Roy, R. N., Gateau, T., and Scannella, S. (2016). "Auditory alarm misperception in the cockpit: an eeg study of inattentional deafness," in *International Conference on Augmented Cognition* (Toronto, ON: Springer), 177–187.

Dehais, F., Tessier, C., Christophe, L., and Reuzeau, F. (2010). "The perseveration syndrome in the pilot's activity: guidelines and cognitive countermeasures," in *Human Error, Safety and Systems Development*, eds P. Palanque, J. Vanderdonckt, and M. Winckler (Berlin, Heidelberg: Springer), 68–80. doi: 10.1007/978-3-642-11750-3_6

Dubb, J., and Boas, D. (2016). *Homer 2 Toolbox, v2.1*. Available online at: https://www.nitrc.org/frs/shownotes.php?release_id=3156

Durantin, G., Scannella, S., Gateau, T., Delorme, A., and Dehais, F. (2015). Processing functional near infrared spectroscopy signal with a kalman filter to assess working memory during simulated flight. *Front. Hum. Neurosci.* 9:707. doi: 10.3389/fnhum.2015.00707

Fairclough, S. H. (2008). Fundamentals of physiological computing. *Interact. Comput.* 21, 133–145. doi: 10.1016/j.intcom.2008.10.011

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K.-R., et al. (2012). Enhanced performance by a hybrid nirs–eeg brain computer interface. *Neuroimage* 59, 519–529. doi: 10.1016/j.neuroimage.2011.07.084

Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8:700. doi: 10.1038/nrn2201

Foy, H. J., Runham, P., and Chapman, P. (2016). Prefrontal cortex activation and young driver behaviour: a fnirs study. *PLoS ONE* 11:e0156512. doi: 10.1371/journal.pone.0156512

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Vol 1. Springer series in statistics. New York, NY: Springer.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Stat. Assoc.* 84, 165–175. doi: 10.1080/01621459.1989.10478752

Funane, T., Kiguchi, M., Atsumori, H., Sato, H., Kubota, K., and Koizumi, H. (2011). Synchronous activity of two people's prefrontal cortices during a cooperative task measured by simultaneous near-infrared spectroscopy. *J. Biomed. Opt.* 16:077011. doi: 10.1117/1.3602853.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., and Dehais, F. (2015). Real-time state estimation in a flight simulator using fnirs. *PLoS ONE* 10:e0121279. doi: 10.1371/journal.pone.0121279

Gevins, A. S., Morgan, N. H., Bressler, S. L., Cutillo, B. A., White, R., Illes, J., et al. (1987). Human nueroelectric patterns predict performance accuracy. *Science* 235, 580–586. doi: 10.1126/science.3810158

Greenblatt, R. E., Pflieger, M., and Ossadtchi, A. (2012). Connectivity measures applied to human brain electrophysiological data. *J. Neurosci. Methods* 207, 1–16. doi: 10.1016/j.jneumeth.2012.02.025

Grinsted, A., Moore, J. C., and Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Process. Geophys.* 11, 561–566. doi: 10.5194/npg-11-561-2004

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Machine Learn. Res.* 3, 1157–1182.

Hart, S. G., and Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hennrich, J., Herff, C., Heger, D., and Schultz, T. (2015). "Investigating deep learning for fnirs based bci," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (Milan: IEEE), 2844–2847.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2013). Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935

Holper, L., Scholkmann, F., and Wolf, M. (2012). Between-brain connectivity during imitation measured by fnirs. *Neuroimage* 63, 212–222. doi: 10.1016/j.neuroimage.2012.06.028

Holper, L., and Wolf, M. (2011). Single-trial classification of motor imagery differing in task complexity: a functional near-infrared spectroscopy study. *J. Neuroeng. Rehabil.* 8:34. doi: 10.1186/1743-0003-8-34

Hong, K.-S., Naseer, N., and Kim, Y.-H. (2015). Classification of prefrontal and motor cortex signals for three-class fnirs–bci. *Neurosci. Lett.* 587, 87–92. doi: 10.1016/j.neulet.2014.12.029

Hutchison, R. M., Womelsdorf, T., Allen, E. A., Bandettini, P. A., Calhoun, V. D., Corbetta, M., et al. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* 80, 360–378. doi: 10.1016/j.neuroimage.2013.05.079

İşbilir, E., Çakir, M. P., Cummins, F., and Ayaz, H. (2016) "Investigating brain-brain interactions of a dyad using fnir hyperscanning during joint sentence reading task," in *3rd International Symposium on Brain and cognitive Science* (Istanbul).

Khan, M. J., and Hong, K.-S. (2015). Passive bci based on drowsiness detection: an fnirs study. *Biomed. Opt. Express* 6, 4063–4078. doi: 10.1364/BOE.6.004063

Kohavi, R., and Sommerfield, D. (1995). "Feature subset selection using the wrapper method: overfitting and dynamic search space topology," in *KDD' 95: Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, eds U. Fayyad and R. Uthurusamy (Montréal, QC: AAAI Press), 192–197.

Lachaux, J.-P., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., et al. (2002). Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Clin. Neurophysiol.* 32, 157–174. doi: 10.1016/S0987-7053(02)00301-5

Leff, D. R., James, D. R., Orihuela-Espina, F., Kwok, K.-W., Sun, L. W., Mylonas, G., et al. (2015). The impact of expert visual guidance on trainee visual search strategy, visual attention and motor skills. *Front. Hum. Neurosci.* 9:526. doi: 10.3389/fnhum.2015.00526

Li, G., Baker, S. P., Grabowski, J. G., and Rebok, G. W. (2001). Factors associated with pilot error in aviation crashes. *Aviat. Space Environ. Med.* 72, 52–58.

Lowe, M., Mock, B., and Sorenson, J. (1998). Functional connectivity in single and multislice echoplanar imaging using resting-state fluctuations. *Neuroimage* 7, 119–132. doi: 10.1006/nimg.1997.0315

Lu, C.-M., Zhang, Y.-J., Biswal, B. B., Zang, Y.-F., Peng, D.-L., and Zhu, C.-Z. (2010). Use of fnirs to assess resting state functional connectivity. *J. Neurosci. Methods* 186, 242–249. doi: 10.1016/j.jneumeth.2009.11.010

Mandel, L., and Wolf, E. (1976). Spectral coherence and the concept of cross-spectral purity. *JOSA* 66, 529–535. doi: 10.1364/JOSA.66.000529

Mirelman, A., Maidan, I., Bernad-Elazari, H., Nieuwhof, F., Reelick, M., Giladi, N., et al. (2014). Increased frontal brain activation during walking while dual tasking: an fnirs study in healthy young adults. *J. Neuroeng. Rehabil.* 11:85. doi: 10.1186/1743-0003-11-85

Molavi, B., and Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiol. Meas.* 33:259. doi: 10.1088/0967-3334/33/2/259

Molavi, B., Gervain, J., Dumont, G. A., and Noubari, H. A. (2012). "Functional connectivity analysis of cortical networks in functional near infrared spectroscopy using phase synchronization," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE* (San Diego, CA: IEEE), 5182–5185.

Moro, S. B., Carrieri, M., Avola, D., Brigadoi, S., Lancia, S., Petracca, A., et al. (2016). A novel semi-immersive virtual reality visuo-motor task activates ventrolateral prefrontal cortex: a functional near-infrared spectroscopy study. *J. Neural Eng.* 13:036002. doi: 10.1088/1741-2560/13/3/036002

Mumaw, R. J., Sarter, N., and Wickens, C. D. (2001). "Analysis of pilot's monitoring and performance on an automated flight deck," in *11th International Symposium on Aviation Psychology* (Columbus, OH).

Myers, J. L., and Arnold, D. (2003). *Research Design and Statistical Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Myers, J. L., and Arnold, D. (2016). *Statistical Summary of Commercial Jet Airplane Accidents*. Seattle, WA: Boeing.

Naseer, N., Noori, F. M., Qureshi, N. K., and Hong, K.-S. (2016). Determining optimal feature-combination for lda classification of functional near-infrared spectroscopy signals in brain-computer interface application. *Front. Hum. Neurosci.* 10:237. doi: 10.3389/fnhum.2016.00237

Obrig, H., Neufang, M., Wenzel, R., Kohl, M., Steinbrink, J., Einhäupl, K., et al. (2000). Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults. *Neuroimage* 12, 623–639. doi: 10.1006/nimg.2000.0657

Raichle, M. E., and Mintun, M. A. (2006). Brain work and brain imaging. *Annu. Rev. Neurosci.* 29, 449–476. doi: 10.1146/annurev.neuro.29.051605.112819

Reis, C., Mestre, C., and Canhão, H. (2013). Prevalence of fatigue in a group of airline pilots. *Aviat. Space Environ. Med.* 84, 828–833. doi: 10.3357/ASEM.3548.2013

Rowley, A., Payne, S., Tachtsidis, I., Ebden, M., Whiteley, J., Gavaghan, D., et al. (2006). Synchronization between arterial blood pressure and cerebral oxyhaemoglobin concentration investigated by wavelet cross-correlation. *Physiol. Meas.* 28:161. doi: 10.1088/0967-3334/28/2/005

Roy, R. N., Bonnet, S., Charbonnier, S., and Campagne, A. (2016a). Efficient workload classification based on ignored auditory probes: a proof of concept. *Front. Hum. Neurosci.* 10:519. doi: 10.3389/fnhum.2016.00519

Roy, R. N., Charbonnier, S., Campagne, A., and Bonnet, S. (2016b). Efficient mental workload estimation using task-independent eeg features. *J. Neural Eng.* 13:026019. doi: 10.1088/1741-2560/13/2/026019

Roy, R. N., and Frey, J. (2016). "Neurophysiological markers for passive brain–computer interfaces," in *Brain–Computer Interfaces 1: Foundations and Methods*, eds L. Bougrain, L. Clerc, and F. Lotte (Indianapolis, IN: Wiley Online Library), 85–100.

Sarter, N. B., Woods, D. D., and Billings, C. E. (1997). Automation surprises. *Handb. Hum. Fact. Ergon.* 2, 1926–1943.

Sasai, S., Homae, F., Watanabe, H., and Taga, G. (2011). Frequency-specific functional connectivity in the brain during resting state revealed by nirs. *Neuroimage* 56, 252–257. doi: 10.1016/j.neuroimage.2010.12.075

Siegel, M., Donner, T. H., and Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nat. Rev. Neurosci.* 13:121. doi: 10.1038/nrn3137

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Steptoe, A., and Bostock, S. (2012). *A Survey of Fatigue and Well-Being among Commercial Airline Pilots*. London: UCL Psychobiology Group.

Tai, K., and Chau, T. (2009). Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface. *J. Eeuroengi. Rehabil.* 6:39. doi: 10.1186/1743-0003-6-39

Tessier, C., and Dehais, F. (2012). Authority management and conflict solving in human-machine systems. *AerospaceLab* 4, 1–10.

Tong, Y., and Frederick, B. D. (2010). Time lag dependent multimodal processing of concurrent fmri and near-infrared spectroscopy (nirs) data suggests a global circulatory origin for low-frequency oscillation signals in human brain. *Neuroimage* 53, 553–564. doi: 10.1016/j.neuroimage.2010.06.049

Toppi, J., Borghini, G., Petti, M., He, E. J., De Giusti, V., He, B., et al. (2016). Investigating cooperative behavior in ecological settings: an eeg hyperscanning study. *PLoS ONE* 11:e0154236. doi: 10.1371/journal.pone.0154236

Toronov, V., Webb, A., Choi, J. H., Wolf, M., Michalos, A., Gratton, E., et al. (2001). Investigation of human brain hemodynamics by simultaneous near-infrared spectroscopy and functional magnetic resonance imaging. *Med. Phys.* 28, 521–527. doi: 10.1118/1.1354627

Torrence, C., and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* 79, 61–78.

Tsunashima, H., and Yanagisawa, K. (2009). Measurement of brain function of car driver using functional near-infrared spectroscopy (fnirs). *Comput. Intell. Neurosci.* 2009:164958. doi: 10.1155/2009/164958

van den Heuvel, M. P., and Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* 17, 683–696. doi: 10.1016/j.tics.2013.09.012

Wickens, C. D., and Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *Int. J. Aviat. Psychol.* 19, 182–199. doi: 10.1080/10508410902766549

Wright, N., and McGown, A. (2001). Vigilance on the civil flight deck: incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters. *Ergonomics* 44, 82–106. doi: 10.1080/00140130150203893

Xiong, J., Parsons, L. M., Gao, J.-H., and Fox, P. T. (1999). Interregional connectivity to primary motor cortex revealed using mri resting state images. *Hum. Brain Mapp.* 8, 151–156.

Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Zander, T. O., Kothe, C., Jatzev, S., and Gaertner, M. (2010). "Enhancing human-computer interaction with input from active and passive brain-computer interfaces," in *Brain-Computer Interfaces* (Cham: Springer), 181–199.

Zhang, X., Noah, J. A., and Hirsch, J. (2016). Separation of the global and local components in functional near-infrared spectroscopy signals using principal component spatial filtering. *Neurophotonics* 3:015004. doi: 10.1117/1.NPh.3.1.015004

frontiers
in Human Neuroscience

Do Event-Related Evoked Potentials Reflect Apathy Tendency and Motivation?

Hiroyuki Takayoshi, Keiichi Onoda and Shuhei Yamaguchi*

Department of Neurology, Shimane University, Izumo, Japan

Apathy is a mental state of diminished motivation. Although the reward system as the foundation of the motivation in the human brain has been studied extensively with neuroimaging techniques, the electrophysiological correlates of motivation and apathy have not been fully explored. Thus, in 14 healthy volunteers, we examined whether event-related evoked potentials (ERP) obtained during a simple number discrimination task with/without rewards reflected apathy tendency and a reward-dependent tendency, which were assessed separately using the apathy scale and the temperament and character inventory (TCI). Participants were asked to judge the size of a number, and received feedback based on their performance in each trial. The P3 amplitudes related to the feedback stimuli increased only in the reward condition. Furthermore, the P2 amplitudes related to the negative feedback stimuli in the reward condition had a positive correlation with the reward-dependent tendency in TCI, whereas the P3 amplitudes related to the positive feedback stimuli had a negative correlation with the apathy score. Our result suggests that the P2 and P3 ERPs to reward-related feedback stimuli are modulated in a distinctive manner by the motivational reward dependence and apathy tendency, and thus the current paradigm may be useful for investigating the brain activity associated with motivation.

Keywords: motivation, apathy, event-related potential, P2, P3

INTRODUCTION

Apathy is one of the representative clinical symptoms with reduced motivation. Apathy is defined as diminished motivation that is not attributable to a disorder of consciousness, cognitive impairment, or emotional distress (Marin, 1990), and it is characterized by an absence of will which results in decreased self-initiated behavior (Berrios and Grli, 1995). It is difficult to judge clinically whether there is an absence of will or not, but Stuss et al. defined apathy as a status characterized by a decreased response to external stimuli (Stuss et al., 2000). Apathy is seen frequently in various neuropsychiatric disorders, but its mechanism has not been fully explored. If apathy can be assessed by physiological measures, the exploration of neural basis of motivation may be understood more deeply.

Electrophysiological indices such as the P3 event-related evoked potential (ERP), feedback-related-negativity (FRN), and stimulus preceding negativity (SPN) are objective measures of cognitive processing, which have excellent temporal resolution for neural activities elicited by external and internal events. Several ERP studies have examined the motivational changes caused by monetary gain or loss, and it is known that some ERP components are particularly sensitive to

valence and the size of a reward. One of these components is P3 and increases in its amplitude are associated with the gain of a larger reward (Yeung and Sanfey, 2004; Sato et al., 2005). However, few studies have examined the direct relationship between P3 and apathy. In particular, the P3 amplitude in a visual oddball task decreased in apathetic patients after stroke (Yamagata et al., 2004). A similar result was found in Parkinson's disease based on a visual oddball task (Kaufman et al., 2016). The P3 amplitude also decreased in patients with anhedonia (Dubal et al., 2000) and depression (Foti and Hajcak, 2009). This evidence suggests that P3 may reflect cognitive processes that are sensitive to an apathetic state.

Another component is the FRN, which was discovered as a negative potential generated by feedback stimuli signifying a false response (Takasawa et al., 1990). The FRN was also elicited by feedback signifying monetary loss (Gehring and Willoughby, 2002). The FRN amplitude was higher when immediately preceding feedback represented monetary gain compared with loss (Masaki et al., 2006), thereby indicating that the FRN is affected by the motivation level in a trial base. The FRN is generated in the anterior cingulate cortex (ACC), and dysfunction of the ACC network is associated with apathy (Onoda and Yamaguchi, 2015). This evidence suggests that the FRN might also reflect the degree of apathy.

In addition, the SPN was associated with reward gain in motivation studies, including a task with feedback signals related to performance (Brunia and Damen, 1988). The SPN was studied in a time production task, and it had a larger amplitude in the case with monetary rewards (Bocker et al., 1994). Therefore, it is possible that the SPN also reflects motivation.

The ERP component is known to correlate with personality traits and affective disorder (Gangadhar et al., 1993; Hansenne, 1999). To make a physiological index of apathy, the effect of other motivation-related factors should be considered simultaneously. Here, we focused on reward dependence, novelty seeking, and depression. Reward dependence is characterized by eager to help and please others, persistent, industrious, warmly sympathetic, sentimental, and sensitive to social cues and personal succor but able to delay gratification with the expectation of eventually being rewarded (Cloninger, 1987). These characteristics suggest that reward dependence could be treated as a motivational trait. Novelty seeking is a temperament associated with exploratory activity in response to novel stimulation, impulsive decision making, extravagance in approach to reward cues, quick loss of temper, and avoidance of frustration (Cloninger et al., 1993). Reward dependence and novelty seeking are related with reward system (Krebs et al., 2009). Moreover, novelty seeking is associated with dopamine function (Cloninger et al., 1994) and its polymorphism (Lusher et al., 2001). Novelty seeking may play a role in motivation. On the other hand, depression is associated with anhedonia and loss of motivation through functional impairment of the

mesolimbic dopamine pathway (Martin-Soelch, 2009). Apathy and depression are often confused, and sometimes both could be seen simultaneously, particularly in neurological disorders (Hama et al., 2011). It would be desirable to distinguish apathy and depression to reveal the neural basis. Therefore, we investigated the relationships between the ERPs and not only apathy but also reward dependence, novelty seeking, and depression.

In this study, we developed a new simple task where the P3, FRN, and SPN components were evaluated in a single session, and motivation was modulated by changing a monetary reward. This task paradigm enabled us to examine the relationships among the electrophysiological measures, novelty seeking, reward dependence, depressive state, and apathy tendency.

METHODS

Subjects

Fourteen neurologically healthy adult volunteers (eight males, six females) were recruited. Their mean age was 25.3 years (standard deviation = 4.1, range = 20–35 years). All subjects had normal vision or corrected to normal vision. This study was approved by the Ethics Committee of Shimane University, and was conducted in accordance with the Declaration of Helsinki.

Questionnaires

Participants completed the apathy scale (Starr et al., 1983; Okada et al., 1998), the temperament and character inventory (TCI) (Cloninger et al., 1993; Kijima et al., 1996), and Zung's self-rating depression scale (Zung, 1965). These questionnaires are self-entry style questionnaires. TCI is a 125-item questionnaire regarding personality developed by Cloninger et al. (4 points scale per item). We obtained scores for reward dependence and novelty seeking because both traits are closely related to motivation. A higher score of novelty seeking represents novelty seekers (Cloninger et al., 1993). A higher score of reward dependence represents more motivated state (Kijima et al., 1996). A higher scores of apathy scale and SDS represent more apathetic state and depressive state, respectively. The score of mean and standard deviation for novelty seeking was 47.1 ± 6.9, for reward dependence was $41.8 \pm 4.5(33–49)$, for harm avoidance was $51.6 \pm 5.9(41–63)$, for apathy score was $11.3 \pm 5.5(2–21)$, and for SDS was $36.4 \pm 8.4(20–52)$. There were several correlations among apathy scale, SDS, and temperaments. Apathy scale was positively correlated with harm avoidance and SDS (Supplementary Table 1).

Procedures

We developed an original task to measure SPN, FRN, and P3 in a single experimental session. Participants were asked to perform a number discrimination task. **Figure 1** shows the protocol for the number discrimination task. This task comprised three conditions (reward, non-reward, and control condition). In each trial, a number excluding five was displayed and participants judged whether the number is smaller than five. Participants were asked to press the left button when the number was smaller than five and to press the right button when the number was

---

Abbreviations: ACC, anterior cingulate cortex; ANOVA, analysis of variance; EEG, electroencephalogram; ERP, event-related potentials; FRN, feedback-related negativity; SPN, stimulus preceding negativity; TCI, temperament and character inventory.

larger than five. The feedback stimulus was presented 2.5 s after the response. When participants correctly responded faster than the criterion time, a positive feedback stimulus was presented with a value of +10 to +90 at an interval of 10. In contrast, when they responded correctly but slower than the criterion time, a negative feedback was presented with a value of −10 to −90 at an interval of 10. The feedback value was altered based on the response speed and accuracy in the trial, which faster responses yielded higher values, and vice versa (see **Figure 1B** for details). If the previous response was correct and faster than the criterion, the next criterion was shortened automatically by 10 ms. Inversely, if the previous response was incorrect or too slow, the following criterion was automatically prolonged by 10 ms. In the case of the reward condition, the positive value was represented by a monetary reward and it was added to the total amount of money acquired. Even if the feedback was negative, the total amount of money acquired was not decreased because the expected total reward was manipulated to be positive in the reward condition. In the non-reward condition, the value of the feedback stimulus indicated the response speed, which did not affect the acquisition of money. In the case of the control condition, the value of the feedback stimulus ranging from +90 to −90 at an interval of 10 was assigned randomly regardless of the response speed. The probabilities of positive and negative feedback were manipulated so they were both kept at 50%. When participants made a wrong response, "incorrect response" was presented as text in all conditions. If no response was made for 0.8 s after the presentation of the number, "no response" was presented as text. The duration of the feedback stimulus was 1.0 s. After feedback, the current total monetary reward was displayed for 1.0 s. The stimulus color differed in each condition (reward: yellow; non-reward: green; control: white). The average time of the inter-trial interval was 2.5 s (range: 2.0–3.0 s). The task comprised five sessions and each session included three blocks (one block per condition; **Figure 1C**). Each block included 20 trials. A break for a few minutes was allowed between the sessions. Participants were given an opportunity to practice 20 trials before the actual task. They were instructed to press a button as quickly as possible. The initial time criterion was calculated based on the mean reaction time for correct responses in the practice section for each participant. Participants were told that the positive feedback value would be larger if they pressed the button as quickly as possible and answered correctly, and that the negative feedback value would be larger if they responded slowly even with a correct response. We also told the participants that they could identify the ongoing condition based on the stimulus color.

ERP Data Acquisition and Signal Processing

Participants were seated ∼1 m from a monitor in a shielded room. Electroencephalographic (EEG) data were acquired using a BrainAmp system with 64-channel electrodes (Brain Products, Brain AMP DC, Germany) (**Figure 1D**). EEG signals were recorded continuously with the bandpass set at 0.01–250 Hz and a sampling frequency of 500 Hz. The reference channel

was Cz, and re-referencing was performed offline based on the average of all recording sites. Noise components including ocular movement were removed by independent component analysis. The continuous EEG was segmented into epochs, including 200 ms pre-stimulus and 800 ms post-stimulus for the target, and feedback stimulus with a bandpass filter of 2–16 Hz to analyze the P2, P3, and FRN components. This filter setting was used to detect more prominent FRN and to remove slow drift with low frequency filter (Onoda et al., 2010). P2 and P3 were identified as positive or negative components in latency windows of 100–250, 200–350, and 300–500 ms, respectively. FRN was measured as a negative component in the latency window of 250–400 ms for the subtraction waveform (negative-positive). The peak amplitude and latency for each component were determined in the same window. To analyze the SPN, epochs from 2,000 ms pre-stimulus to 500 ms post-stimulus were extracted from the EEG with a bandpass filter of 0.016–30 Hz. The baseline for the SPN was defined as the time window from −1,500 to −1,000 ms before the feedback stimulus. The mean amplitude of the SPN was measured between 1000 ms pre-stimulus and stimulus onset.

Statistics

Behavioral measures were subjected to repeated one-way analysis of variance (ANOVA) with the condition. The primary analysis models for the amplitude and latency of the ERP components comprised repeated measures ANOVA with two or three factors (channel × condition, or channel × condition × feedback valence). The Greenhouse–Geisser correction was applied to ANOVA. In the *post-hoc* test, the Bonferroni method was employed for multiple comparisons. The statistical significance threshold was set to $p < 0.05$.

RESULTS

According to the behavioral data, the main effect of condition on the reaction time was significant [$F_{(2, 26)} = 7.70$, $\varepsilon = 0.68$, $p = 0.008$, d = 0.37], where the reaction time to targets was faster for the reward condition compared with the non-reward and control conditions ($ps < 0.05$, **Table 1**). There was no significant main effect on the error rate [$F_{(2, 26)} = 0.72$, n.s.]. The mean total monetary reward was 1912 ± 698 yen.

The grand average waveforms are illustrated in **Figure 2**. P2 and P3 were elicited for both the target and feedback stimuli, and SPN appeared to precede the feedback stimuli. These components differed in their amplitude and latency depending on the condition or feedback valence.

The target P2 was the largest at Cz, did not exhibit any significant main effects or interaction in terms of their amplitude and latency ($Fs < 2.4$, for P2). The peak amplitude of target P3, which was largest at Cz and Pz, was mainly affected by the condition [$F_{(2, 26)} = 4.79$, $\varepsilon = 0.75$, $p = 0.03$, d = 0.27], where the amplitude for the reward condition was larger than that for the control condition [$p = 0.04$, **Figure 3A**]. Similar to the amplitude, the latency was also mainly affected significantly by the condition [$F_{(2, 26)} = 4.7$, $\varepsilon = 0.84$, $p = 0.024$, d = 0.27], where the latency was shorter for the reward condition than the control condition ($p = 0.007$). The mean amplitude of SPN was
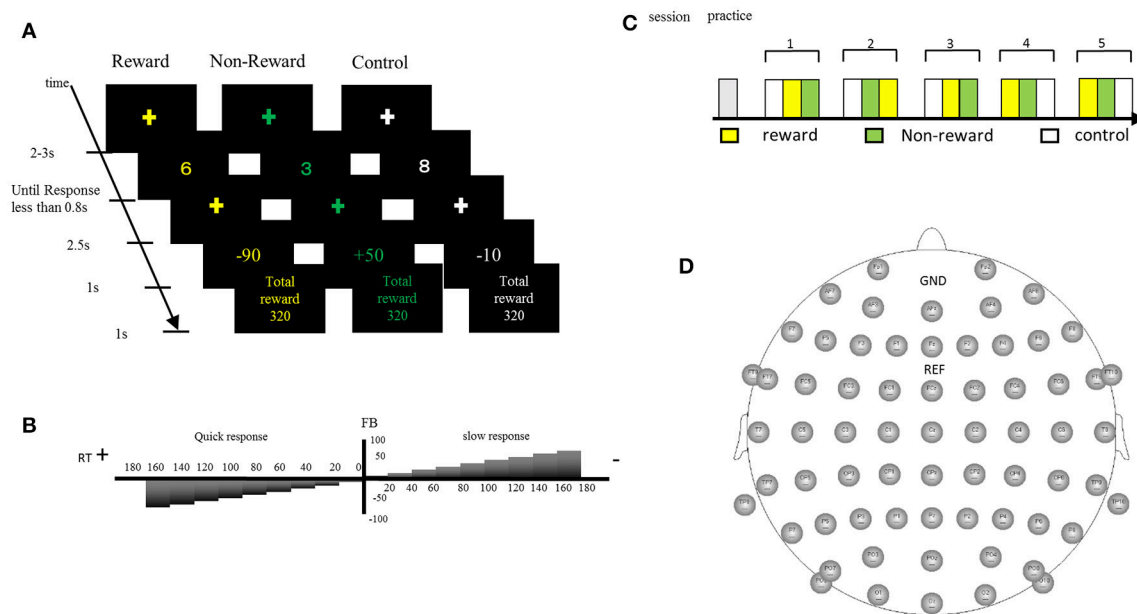
FIGURE 1 | (A) In the number discrimination task, participants pressed the assigned button after judging the size of the number as more than five or less. A monetary reward (reward condition) or reaction speed (non-reward condition) was presented as a feedback stimulus. In the control condition, the feedback was a random number. **(B)** The feedback value increased when the button was pressed faster than a criterion, whereas it became smaller when the button was pressed slower than the criterion. **(C)** Sequence of the stimulus sessions. The task comprised five sessions and each session included three blocks. Participants practiced 20 trials before starting the task. **(D)** Electrode positions.

TABLE 1 | Response time and error rate in rewarded discrimination task.

	Reward	Non-reward	Control	Statistic
RT(ms)	372 ± 35	388 ± 28	395 ± 28	R<N, C
Error rate	2.2%	1.9%	1.75%	n.s.

R, reward; N, non-reward; C, control P < 0.05; Statistics, repeated ANNOVA.

not significantly influenced by the main effects or interactions ($Fs < 2.0$, **Figure 3B**).

The ERPs for the feedback stimulus were analyzed using three-way ANOVA. The feedback P2 had a significant main effect of channel [$F_{(2, 26)} = 5.54$, $\varepsilon = 0.64$, $p = 0.01$], where the amplitude was largest at Cz (**Figure 3C**). Regarding for the FRN, there were neither main effects nor interactions ($Fs < 1.3$, **Figure 3D**). The feedback P3 had a significant main effect of condition [$F_{(2, 26)} = 52.9$, $\varepsilon = 0.84$, $p < 0.001$, d = 0.80], where the largest amplitude was in the reward condition and the smallest amplitude in the control condition ($ps < 0.05$, **Figure 3E**). The interactions between channel × valence/condition were also significant ($Fs > 3.45$, $\varepsilon = 0.70/0.87$, $ps < 0.04$, ds > 0.21). The *post-hoc* test showed that the P3 amplitude at Fz was larger for the negative feedback compared with the positive feedback ($p < 0.05$). Feedback P3 amplitude was the largest at Pz in the reward condition. There was a main effect of valence on latency, where the P3 peak latency was shorter for the positive feedback than the negative feedback [$F_{(1, 13)} = 6.07$, $p = 0.03$, d = 0.32].

Next, we examined the correlations between the ERP components, and the individual psychological and affective

characteristics (**Figure 4**, Supplementary Tables 1, 2). To reduce the flood of information about ERP measures in supplementary table, we averaged ERP measures across three conditions. Target P3 amplitude showed negative correlation with reaction time ($ps < 0.05$). The P2 amplitude at Cz had positive correlations with reward dependence for negative feedback in the reward condition, for positive feedback in the non-reward condition, and for positive feedback in the control condition ($ps < 0.05$). Furthermore, the P3 amplitude at Pz had negative correlations with apathy scale for positive feedback in the reward condition, for positive and negative feedback in the non-reward condition, and for positive feedback in the control condition ($ps < 0.02$).

DISCUSSION

In this study, we examined whether ERP components can be employed as objective measures of apathy and motivation by using a newly developed number discrimination task with or without rewards. According to the behavioral analysis, the reaction time to targets was faster in the reward condition than in the non-reward and control conditions, thereby indicating that the participants were relatively motivated by the monetary reward. We found larger ERP components for the target and feedback stimuli in the reward condition compared with other conditions, which suggests that increased neural activities are associated with enhanced motivation.

We demonstrated that the feedback P2 amplitude was positively correlated with reward dependence, and the feedback P3 amplitude was negatively correlated with the apathy score.
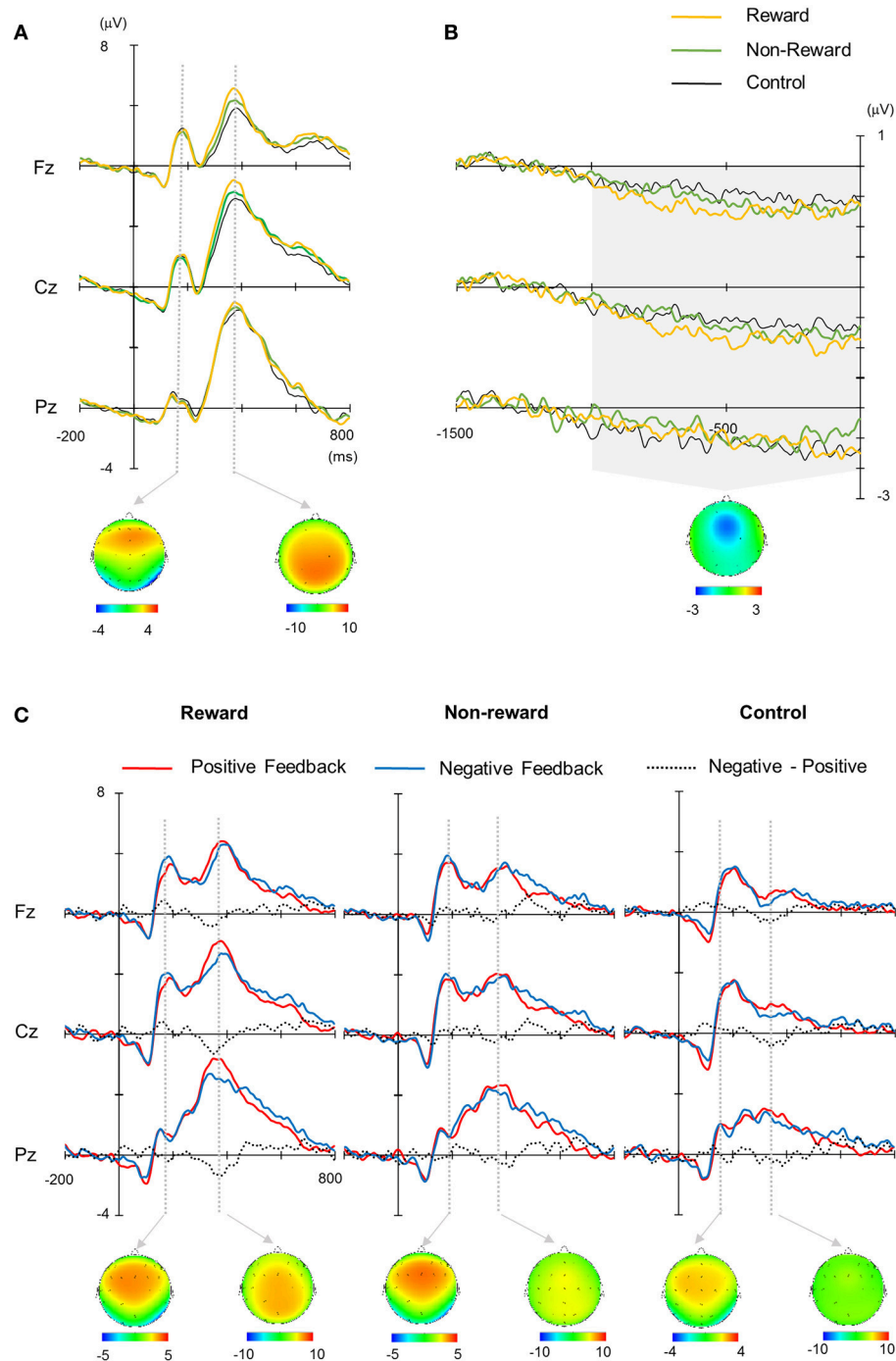
FIGURE 2 | (A) Grand average waveforms for target stimuli. **(B)** Stimulus preceding negativity (SPN). **(C)** Grand average waveforms for feedback stimuli. Each topography was obtained across conditions (mean average of the reward, non-reward, and control conditions for target stimuli **(A)**, positive and negative conditions for feedback stimuli **(C)** at the peak of the grand average waveform. The latency for depicting each topography was 188 ms for target P2, 372 ms for target P3, 188 ms for feedback P2, and 364 ms for feedback P3. The topography of SPN was made from the mean amplitude between 1,000 ms pre-stimulus and stimulus onset **(B)**.

These results imply that the feedback P2 and P3 reflected the motivation. Other ERP components, i.e., SPN and FRN, had no significant relationships with the motivational measures.

The feedback P2 was clearly elicited in all conditions in this study. P2 is considered to be a stimulus-dependent component related to an early stage of information processing (Portella et al., 2012). Potts et al. reported that the frontal P2 was the largest when the reward was unpredictable and the generator was medial frontal cortex associated with reward system (Potts et al., 2006). This evidence indicates that a larger P2 is often observed when
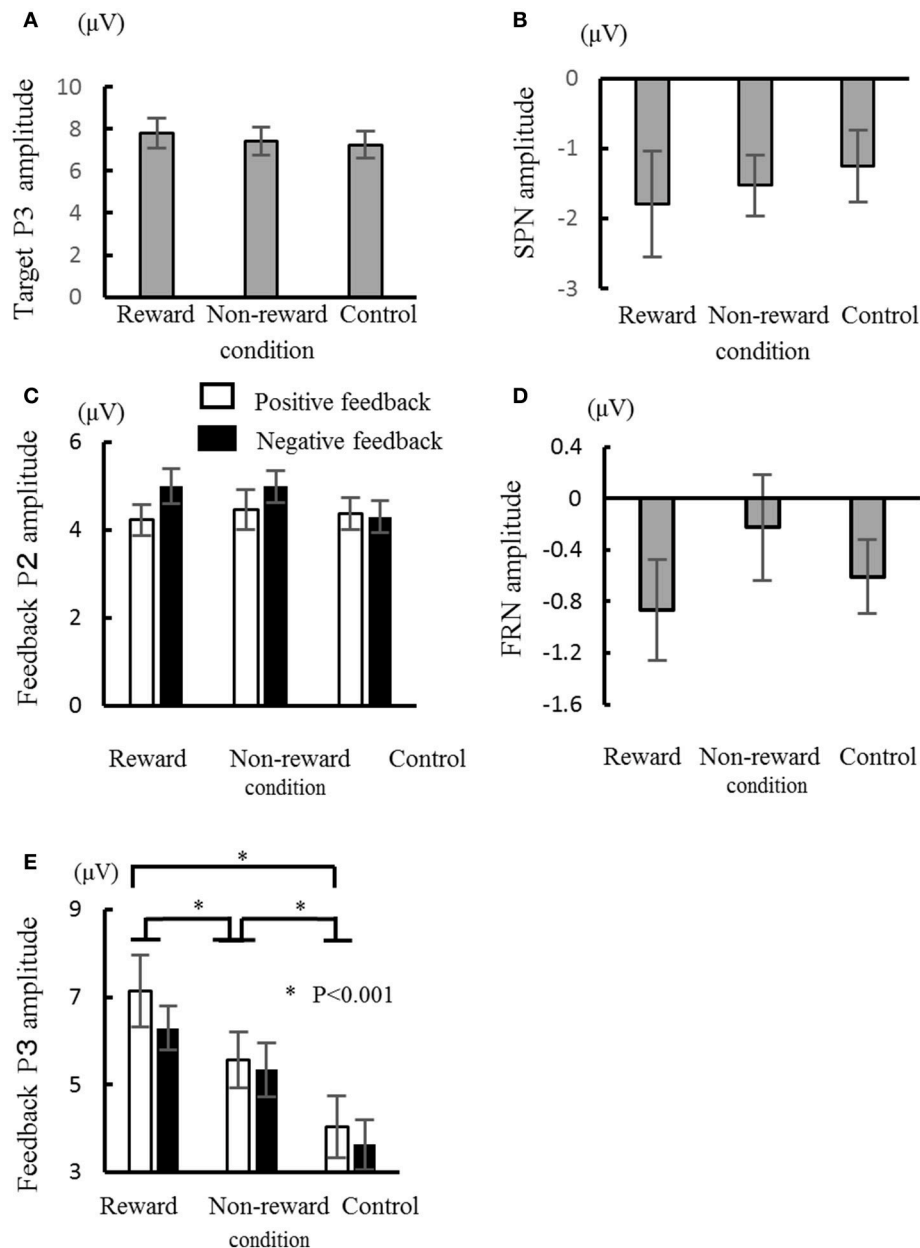
FIGURE 3 | Comparisons of the ERP components in the three conditions: target P3 amplitude **(A)**, stimulus preceding negativity (SPN) amplitude **(B)**, feedback P2 amplitude **(C)**, feedback-related negativity (FRN) amplitude **(D)**, and feedback P3 amplitude **(E)**. Error bars represent the standard error. *$P < 0.001$.

attention is preferentially allocated to a particular stimulus, such as an imperative stimulus or performance feedback (Lackner et al., 2014). In this study, feedback P2 was correlated with reward dependence. Our result suggests that P2 amplitude increases through higher attention based on higher reward dependence. Moreover, close relationships between affective state/personality trait and the P2 component has also been reported. Regarding affective state, higher P2 amplitude was seen in shy adolescents, in individuals with anxiety disorder, and individuals with depression (Kemp et al., 2010; Han et al., 2014; Lackner et al., 2014). Affective state influenced higher attention and is explained

by attention bias (Han et al., 2014) and disruption of selective attention (Kemp et al., 2010). In this study, depressive state was not associated with P2. This may be because the task does not cause affective process markedly and the degree of depressive state was mild. On the other hand, there are several studies regarding the association of reward system and reward dependence. Reward dependence was correlated with gray matter volumes in the caudate nucleus (Iidaka et al., 2006), orbitofrontal cortex, and temporal lobe (Van Schuerbeek et al., 2011); BOLD activity of substantia nigra/ventral tegmental area (Krebs et al., 2009); and opioid receptor availability in striatum and nucleus
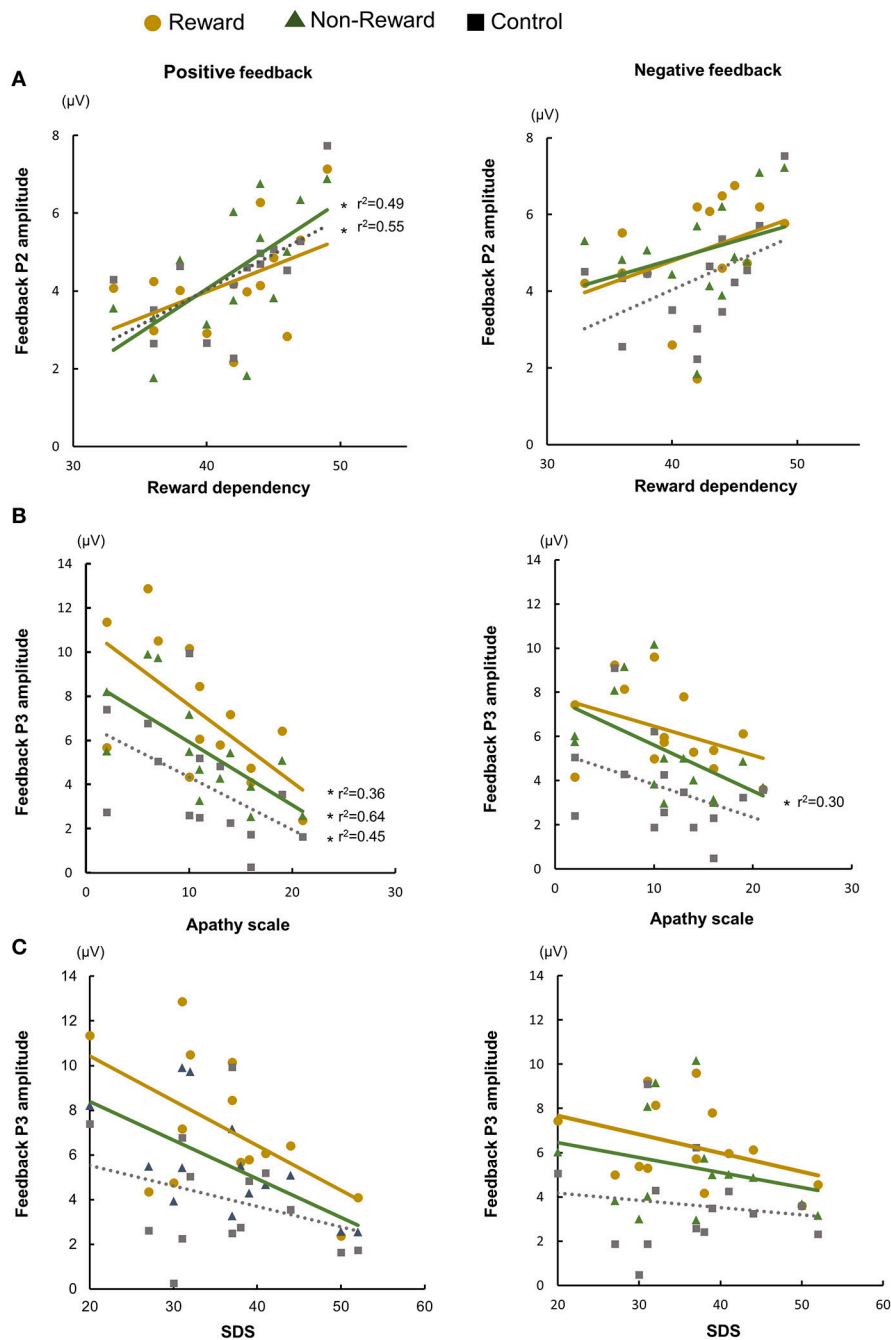
FIGURE 4 | Relationship between temperaments/individual affective state and ERP amplitude: feedback P2 and reward dependency **(A)**, feedback P3 and apathy scale **(B)**, and feedback P3 and Zung's self-rating depression (SDS) scale **(C)**.

accumbens (Schreckenberger et al., 2008). These results indicate that reward dependence is associated with the reward system based on the fronto–striatal circuit. The fronto–striatal circuit may modulate P2 activity via attentional deployment.

In addition, we examined whether the P3 component is modulated by individual temperament and affective state. We found a negative correlation between the feedback P3 amplitude and score of apathy scale. P3 component is usually separated

into P3a and P3b (Snyder and Hillyard, 1976). P3a is elicited by novelty or salient stimuli, for example, in an oddball task (Courchesne et al., 1975; Knight, 1984), and distributed over the fronto-central area (Conroy and Polich, 2007), suggesting its association with the frontal attention system. P3b is elicited by target stimuli in an oddball task. This component is generated partly from temporo–parietal junction (Conroy and Polich, 2007) and relates to attention and memory processing. P3 seen in a

gambling task is related to motivational salience in feedback processing (Nieuwenhuis et al., 2005; Yeung et al., 2005). The feedback P3 amplitude changes depending on reward expectancy and size and the feedback value (Wu and Zhou, 2009). We consider target and feedback P3 as P3b because of the task demands and the topography. Target P3 is associated with target evaluation, feedback anticipation, and encoding contextual valence. On the other hand, feedback P3 is enhanced for the outcome with large value compared to small value and is involved in the late stage of outcome processing for motivational salience rather than contextual valence (Zheng et al., 2017). In our study, target P3 amplitude was increased in reward condition and showed correlation with reaction time but was not correlated with temperament or affective state. Referring to the study of Zheng et al. feedback P3 is related to outcome evaluation for motivational salience, and our results support their notion. We speculate that feedback P3 could be a physiological marker as motivational state.

Several studies have investigated the association between emotion/affection and the P3 component, where they demonstrated that the P3 amplitude decreased in individuals with anhedonia (Dubal et al., 2000) and depression (Foti and Hajcak, 2009; Mathis et al., 2014), which are often accompanied by apathy. In our study, we found no significant correlation between depressive state and the P3 amplitude, thereby suggesting that the P3 component may reflect apathy more directly rather than depression. Similar results were obtained for Parkinson's disease (Mathis et al., 2014), Alzheimer's disease (Daffner et al., 2001), and head trauma (Daffner et al., 2000), where these studies measured the ERP using a visual or auditory oddball task. The P3a arising mainly from the prefrontal area was also correlated with apathy in subcortical stroke patients (Yamagata et al., 2004).

Previous studies have suggested that the SPN and FRN are associated with reward expectation (Bocker et al., 1994; Pfabigan et al., 2011). The SPN amplitude depends on the amount of information with an affective or motivational value carried by the feedback stimulus (Bocker et al., 1994). The FRN is sensitive to unexpected negative feedback but also to unexpected positive feedback, which suggests that the FRN reflects expectancy and the valence of feedback. However, meaningful results were not obtained in the SPN and FRN in the current study. After the participants pressed a button, they made a prediction regarding the outcome, which would have been informed by the feedback received. The probabilities of positive and negative feedback were each fixed at 50% in this study. The probability could have influenced their surprising or disappointing reaction to feedback. It is possible that no significant changes were found in the SPN and FRN because the anticipation and expectation of the outcome were attenuated by the uncertainty of the feedback stimuli.

There was some limitations in our study. Firstly, it was conducted with healthy volunteers; therefore, the degree of apathy was mild even if they were apathetic. Severe apathy is characterized by decreased mental or behavioral reactions; therefore, although the current task was simple and easy to perform, a task that requires responses might not be suitable for studying severe apathy. Thus, we cannot be certain that similar results would be obtained in subjects with severe apathy. Secondly, the number of participants was not adequate for the correlation analysis between subjective measures and ERPs. High reliability for TCI was obtained in the English (Cloninger et al., 1994) and Japanese versions (Takeuchi et al., 2011). We also confirmed the reliability and validity of the apathy scale (Okada et al., 1998). Moreover, the stability of P2 and P3 (Thigpen et al., 2017) is known and high correlation was reported in the test-retest (McEvoy et al., 2000; Williams et al., 2005). Therefore, because there is robustness in these indicators, the results of correlation study seem acceptable even though the number of participants is not adequate for the analysis. Thirdly, there were several correlations between temperaments and states. Therefore, it is difficult to judge whether temperaments affect ERP components or individual affective states. Further studies are necessary to validate our findings before the clinical use of this method. It is desirable to generate tasks that can evaluate intrinsic, extrinsic, and novel motivation to clarify the neural basis of motivation.

In summary, the P2 and P3 may have distinct associations with motivation, where P2 reflects attention that is modulated by motivation and P3 reflects apathy more directly. The current stimulus paradigm may be useful for investigating the brain activity associated with apathy.

ETHICS STATEMENT

This study was approved by the Ethics Committee of Shimane University, and was conducted in accordance with the Declaration of Helsinki. All participants gave written informed consent.

AUTHOR CONTRIBUTIONS

HT, KO, and SY make substantial contributions to conception and design, and acquisition of data, and analysis and interpretation of data and also participate in drafting the article or revising it critically for important intellectual content. Every authors give final approval of the version to be submitted and any revised version. Every authors give agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum. 2018.00011/full#supplementary-material

REFERENCES

Berrios, G. E., and Grli, M. (1995). Abulia and impulsiveness revisited: a conceptual history. *Acta Psychiatr. Scand.* 92, 161–167. doi: 10.1111/j.1600-0447.1995.tb09561.x

Böcker, K. B., Brunia, C. H., and van den Berg-Lenssen, M. M. (1994). A spatiotemporal dipole model of the stimulus preceding negativity (SPN) prior to feedback stimuli. *Brain Topogr.* 7, 71–88. doi: 10.1007/BF01184839

Brunia, C. H., and Damen, E. J. (1988). Distribution of slow brain potentials related to motor preparation and stimulus anticipation in a time estimation task. *Electroencephalogr. Clin. Neurophysiol.* 69, 234–243. doi: 10.1016/0013-4694(88)90132-0

Cloninger, C. R. (1987). A systematic method for clinical description and classification of personality variants: a proposal. *Arch. Gen. Psychiatry* 44, 573–588. doi: 10.1001/archpsyc.1987.01800180093014

Cloninger, C. R., Svrakic, D. M., and Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Arch. Gen. Psychiatry* 50, 975–990. doi: 10.1001/archpsyc.1993.01820240059008

Cloninger, C. R., Thomas, R. P., and Dragan, M. S. (1994). *The Temperament and Character Inventory* (TCI): *A Guide to Its Development and Use*. St. Louis, MO: Washington University.

Conroy, M. A., and Polich, J. (2007). Normative variation of P3a and P3b from a large sample. *J. Psychophysiol.* 21, 22–32. doi: 10.1027/0269-8803.21.1.22

Courchesne, E., Hillyard, S. A., and Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalogr. Clin. Neurophysiol.* 39, 131–143. doi: 10.1016/0013-4694(75)90003-6

Daffner, K. R., Mesulam, M. M., Scinto, L. F. M., Acar, D., Calvo, V., Faust, R., et al. (2000). The central role of the prefrontal cortex in directing attention to novel events. *Brain* 123, 927–939. doi: 10.1093/brain/123.5.927

Daffner, K. R., Rentz, D. M., Scinto, L. F. M., Faust, R., Budson, A. E., and Holcomb, P. J. (2001). Pathophysiology underlying diminished attention to novel events in patients with early AD. *Neurology* 56, 1377–1383. doi: 10.1212/WNL.56.10.1377

Dubal, S., Pierson, A., and Jouvent, R. (2000). Focused attention in anhedonia: a P3 study. *Psychophysiology* 37, 711–714. doi: 10.1111/1469-8986.3750711

Foti, D., and Hajcak, G. (2009). Depression and reduced sensitivity to non-rewards versus rewards: evidence from event-related potentials. *Biol. Psychol.* 81, 1–8. doi: 10.1016/j.biopsycho.2008.12.004

Gangadhar, B. N., Ancy, J., Janakiranaiah, N., and Umapathy, C. (1993). P300 amplitude in non-bipolar, melancholic depression. *J. Affect. Disord.* 28, 57–60. doi: 10.1016/0165-0327(93)90077-W

Gehring, W. J., and Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science* 295, 2279–2282. doi: 10.1126/science.1066893

Hama, S., Yamashita, H., Yamawaki, S., and Kurisu, K. (2011). Post-stroke depression and apathy: interactions between functional recovery, lesion location, and emotional response. *Psychogeriatrics* 11, 68–76. doi: 10.1111/j.1479-8301.2011.00358.x

Han, H. Y., Gan, T., Li, P., Li, Z. J., Guo, M., and Yao, S. M. (2014). Attentional bias modulation by reappraisal in patients with generalized anxiety disorder: an event-related potential study. *Brazil. J. Med. Biol. Res.* 47, 576–583. doi: 10.1590/1414-431X20143622

Hansenne, M. (1999). P300 and personality: an investigation with the Cloninger's model. *Biol. Psychol.* 50, 143-55.

Iidaka, T., Matsumoto, A., Ozaki, N., Suzuki, T., Iwata, N., Yamamoto, Y., et al. (2006). Volume of left amygdala subregion predicted temperamental trait of harm avoidance in female young subjects. A voxel-based morphometry study. *Brain Res.* 1125, 85-93. doi: 10.1016/j.brainres.2006.09.015

Kaufman, D. A. S., Bowers, D., Okun, M. S., Van Patten, R., and Perlstein, W. M. (2016). Apathy, novelty processing, and the P3 potential in Parkinson's disease. *Front. Neurol.* 7:95. doi: 10.3389/fneur.2016.00095

Kemp, A. H., Benito, L. P., Quintana, D. S., Clark, C. R., McFarlane, A., Mayur, P., et al. (2010). Impact of depression heterogeneity on attention: an auditory oddball event related potential study. *J. Affect. Disord.* 123, 202–207. doi: 10.1016/j.jad.2009.08.010

Kijima, N., Saito, R., Takeuchi, M., Yoshino, A., Ono, Y., Kato, M., et.al. (1996). Cloninger's seven-factor model of temperament and character and Japanese version of Temperament and Character Inventory (TCI). *Arch. Psychiatr. Diagnosis. clin. Eval.* 7, 379–399.

Knight, R. T. (1984). Decreased response to novel stimuli after prefrontal lesions in man. *Electroencephalogr. Clin. Neurophysiol.* 59, 9–20. doi: 10.1016/0168-5597(84)90016-9

Krebs, R. M., Schott, B. H., and Düzel, E. (2009). Personality traits are differentially associated with patterns of reward and novelty processing in the human substantia nigra/ventral tegmental area. *Biol. Psychiatry* 65, 103–110. doi: 10.1016/j.biopsych.2008.08.019

Lackner, C. L., Santesso, D. L., Dywan, J., Wade, T. J., and Segalowitz, S. J. (2014). Event-related potentials elicited to performance feedback in high-shy and low-shy adolescents. *Infant Child Dev.* 23, 283–294. doi: 10.1002/icd.1865

Lusher, J. M., Chandler, C., and Ball, D. (2001). Dopamine D4 receptor gene (DRD4) is associated with Novelty Seeking (NS) and substance abuse: the saga continues. *Mol. Psychiatry* 6, 497–499. doi: 10.1038/sj.mp.4000918

Marin, R. S. (1990). Differential-diagnosis and classification of apathy. *Am. J. Psychiatry* 147, 22–30. doi: 10.1176/ajp.147.1.22

Martin-Soelch, C. (2009). Is depression associated with dysfunction of the central reward system? *Biochem. Soc. Trans.* 37(Pt 1), 313–317. doi: 10.1042/BST0370313

Masaki, H., Takeuchi, S., Gehring, W. J., Takasawa, N., and Yamazaki, K. (2006). Affective-motivational influences on feedback-related ERPs in a gambling task. *Brain Res.* 1105, 110–121. doi: 10.1016/j.brainres.2006.01.022

Mathis, S., Neau, J.-P., Pluchon, C., Fargeau, M.-N., Karolewicz, S., Iljicsov, A., et al. (2014). Apathy in Parkinson's disease: an electrophysiological study. *Neurol. Res. Int.* 2014:290513. doi: 10.1155/2014/290513

McEvoy, L. K., Smith, M. E., and Gevins, A. (2000). Test-retest reliability of cognitive EEG. *Clin. Neurophysiol.* 111, 457–463. doi: 10.1016/S1388-2457(99)00258-8

Nieuwenhuis, S., Aston-Jones, G., and Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychol. Bull.* 131, 510–532. doi: 10.1037/0033-2909.131.4.510

Okada, K., Kobayashi, S., Aoki, K., Suyama, N., and Yamaguchi, S. (1998). Assessment of motivational loss in poststroke patients using the Japanese version of Starkstein's Apathy Scale. *Nosotchu* 20, 318–323. doi: 10.3995/jstroke.20.318

Onoda, K., Abe, S., and Yamaguchi, S. (2010). Feedback-related negativity is correlated with unplanned impulsivity. *Neuroreport* 21, 736–739. doi: 10.1097/WNR.0b013e32833bfd36

Onoda, K., and Yamaguchi, S. (2015). Dissociative contributions of the anterior cingulate cortex to apathy and depression: topological evidence from resting-state functional MRI. *Neuropsychologia* 77, 10–18. doi: 10.1016/j.neuropsychologia.2015.07.030

Pfabigan, D. M., Alexopoulos, J., Bauer, H., and Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology* 48, 656–664. doi: 10.1111/j.1469-8986.2010.01136.x

Portella, C., Machado, S., Arias-Carrion, O., Sack, A. T., Silva, J. G., Orsini, M., et al. (2012). Relationship between early and late stages of information processing: an event-related potential study. *Neurol. Int.* 4:e16. doi: 10.4081/ni.2012.e16

Potts, G. F., Martin, L. E., Burton, P., and Montague, P. R. (2006). When things are better or worse than expected: the medial frontal cortex and the allocation of processing resources. *J. Cogn. Neurosci.* 18, 1112–1119. doi: 10.1162/jocn.2006.18.7.1112

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., et al. (2005). Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport* 16, 407–411. doi: 10.1097/00001756-200503150-00020

Schreckenberger, M., Klega, A., Gründer, G., Buchholz, H. G., Scheurich, A., Schirrmacher, R., et al. (2008). Opioid receptor PET reveals the psychobiologic correlates of reward processing. *J. Nucl. Med.* 49, 1257–1261. doi: 10.2967/jnumed.108.050849

Snyder, E., and Hillyard, S. A. (1976). Long-latency evoked potentials to irrelevant, deviant stimuli. *Behav. Biol.* 16, 319–331. doi: 10.1016/S0091-6773(76)91447-4

Starr, L. B., Robinson, R. G., and Price, T. R. (1983). Reliability, validity, and clinical utility of the social functioning exam in the assessment of stroke patients. *Exp. Aging Res.* 9, 101–106. doi: 10.1080/03610738308258434

Stuss, D. T., Van Reekum, R., and Murphy, K. J. (2000). "Differentiation of states and causes of apathy," in *The Neuropsychology of Emotion*, ed J. Borod (New York, NY: Oxford University Press), 340–363.

Takasawa, N., Takino, R., and Yamazaki, K. (1990). Event-related potentials derived from feedback tones during motor learning. *Jpn. J. Physiol. Psychol.* Psychophysiol. 8, 95–101. doi: 10.5674/jjpppp1983.8.95

Takeuchi, M., Miyaoka, H., Tomoda, A., Suzuki, M., Lu, X., and Kitamura, T. (2011). Validity and reliability of the Japanese version of the Temperament and Character Inventory: a study of university and college students. *Compr. Psychiatry* 52, 109–117. doi: 10.1016/j.comppsych.2010.04.002

Thigpen, N. N., Kappenman, E. S., and Keil, A. (2017). Assessing the internal consistency of the event-related potential: an example analysis. *Psychophysiology* 54, 123–138. doi: 10.1111/psyp.12629

Van Schuerbeek, P., Baeken, C., De Raedt, R., De Mey, J., and Luypaert, R. (2011). Individual differences in local gray and white matter volumes reflect differences in temperament and character: a voxel-based morphometry study in healthy young females. *Brain Res.* 1371:32–42. doi: 10.1016/j.brainres.2010.11.073

Williams, L. M., Simms, E., Clark, C. R., Paul, R. H., Rowe, D., and Gordon, E. (2005). The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker." *Int. J. Neurosci.* 115, 1605–1630 doi: 10.1080/00207450590958475

Wu, Y., and Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Res.* 1286, 114–122. doi: 10.1016/j.brainres.2009.06.032

Yamagata, S., Yamaguchi, S., and Kobayashi, S. (2004). Impaired novelty processing in apathy after subcortical stroke. *Stroke* 35, 1935–1940. doi: 10.1161/01.STR.0000135017.51144.c9

Yeung, N., Holroyd, C. B., and Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cereb. Cortex* 15, 535–544. doi: 10.1093/cercor/bhh153

Yeung, N., and Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *J. Neurosci.* 24, 6258–6264. doi: 10.1523/JNEUROSCI.4537-03.2004

Zheng, Y., Li, Q., Zhang, Y., Shen, H., Gao, Q., and Zhou, S. (2017). Reward processing in gain versus loss context: an ERP study. *Psychophysiology* 54, 1040-1053 doi: 10.1111/psyp.12855

Zung, W. W. K. (1965). A self-rating depression scale. *Arch. Gen. Psychiatry* 12, 63–70. doi: 10.1001/archpsyc.1965.01720310065008

Isolating Discriminant Neural Activity in the Presence of Eye Movements and Concurrent Task Demands

Jon Touryan[1]*, Vernon J. Lawhern[1], Patrick M. Connolly[2], Nima Bigdely-Shamlo[3] and Anthony J. Ries[1]

[1] U.S. Army Research Laboratory, Future Soldier Technologies Division, Human Research and Engineering Directorate, Aberdeen Proving Ground, Aberdeen, MD, United States, [2] Teledyne Scientific Company, Durham, NC, United States, [3] Qusp Labs, San Diego, CA, United States

A growing number of studies use the combination of eye-tracking and electroencephalographic (EEG) measures to explore the neural processes that underlie visual perception. In these studies, fixation-related potentials (FRPs) are commonly used to quantify early and late stages of visual processing that follow the onset of each fixation. However, FRPs reflect a mixture of bottom-up (sensory-driven) and top-down (goal-directed) processes, in addition to eye movement artifacts and unrelated neural activity. At present there is little consensus on how to separate this evoked response into its constituent elements. In this study we sought to isolate the neural sources of target detection in the presence of eye movements and over a range of concurrent task demands. Here, participants were asked to identify visual targets (Ts) amongst a grid of distractor stimuli (Ls), while simultaneously performing an auditory N-back task. To identify the discriminant activity, we used independent components analysis (ICA) for the separation of EEG into neural and non-neural sources. We then further separated the neural sources, using a modified measure-projection approach, into six regions of interest (ROIs): occipital, fusiform, temporal, parietal, cingulate, and frontal cortices. Using activity from these ROIs, we identified target from non-target fixations in all participants at a level similar to other state-of-the-art classification techniques. Importantly, we isolated the time course and spectral features of this discriminant activity in each ROI. In addition, we were able to quantify the effect of cognitive load on both fixation-locked potential and classification performance across regions. Together, our results show the utility of a measure-projection approach for separating task-relevant neural activity into meaningful ROIs within more complex contexts that include eye movements.

Keywords: fixation-related potentials, EEG, eye tracking, target detection, cognitive load

INTRODUCTION

Goal-directed eye movements are a ubiquitous component of everyday life and integral to our perception of the world. Over recent decades, numerous visual search studies have used eye movement patterns to better understand perceptual and attentional processes that underlie human vision (Kowler, 2011). In contrast, the majority of human electrophysiological studies of visual

search continue to use fixation constrained paradigms, artificially limiting the natural linkage between attentional shifts and subsequent eye movements. Thus, extending these paradigms into a framework of overt visual search would enable the validation of attentional models in a more natural context. However, a number of potential confounds and analytical challenges emerge when interpreting electroencephalography (EEG) in the presences of eye movements (Nikolaev et al., 2016). One of the primary confounds is the large eye movement related signals around the events of interest, namely saccades and fixations. These include the corneo-retinal and saccadic spike potentials along with eyelid artifacts. Importantly, the magnitude of these signals systematically scale with the direction, amplitude, and velocity of the saccade. Furthermore, saccade features themselves can systematically vary with task design or conditions. To address this concern, a number of methods have been developed to account for the effects of saccade sequence (Dandekar et al., 2012b) or isolate eye movement related signals within the EEG record (Plöchl et al., 2012). These approaches have been successful to the degree that they were able to reveal task-relevant activity, such as the P3 component, that may otherwise have been conflated with eye movement related artifacts (Dandekar et al., 2012a; Devillez et al., 2015).

Robust saccade detection and quantification presents another methodological challenge. Previously, EEG studies often relied on an explicit electrooculography (EOG) measurement, horizontal or vertical, for detecting the onset of a saccade (Gaarder et al., 1964; Thickbroom et al., 1991; Kazai and Yagi, 1999). The benefit of using this signal is both the high temporal resolution and the *de facto* alignment with the EEG record. Unfortunately, the lack of precision in determining the direction and distance of saccades limits these studies to paradigms with a small number of predetermined fixation locations. However, recent advances in the speed and accuracy of infrared eye-tracking technology has made it possible to link gaze position with neural activity at both high spatial and temporal resolution. This has led to a growing number of studies that explore the neural correlates of target detection during visual search, in both controlled (Brouwer et al., 2013) and free-viewing paradigms (Kamienkowski et al., 2012; Dias et al., 2013; Jangraw et al., 2014; Kaunitz et al., 2014; Ušćumlić and Blankertz, 2016; Wenzel et al., 2016).

In addition to the above measurement and signal processing challenges, there is the more nuanced task of interpreting brain activity in the context of planned and executed eye movements (Nikolaev et al., 2016). This remains a significant obstacle for studies focusing on both perceptual and cognitive phenomena. First, there is the task of quantifying or controlling for stimulus properties. When the eyes are free to move, stimuli impinging on the retina will necessarily vary across conditions and participants, even when gaze position is guided by the task sequence. In more controlled settings, experimental design can ensure that small differences in eye position do not significantly bias the statistics of the stimuli. However, this becomes more challenging for the ultimate goal of free-viewing in natural scenes where spatial frequency, orientation, and chromatic distributions can vary widely within a single image. Likewise, there is the challenge

of separating saccade planning and execution from the perceptual or cognitive signal of interest. Even when utilizing high density EEG and source localization techniques, the spatial resolution of the saccadic preparatory signals is limited. Thus, accounting for these signals via subtraction across equated conditions (Nikolaev et al., 2013), regression (Dandekar et al., 2012b), or other techniques (Dias et al., 2013) is an important factor for the interpretation of para-saccadic neural activity.

Despite these recent advancements, there remains a need for development and validation of methods for the quantification of both perceptual and cognitive phenomena in the presence of eye movements. Part of this process is the evaluation of novel analytical approaches within paradigms that enable a more direct comparison to related fixation-constrained studies. Similarly, any particular method may only address some of the above challenges while still providing valuable insight when applied within the appropriate constraints or combined with other techniques. It is within this context that we propose the following approach for separating neural activity into meaningful regions of interest (ROIs) in the presence of eye movements. To evaluate our approach, we utilized data from a previously publish study (Ries et al., 2016) that employed a dual-task paradigm, *visual target detection* and *auditory N-back*, to quantify the effect of working memory load on the lambda response. The primary observation from this study was a small but significant reduction in the lambda amplitude with increasing cognitive load.

Here, we were able to separate the neural response to each fixation into six ROIs by applying a technique that linearly combines activity from independent sources based on their equivalent dipole location. Within each ROI we show a distinct neural response that, to varying degrees, discriminated target from non-target fixations and was differentially modulated by cognitive load. While the task design mitigated the overlapping response from adjacent saccades, common in free-viewing visual search, this approach is a substantive step in the interpretation of fixation-related brain activity. When combined with GLM-based techniques for the deconvolution of overlapping FRPs, this approach can be applied to more natural contexts where the interplay between bottom-up and top-down neural activity is not well understood.

MATERIALS AND METHODS

The experiment used in this study has been described in a previous publication (Ries et al., 2016). Here, we provide a summary of stimuli and procedure, followed by a more detailed description of the novel ROI analysis method.

Participants

Fourteen participants volunteered for the study; all participants were right-handed males with an average age of 32.8 years. All participants had 20/20 vision or corrected to 20/20 vision. This study was conducted in accordance with the U.S. Army Research Laboratory's IRB requirements (32 CFR 219 and DoDI 3216.02). The voluntary, fully informed consent of research participants was obtained in written form. The study was reviewed and

approved by the U.S. Army Research Laboratory's IRB before the study began.

Stimuli and Procedure

Participants performed a guided visual target detection task on a 7 × 7 grid (23.9° × 23.9° visual angle) of equally spaced and variably oriented "T" or "L" characters (1.1° visual angle) presented on a low contrast 1/f noise background from a viewing distance of approximately 65 cm (**Figure 1**). Eye fixations were guided across the grid by a red annulus (2.3° visual angle) that randomly surrounded one of the characters for a duration of 1 s before moving to the next randomly selected character. Participants were instructed to saccade to and fixate on the character in the center of the red annulus and to press a button (left hand) only when a "T" (visual target) was present. Visual target characters appeared on 10% of trials. Participants were instructed to maintain fixation on the character until the next red annulus appeared. All red annuli surrounding a non-target "L" were at least two characters from any "T" present on the grid to minimize peripheral detection. The guided visual target detection task was performed in one of five conditions: visual alone (silent condition), while ignoring binaurally presented digits (numbers 0–9), or while using the auditory digits in a 0, 1, or 2-back working memory task. The digit "0" was only used in the 0-Back condition where it served as the auditory target. Auditory stimuli were presented every 2 s with a 500 ms offset from a shift in the red annulus location. Participants were instructed to make a button press (right hand) for auditory targets, which occurred on 20% of trials. Thus, the same number of targets appeared in both tasks during the 0-back, 1-back, and 2-back conditions. Participants performed two consecutive blocks of the same condition (silent, ignore, 0-back, 1-back, 2-back) with the condition order counterbalanced. Each of the 10 blocks had a duration of 200 s with self-paced rest periods between blocks. Participants were given practice in each N-back condition, prior to experimental data collection, until they reached above chance performance.

Eye Tracking

Eye-tracking data were sampled at 250 Hz using the SMI RED 250 system (Teltow, Germany). A 15-point calibration was performed prior to the practice and experimental blocks. A *post-hoc* model was fit to the eye-tracking data for each participant to increase accuracy of the gaze position estimate. Briefly, we used the expected eye position (i.e., location of the red annulus) to fit a quadratic regression model for both the horizontal and vertical gaze position vectors (**Figure 2**). A temporal lag (250 ms) was applied to the expected location (red annulus) to account for the delay between annulus onset and subsequent fixation.

Saccades and fixations were detected in the eye-tracking data using a velocity-based algorithm (Engbert and Mergenthaler, 2006; Dimigen et al., 2011). Saccades and fixations were detected using a velocity factor of 6 (standard deviations of the velocity distributions), minimum saccade duration of 20 ms, minimum fixation duration of 350 ms. If two saccades occurred within a 350 ms window, only the fixation corresponding to the largest
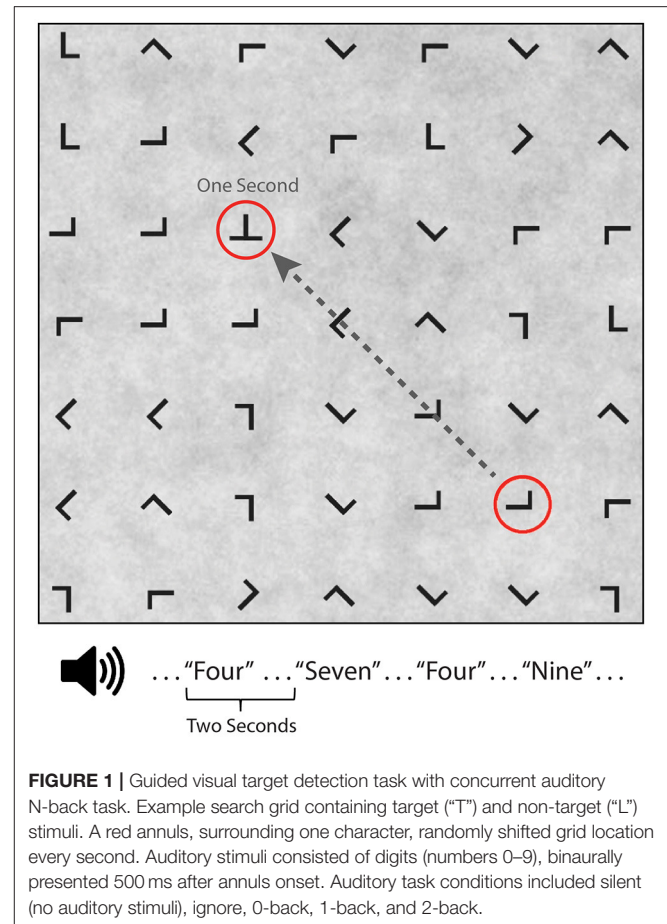

FIGURE 1 | Guided visual target detection task with concurrent auditory N-back task. Example search grid containing target ("T") and non-target ("L") stimuli. A red annuls, surrounding one character, randomly shifted grid location every second. Auditory stimuli consisted of digits (numbers 0–9), binaurally presented 500 ms after annuls onset. Auditory task conditions included silent (no auditory stimuli), ignore, 0-back, 1-back, and 2-back.

saccade was preserved. Fixations were only considered task-relevant or "valid" if they were within 3 degrees of the current stimulus location. These criteria were chosen to focus analyses on the first saccade onto the new stimulus (red annulus) location.

Electroencephalography and Feature Extraction

Electrophysiological signal acquisition and analysis steps are outlined in **Figure 3**. EEG recordings were digitally sampled at 1,024 Hz from 64 scalp electrodes over the entire scalp using a BioSemi Active Two system (Amsterdam, Netherlands). External leads were placed on the outer canthi, and above and below the orbital fossa of the right eye to record electrooculography (EOG). EEG was referenced offline to the average mastoids, down-sampled to 256 Hz (f_s), and digitally high-pass filtered above 1 Hz using the EEGLAB toolbox (Delorme and Makeig, 2004). Large artifacts were detected using a previous described technique (Touryan et al., 2016). Briefly, EEG sessions were segmented into high-resolution 100 ms epochs, with a 10 ms step size. Epochs were marked as high noise if the average power between 90 and 120 Hz was greater than three standard deviations above the mean for all epochs. These epochs were then removed and the remaining EEG record was lowpass filtered below 50 Hz.
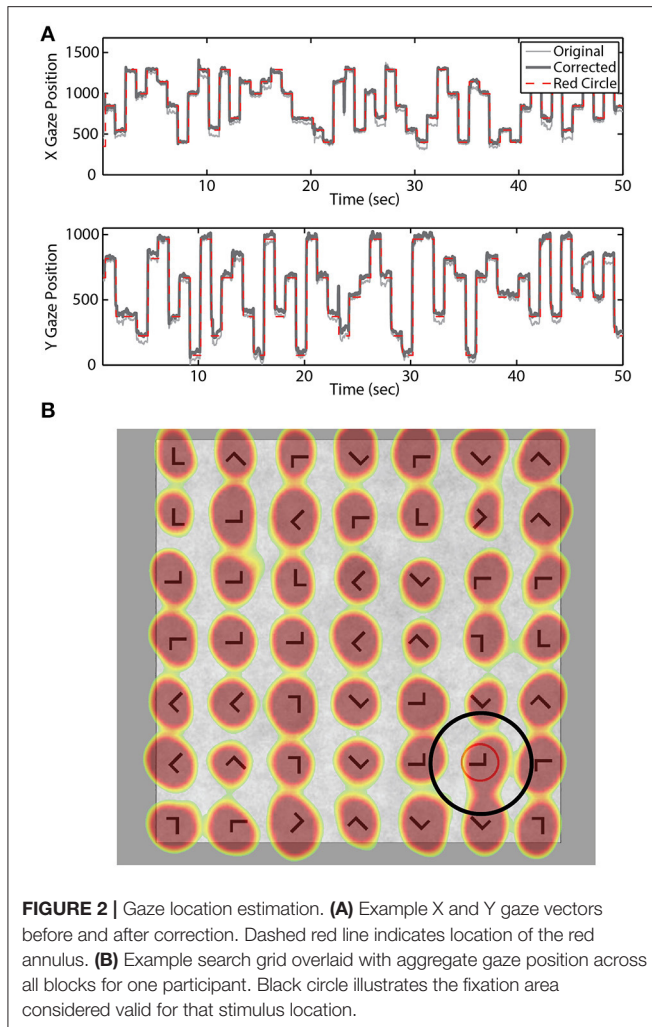
FIGURE 2 | Gaze location estimation. **(A)** Example X and Y gaze vectors before and after correction. Dashed red line indicates location of the red annulus. **(B)** Example search grid overlaid with aggregate gaze position across all blocks for one participant. Black circle illustrates the fixation area considered valid for that stimulus location.
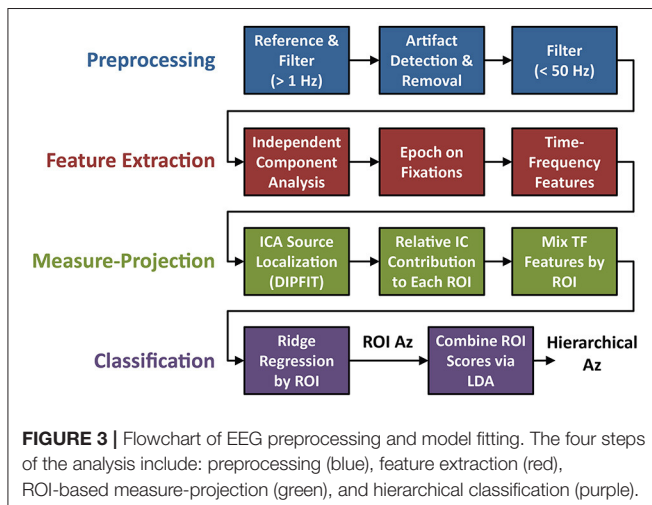
FIGURE 3 | Flowchart of EEG preprocessing and model fitting. The four steps of the analysis include: preprocessing (blue), feature extraction (red), ROI-based measure-projection (green), and hierarchical classification (purple).

Each "clean" EEG session was decomposed into independent components using the Extended Infomax ICA algorithm implemented in EEGLAB (Delorme and Makeig, 2004). The

equivalent dipole locations of these independent sources were then estimated using the EEGLAB implementation of DIPFIT (Scherg, 1990; Pascual-Marqui et al., 1994). IC activation epochs were extracted around each valid fixation using a temporal window spanning 300 ms before and 1,000 ms after fixation onset. Time-frequency features were also calculated for each epoch using a wavelet transform (Torrence and Compo, 1998). Specifically, we used the Morlet wavelet function:

$$\psi_0(t) = c\pi^{-1/4}e^{i\omega_0 t}e^{-t^2/2} \tag{1}$$

where ω_0 is the central frequency and c the normalization constant. This function was used to create a basis set of 30 wavelets covering the available frequency range with minimum scale of $2/f_s$ and a discrete step size of 0.25 (wavelet transform software available at http://paos.colorado. edu/research/wavelets/). After the wavelet transform, the spectral power of each epoch was computed via multiplication with the complex conjugate of the corresponding epoch. While this time-frequency decomposition included frequencies from 1 to 128 Hz, only frequencies below 32 Hz were included in subsequent analyses.

To isolate activity in brain regions of interest (ROIs), the above IC activation epochs were linearly mixed based on equivalent dipole location using the initial steps of measure-projection analysis (Bigdely-Shamlo et al., 2013). ICs with equivalent dipoles outside of the MNI model brain volume were identified and excluded from analysis (see Supplementary Section 3). These ICs often corresponded to corneo-retinal potentials (i.e., EOG) or muscle artifacts (i.e., EMG). The remaining k IC processes were preserved and their corresponding fixation-locked activation epochs used as the "measure" for each dipole location in the mixing process. Specifically, the fixation-locked activations or measures can be indexed as $M_i, i = 1 \ldots k$ for each IC, and the equivalent dipole location $x \in V \subset R^3$ indexed as $D(x_i), i = 1 \ldots k$. Importantly, there exists uncertainty in dipole localization arising from errors in tissue conductivity parameters, electrode co-registration, noise in the IC estimate process, and between-subject variability in the location of equivalent functional cortical areas. To capture this uncertainty in the mixing process we can instead model each equivalent dipole as a spherical (3-D) Gaussian with uniform covariance σ^2, centered at the estimated dipole location x_i. The spherical Gaussian is truncated at $t*\sigma$ to minimize the erroneous influence of distant dipoles in sparsely populated regions. Thus, the probability of dipole $D(x_j)$ being located at position $y \in V$ now becomes $P_j(y) = TN(y; x_j, \sigma^2, t)$, where TN is a truncated normal distribution centered at x_j. Then for an arbitrary location $y \in V$, the expected value of the measure becomes:

$$E\{M(y)\} = \langle M(y) \rangle = \frac{\sum_{i=1}^{k} P_i(y) M_i}{\sum_{i=1}^{k} P_i(y)} \tag{2}$$

Where $M(y)$ is the combined fixation-locked activity at location y from all proximal ICs. We used this approach to calculate the aggregate measure $\langle M(y) \rangle$, either fixation-related potential (FRP) or time-frequency spectrum, for specified regions of the brain

volume. For this study, six *a priori* ROIs (**Figure 4**) were defined using the Measure Projection Toolbox (http://sccn.ucsd.edu/wiki/MPT). Each ROI consisted of all regions of LONI LPBA40 atlas (Shattuck et al., 2008) that included the corresponding anatomical label (e.g., "occipital"). The only additional parameter σ (standard deviation of the Gaussian distribution) in this calculation was set to 12 mm. This value produced smooth spatial distributions in each ROI given the relatively small number of participants (N = 14). The six ROIs included: *occipital, fusiform, temporal, parietal, cingulate,* and *frontal* cortices.

Hierarchical Classification

For the classification step, we used a two-stage hierarchical approach to dissociate target and non-target fixation epochs. In the first stage, ridge regression (MATLAB® *ridge* function) was applied separately to the time-frequency epochs from each ROI. Specifically, we applied regularized regression to the entire temporal epoch and frequencies up to 32 Hz. The regularization parameter was determined via calculating the effective degrees of freedom as a function of lambda (λ):

$$df\left(\lambda\right) = tr\left(X\left(X^TX + \lambda I\right)^{-1}X^T\right) = \sum_{j=1}^{m}\frac{d_j^2}{d_j^2 + \lambda} \quad (3)$$

where d_j are the singular values of the *n x m* data matrix *X*. As these functions were roughly similar across ROIs (Supplementary Figure 3), we selected a hyperparameter value (Lemm et al., 2011) such there were approximately 1 target and 10 non-target observations per degree of freedom (see Model Considerations for a discussion of the validity of this approach). However, the exact value had minimal effect on the results (see Supplementary Section 2). The second stage utilized the regression output, or latent variable estimate, from the six ROIs to provide a single classification score and label for each fixation epoch. In this second stage, we employed linear discriminant analysis (LDA; MATLAB® *fitcdiscr* function) and coefficients for both stages were fit within a single 5-fold cross-validation scheme. Area under the ROC curve (Az) was calculated for each ROI, as well as for the second-stage LDA classifier. Finally, for direct inference into the discriminant neural activity we calculated the forward model for each ROI (Haufe et al., 2014). Specifically, the

regression weights (*W*) were used to estimate the forward model (*A*), such that $A = \Sigma_X W \Sigma_{\hat{s}}$, where Σ_X and $\Sigma_{\hat{s}}$ are the empirical data and score covariance respectively.

To facilitate comparison with other approaches, we included two techniques commonly used for single-trial classification of EEG. Both methods were applied directly to the filtered EEG data (64 channels) using the same fixation epochs described above. First, Hierarchical Discriminant Components Analysis (HDCA) was applied with each epoch divided into 8 equal-sized temporal windows (Gerson et al., 2006). Second, we used the xDAWN algorithm (Rivet et al., 2009) to identify the 8 most discriminant spatial filters followed by a Bayesian linear discriminant analysis, collectively referred to as XD+BLDA (Cecotti et al., 2011). Area under the ROC curve was calculated for both of these classifiers on all participants.

RESULTS

Behavioral and Ocular Measures

Detailed behavioral analysis of this study has been previously reported (Ries et al., 2016), however the relevant statistics are summarized below for comparison with the classification results. Reaction time and accuracy were analyzed separately for the visual and auditory tasks using a one-way repeated measures ANOVA (Greenhouse-Geisser correction reported where appropriate). The primary factor was auditory task condition, which had five levels in the visual task (Silent, Ignore, 0-Back, 1-Back, 2-Back), and three levels in the auditory task (0-Back, 1-Back, 2-Back). There was a trend for decreased accuracy in the visual task as a function of cognitive load (i.e., auditory N-back level); however this was not statistically significant (**Table 1**). We did observe a highly significant effect of cognitive load on reaction time (RT) in the visual task [$F_{(2.73, 35.44)} = 29.24$, $p < 0.001$, $\eta^2 = 0.69$] showing that visual target RT increased as a function of cognitive load. Likewise, analysis of the auditory task showed both a significant decrease in accuracy [$F_{(1.61, 20.96)} = 6.74$, $p < 0.01$, $\eta^2 = 0.34$] and increase in RT [$F_{(1.60, 20.79)} = 17.64$, $p < 0.001$, $\eta^2 = 0.58$] with increasing auditory task demands. While the behavioral results showed that auditory working memory load had a significant negative impact on visual task performance, exhibited through increased RT, the near-ceiling accuracy likely mitigated any decline of this corresponding metric. Together, the behavioral results suggest that participants were not exclusively favoring one modality as
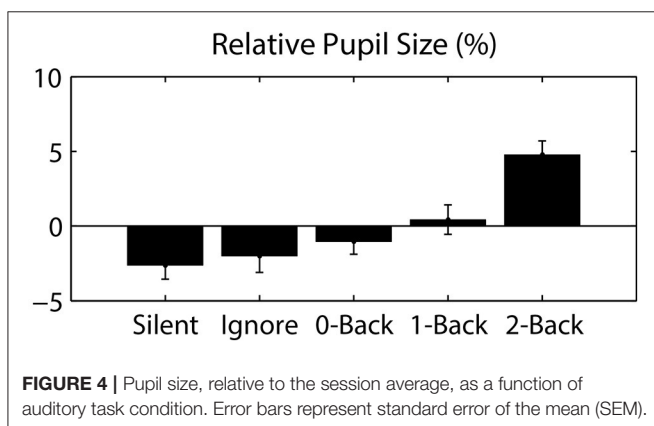
FIGURE 4 | Pupil size, relative to the session average, as a function of auditory task condition. Error bars represent standard error of the mean (SEM).

TABLE 1 | ANOVA statistics for accuracy and RT in visual and auditory tasks.

Factor	df	F	p	η^2
VISUAL TARGET (BY CONDITION*)				
Accuracy	2.67,34.77	2.43	0.088	0.16
RT	2.73,35.44	29.24	<0.001	0.69
AUDITORY TARGET (BY CONDITION*)				
Accuracy	1.61,20.96	6.74	0.008	0.34
RT	1.60,20.79	17.64	<0.001	0.58

*Auditory N-back level: Silent, Ignore, 0-Back, 1-Back, 2-Back.

performance declined in both the visual and auditory tasks with increased cognitive load.

Since eye movements were constrained by the nature of the visual task (guided target detection), the majority of ocular metrics did not significantly differ across blocks or conditions. As expected, we found no significant difference in fixation duration (0.962 ± 0.053 s; mean $\pm$ STD) or saccade distance (13.316 ± 1.072 degrees) across auditory task conditions. However, we did observe a large change in pupil dilation as a function of cognitive load (**Figure 4**). Specifically, we calculated the average pupil size in each condition relative to the average size across each participant's entire session. This relative pupil-size metric was significantly modulated by auditory task condition [$F_{(3.04, 39.57)}$ = 7.98, $p < 0.001$, $\eta^2 = 0.38$], exhibiting an increase in size with working memory load. Due to static luminance and counterbalanced condition order (see Materials and Methods), this modulation was unlikely to be a consequence of either changes in luminance or time-on-task (Beatty, 1982). Thus, our results indicate that the task-induced cognitive load increased the arousal level of participants, as has been shown in similar paradigms (Kahneman and Beatty, 1966).

Fixation-Related Potentials (FRPs) by ROI

We calculated FRPs for each brain region by combining independent components activations within fixation epochs, using a ROI-based measure-projection approach (ROI-MPA). An IC's contribution to a given ROI was determined by the overlap between the anatomically defined region and the equivalent dipole Gaussian density function (see Materials and Methods). Importantly, by excluding equivalent dipoles located outside of this brain volume this approach attempts to minimize the influence of non-brain signals, such as those generated by eye movements, from the ROIs (see Supplementary Section 3). **Figure 5** shows the grand average FRPs from each ROI: *occipital, fusiform, temporal, parietal, cingulate,* and *frontal* cortices. To account for the differing number of included ICs, FRPs from each participant were uniformly scaled by total variance and are shown in arbitrary units. All Included epochs were from valid fixations (within 3 degrees of the current stimulus) and free of large artifacts. The average number of target and non-target epochs, by condition, are shown in **Table 2**.

The ROI FRP waveforms shown in **Figure 5** exhibit a clear distinction across brain regions. Both target and non-target FRPs show a temporal progression through the visual cortices (occipital, fusiform, temporal) and reflect known electrophysiological signatures, such as the P1 or lambda component. Importantly, the distinction between target and non-target FRPs is evident in most ROIs. To identify periods of significant difference in the FRP waveforms we used a paired t-test at causal time points in each ROIs (255 time points $\times$ 6 ROIs). A single false-discovery-rate correction for multiple comparisons was then applied to all p-values (Benjamini and Hochberg, 1995). As expected, visual cortices show this distinction in earlier epochs, consistent with the visual mismatch negativity (vMMN: 150–250 ms), while the parietal and cingulate cortex exhibit a clear late positive deflection, indicative of the P3 component. In contrast, the frontal cortex shows little saccade-related EOG artifact that would be expected to dominate frontal electrodes (e.g., Fz).
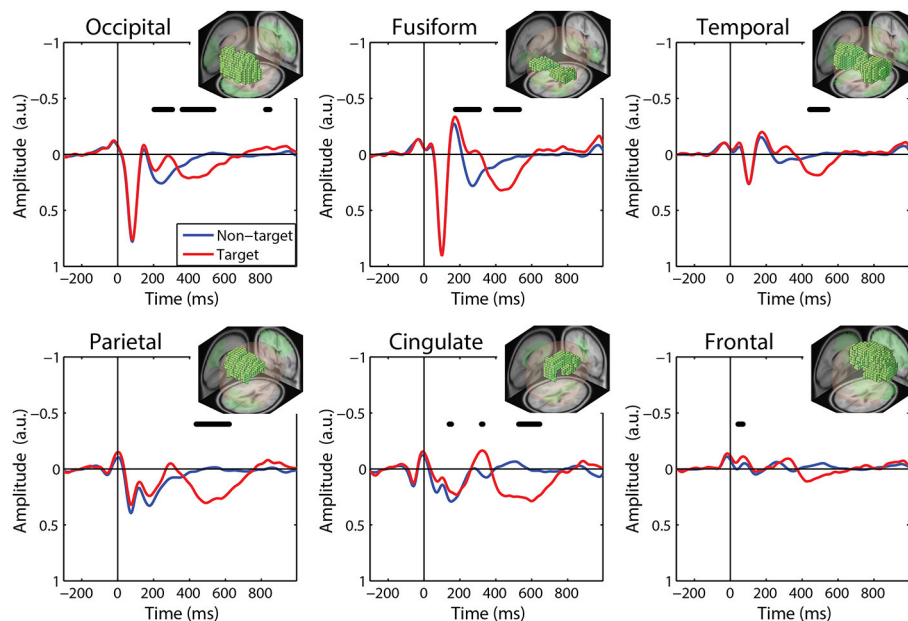
FIGURE 5 | Grand average FRP by ROI. Target and non-target waveforms from each ROI (black line indicates a significant difference, $p < 0.01$). All voxels included in the ROI are shown as the inset. FRP response consists of a linear sum of IC activations weighted by their contribution to the corresponding ROI and are shown in arbitrary units. Note: ICs with equivalent dipoles located outside of the brain volume, such as those produced by EOG, are not aggregated in the ROI FRPs.

Condition	Target epochs	Non-target epochs
Silent	21 (5)	209 (49)
Ignore	21 (7)	209 (66)
0-Back	23 (5)	212 (40)
1-Back	21 (5)	218 (48)
2-Back	23 (7)	203 (63)
Total	108 (20)	1,051 (211)

Values represent the mean, standard deviation in parentheses.

For comparison to the standard approach, we also calculated target and non-target FRPs for electrodes that most directly correspond to each ROI. **Figure 6** shows the grand average FRP from these corresponding electrodes, using the same fixation epochs as above (**Table 2**). For occipital regions, the electrode and ROI FRPs are quite similar, as these electrodes are least affected by changes in the corneo-retinal potential and other saccade related activity. However, EOG artifact increasingly dominates the anterior regions, especially frontal electrodes (e.g., Fz). This can lead to difficulty in dissociating neural from EOG phenomena in more cognitive processes. For these grand averages, the number of target epochs was an order of magnitude lower than that for non-target epochs. However, a similar result was found for both ROI and electrode FRPs when these numbers were equated by randomly sampling a subset of non-target epochs (Supplementary Figures 1, 2).

Finally, to quantify the effect of auditory task condition (i.e., working memory load) on the ROI FRP we performed the following analysis. We measured the amplitude of the FRP for auditory conditions at either end of the difficulty spectrum: *Ignore* and *2-Back*. These were chosen as representative of low and high cognitive load conditions; although similar results were found when comparing the *Silent* and *2-Back* conditions. To capture the P3 waveform, we calculated the average amplitude within a 300–700 ms post-fixation window. We then applied a two-way repeated measures ANOVA, with factors ROI and condition, to quantify the effect of cognitive load on this components (**Table 3**). As expected, there was a strong effect of auditory task condition [$F_{(1, 65)} = 22.45, p < 0.001, \eta^2 = 0.15$] with the amplitude of the P3 being significantly smaller during high, relative to low, working memory load.

Classification by ROI

For single-trial classification, we used fixation-locked time-frequency features from each ROI. Before linearly mixing IC activations, we first applied a Morlet wavelet transform to each epoch. We then calculated the spectral power of the wavelet transform before combining these time-frequency epochs. **Figure 7A** shows the grand average spectral FRPs for target epochs from each ROI. These average time-frequency responses, analogous to event related spectral perturbations (ERSP), show a similar time course as the FRPs above. Visual cortices have an early, mid-frequency (alpha band) component that is the spectral equivalent of the lambda response. Similarly, the parietal, cingulate, and frontal cortices are dominated by

a later lower frequency (delta band) activity, reflecting the P3 component.

To classify target from non-target fixation epochs, we used ridge regression on these high-dimensional time-frequency features. We constructed separate classifiers for each ROI that utilized spectral information, below 32 Hz, from the entire fixation epoch. The forward models for each ROI are shown in **Figure 7B**. Again, the time course is similar to the grand average FRPs, where visual cortices have discriminant activity with smaller latencies and higher-frequency components. The marginal activations (**Figure 7C**) provide a more direct view of the temporal profile of the discriminant activity.

The relative discriminant power of each ROI was quantified by using classifier performance in a two-way repeated measures ANOVA, with factors ROI and auditory task condition (**Table 4**). We found a significant modulation of the area under the ROC curve (Az) by region [$F_{(2.41, 156.73)} = 12.84, p < 0.001, \eta^2 = 0.15$]. The average performance across all ROIs and participants was 0.741 ± 0.068 (**Figure 8**), substantially below behavioral performance in the visual detection task (average accuracy = 0.974 ± 0.030).

Integration across regions required a second-stage classifier applied to the output of the ROI regression step. For each epoch, the output from the ROI classifiers (i.e., vector of six classification scores) were combined using a linear discriminant function. Not surprisingly, this hierarchical approach resulted in significantly better performance (Az: 0.851 ± 0.096) than the individual ROI classifiers ($p < 0.001$; Wilcoxon signed rank test). Interestingly, there was a wide range in classifier performance across participants with Az values ranging from 0.708 to 0.947, indicating that for some individuals our approach was able to identify visual targets at an accuracy similar to behavioral performance. This was despite ongoing neural activity related to the concurrent auditory task as well as the planning and execution of eye movements.

This hierarchical approach compared favorably to other common classification techniques (**Figure 8B**). Specifically, we applied Hierarchical Discriminant Components Analysis (Gerson et al., 2006) to the EEG channel data using the same epochs as above. We also applied the xDAWN filtering algorithm (Rivet et al., 2009) followed by Bayesian linear discriminant analysis, or XD+BLDA (Cecotti et al., 2011). HDCA and XD+BLDA classification accuracies were similar to our hierarchical approach with HDCA having slightly higher overall performance ($p = 0.013$).

While we were able to classify visual target from non-target stimuli during a concurrent auditory task, there was a significant modulation of ROI classification performance as a function of cognitive load (**Table 4**). Here, this modulation was the inverse of that observed in the relative pupil size. Classification performance decreased with increasing auditory task demands; except in the silent and ignore condition in which performance was similar. At the hierarchical stage, while the target scores were significantly modulated by condition (**Figure 9B**), the classification performance was not (**Table 4**, **Figure 9A**). Much like behavioral accuracy in the visual task, the hierarchical classifier performance remained relatively constant
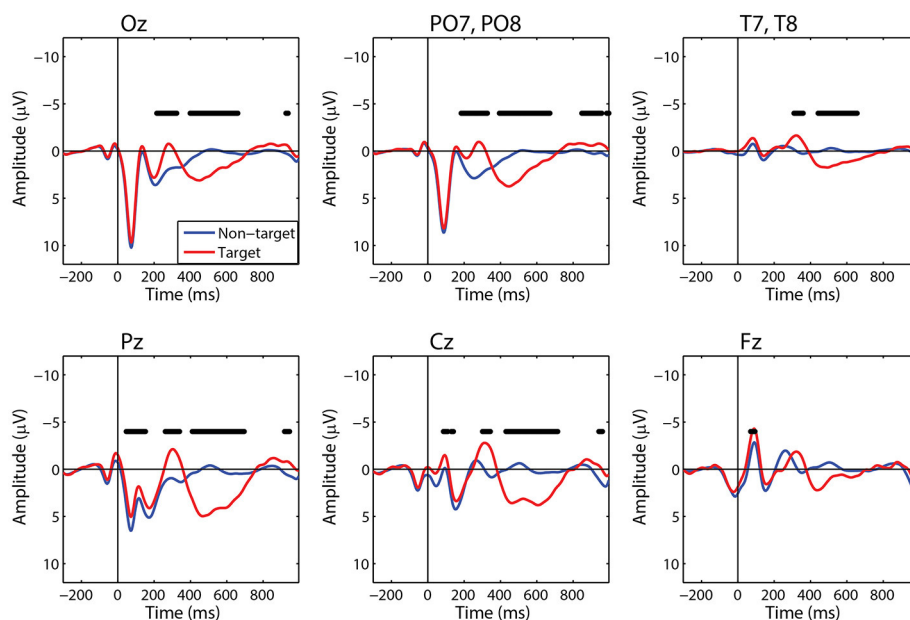
FIGURE 6 | Grand average FRP by electrode. Target and non-target waveforms from electrodes corresponding to each ROI (black line indicates a significant difference, $p < 0.01$).

across auditory task conditions. However, increasing working memory load clearly affected both the neural activity and pupil diameter in a manner consistent with increased arousal (Murphy et al., 2011).

DISCUSSION

In this study, we used a novel approach to determine if the neural response associated with visual target detection could be separated into meaningful components in the presence of eye movements and concurrent task demands. Here, we employed a single framework to isolate neural from non-neural activity and to separate the FRP into cortical regions. Using common statistical techniques we were then able to classify target from non-target FRPs across ROIs on a single-trial basis at a level similar to state-of-the-art machine learning algorithms. By doing so, we were able to show a clear time-course of discriminant activity associated with target detection as well as the modulating effect of cognitive load. While our task design mitigated the overlapping response from previous or subsequent fixations, the results demonstrate the potential for separating task-relevant neural activity in more complex contexts that include eye movements and concurrent tasks.

The EEG analysis framework described here is both specific enough to separate activity by ROI and sensitive enough to evaluate the effects of cognitive load. While more traditional channel-based approaches of FRP analysis may be able to separate these effects by scalp location (e.g., Oz vs. Pz), the inference into the constituent neural sources remains more difficult. Likewise, channel-based approaches require an explicit EOG mitigation or removal process, using ICA (Plöchl et al.,

TABLE 3 | ANOVA statistics for P3 amplitude in the visual task.

Factor	df	F	p	η^2
P3 AMPLITUDE (300–700 ms)				
Condition*	1,65	22.45	< 0.001	0.15
ROI	2.13,27.70	11.55	< 0.001	0.38
Interaction	5,65	1.69	0.150	0.05

*Auditory N-back level: Ignore, 2-Back.

2012) or other techniques (Parra et al., 2005). For example, a number of recent studies have utilized ICA for the identification and removal of EOG components (Nikolaev et al., 2011, 2013; Devillez et al., 2015). However, these studies typically included a manual or semi-manual step for the identification of ICs related to corneo-retinal potentials and eyelid artifacts (although see Mognon et al., 2011; Plöchl et al., 2012). In contrast, our approach uses equivalent dipole locations to include or exclude particular ICs. While there is ongoing debate as to the accuracy of source localization techniques such as LORETA, there is growing evidence that suggests independent sources are indeed dipolar (Delorme et al., 2012). Fortunately, eye movement related ICs typically explained large fractions of the total signal variance and resolve to equivalent dipoles outside the brain volume with relatively little residual error.

In comparison, IC clustering results (Makeig et al., 2002) are highly dependent on the choice of the clustering parameters (in many cases up to 12 tunable parameters without a clear physiological interpretation for each parameter) and provide no guarantees in terms of producing clusters at particular ROIs. However, the ROI-based measure projection approach
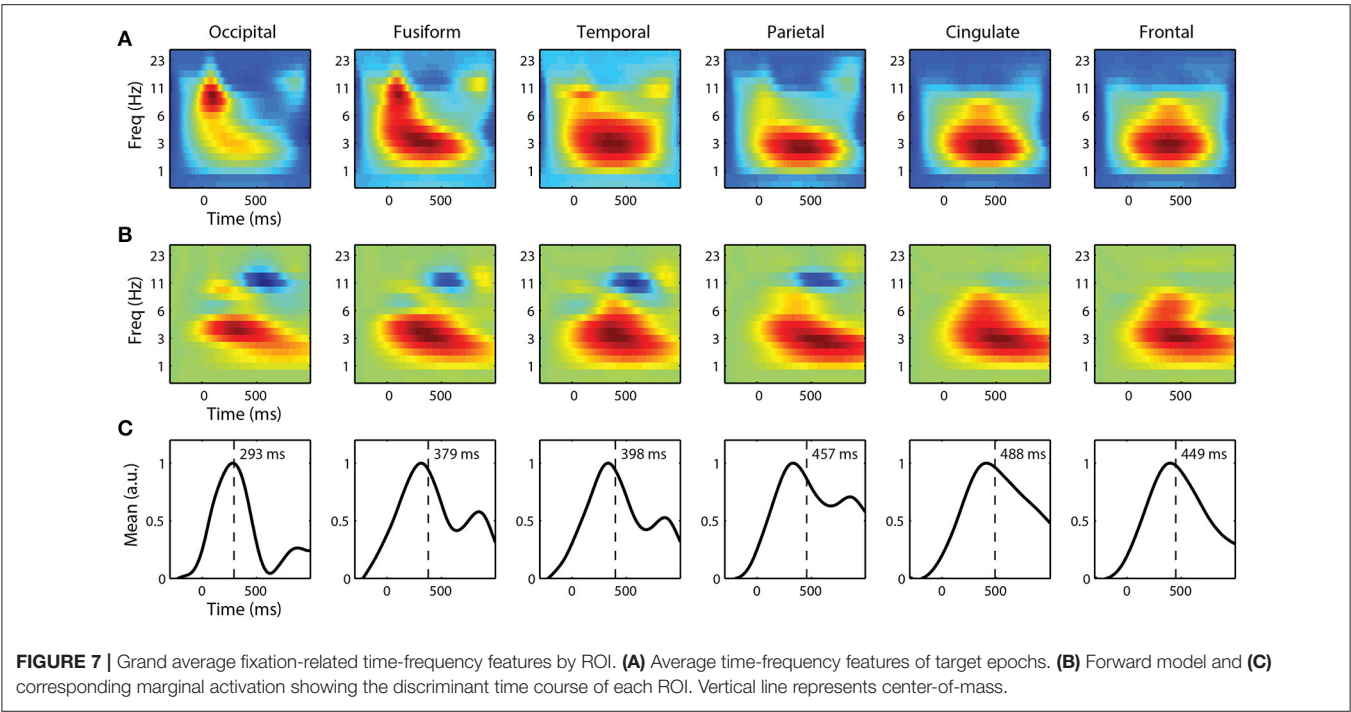
FIGURE 7 | Grand average fixation-related time-frequency features by ROI. **(A)** Average time-frequency features of target epochs. **(B)** Forward model and **(C)** corresponding marginal activation showing the discriminant time course of each ROI. Vertical line represents center-of-mass.

TABLE 4 | ANOVA statistics for classifier performance in the visual task.

Factor	df	F	p	η^2
ROI CLASSIFIER PERFORMANCE (Az)				
Condition*	2.41,156.73	12.84	< 0.001	0.15
ROI	2.40,124.59	1.89	0.147	0.03
Interaction	8.94,116.23	1.07	0.388	0.06
HIERARCHICAL CLASSIFIER PERFORMANCE (Az)				
Condition*	2.04,26.55	1.16	0.329	0.08
HIERARCHICAL CLASSIFIER SCORE (TARGET)				
Condition*	2.66,34.60	3.63	0.026	0.22

Auditory N-back level: Silent, Ignore, 0-Back, 1-Back, 2-Back.

(ROI-MPA) is able to focus the analysis on selected ROIs while utilizing only a single parameter with a physiological interpretation (i.e., the expected spatial uncertainty in IC dipole localization). Likewise, IC clusters are not well-suited for single-trial analysis. Simply averaging the single-trial activity of the ICs contained in each cluster would not properly account the spatial distribution of dipole locations. For example, ICs adjacent to the cluster boundary would be excluded (weighted zero) while ICs just inside the cluster weighted at unity. ROI-MPA directly incorporates this spatial information and the ROI structure by forming a weighted sum based on IC spatial probability overlap with each ROI.

Eye Tracking, Pupillometry, and EEG

A growing number of studies are combining eye tracking with EEG to enable the exploration for neural activity during visual search. While the task employed here was not a visual search
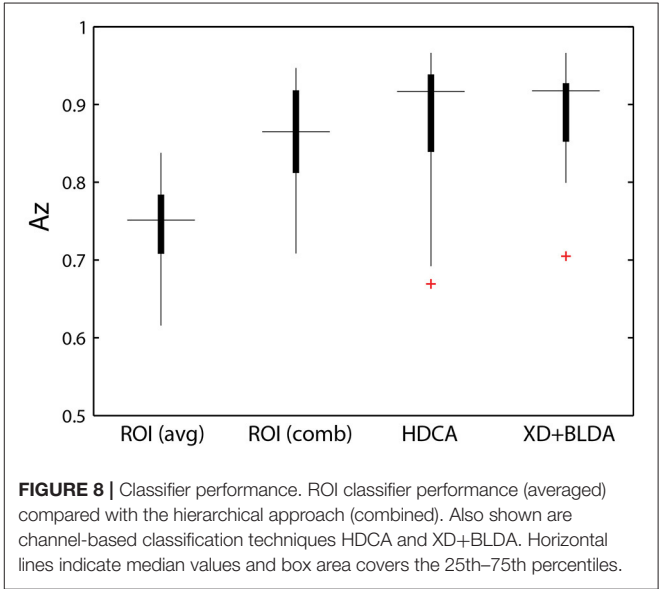
FIGURE 8 | Classifier performance. ROI classifier performance (averaged) compared with the hierarchical approach (combined). Also shown are channel-based classification techniques HDCA and XD+BLDA. Horizontal lines indicate median values and box area covers the 25th–75th percentiles.

paradigm, our results demonstrate the ability to acquire and utilize gaze position to detect saccades and quantify evoked neural activity. Importantly, our experimental configuration employed a head-free tracking system (SMI RED 250) without a requisite chinrest. This configuration facilitates visual search paradigms or related tasks requiring a large field of view that may naturally engender small head movements. To improve the spatial accuracy of such a system, our analysis included a *post-hoc* calibration of gaze position. Specifically, we utilized task information to infer gaze position when adjusting the offline
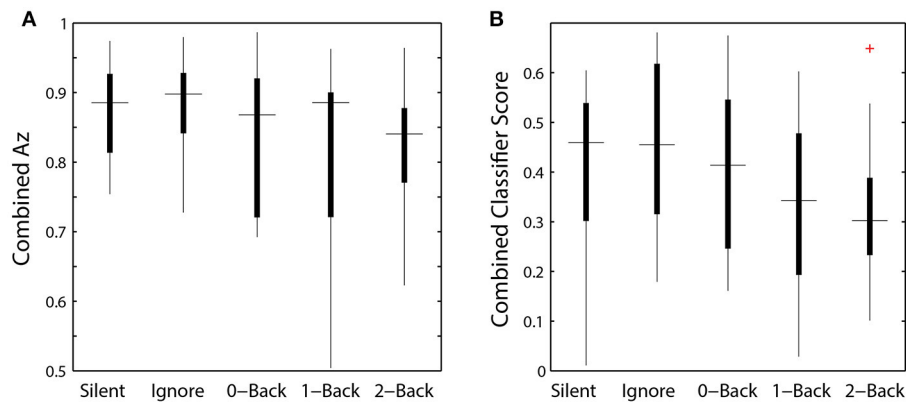
FIGURE 9 | Hierarchical classifier performance and task difficulty. **(A)** Classifier performance as a function of auditory task condition. **(B)** Classifier target scores as a function of auditory task condition. Horizontal lines indicate median values and box area covers the 25th–75th percentiles.

calibration model. While this type of information may not always be available, experimenters can and should use an opportunistic calibration approach during periods where gaze position can reasonably be inferred (e.g., prior to trial initiation or visual target detection).

An additional benefit derived from the inclusion of eye tracking is the coincident measure of pupil size. For example, the change in pupil diameter shown here suggests that increased working memory load resulted in an increase in arousal level, an important modulator of cognitive performance. Since arousal is largely regulated through the norepinephrine system via the locus coeruleus (LC), a nucleus within the dorsal pons, it cannot be measured directly via EEG. However, several studies have shown that pupil dilation can be used as a proxy for LC activity (Aston-Jones and Cohen, 2005; Murphy et al., 2011; Hong et al., 2014). Furthermore, the LC receives input from anterior cingulate and dorsolateral prefrontal cortex and some studies have suggested that the LC system underlies the parietal P3 ERP, specifically the P3b (Nieuwenhuis et al., 2005, 2011).

Overall, our results confirm the localization of the P3 to parietal ROIs and show a significant effect of arousal on both pupil diameter and P3 amplitude (Murphy et al., 2011). In addition, target FRPs in the posterior ROIs show a negative deflection, relative to non-target FRPs, beginning around 200 ms post-fixation. This difference is consistent with the visual mismatch negativity (vMMN); a negative posterior deflection elicited by an infrequent (deviant) visual stimulus presented in a homogenous sequence of frequent (standard) stimuli (Czigler et al., 2002). In particular, this difference is consistent with the later components of the vMMN associated with memory-comparison-based change detection (Kimura et al., 2009). However, since the infrequent stimuli (Ts) are also task-relevant, it is difficult to dissociate this vMMN from the attentional orienting component of the P3 (Polich, 2007).

Interestingly, the frontal ROI and Fz electrode showed a small but significant difference between target and non-target fixations at an early latency (approximately 80 ms). Previous studies have shown activity associated with peripheral detection in frontal-parietal regions early in and even prior to target

fixations (Dias et al., 2013; Devillez et al., 2015). In this paradigm, target stimuli were never immediately adjacent to the current grid location (red annulus), making the peripheral detection of an upcoming target unlikely. However, it would be reasonable for participants to anticipate a target fixation after a sequence of non-target stimuli were encountered. This phenomenon illustrates the manifold difficulties in the interpretation of eye movement related activity given the dependencies between eye movement behavior (e.g., saccade size, fixation duration) and the elicited response. Similarly, in free-viewing contexts there remains the additional challenge of separating overlapping responses from adjacent saccades and fixations. Fortunately, several techniques have now been proposed to address this potential confound using regression or GLM-based approaches (Burns et al., 2013; Smith and Kutas, 2015; Kristensen et al., 2017).

Model Considerations

The ICA-based, hierarchical classification algorithm described here is not ideally suited for real-time application or meant as an alternative to other channel-based approaches (e.g., HDCA). Rather, the goal of this study was to identify the discriminant neural response in each ROI and to quantify the effect of cognitive load on that response. As such, we did not take additional steps to separate data in our cross-validation scheme. Due to data limitations, ICA was applied to the entire EEG record for each participant rather than independently to each training set. Since ICA is an unsupervised technique, however, the potential for overfitting is limited. Additionally, we selected the hyperparameter by balancing the effective degrees of freedom with the number of data points (fixation epochs) rather than through a separate cross-validation step. Here, the exact choice of parameter did not substantially influence the results (see Supplementary Section 2). The added separation of these additional cross-validation steps would significantly reduce the amount of data and the quality of the estimated discriminate functions without providing any additional insight into the neural processes.

Implications for BCI

Importantly, our hierarchical classification scheme was shown to perform at a level similar to other state-of-the-art machine learning algorithms such as HDCA and XD-BLDA. This result suggests that our approach captured the majority of the task-related variance within the EEG record. While the average FRPs are useful and exhibit an effect of cognitive load, single-trial classification techniques can reveal additional discriminative activity (Brouwer et al., 2012). The forward model of each ROI (**Figure 7**) reveal both the time course and spectral characteristics of the discriminant neural response. Thus, while our approach imposes an additional computational burden, compared with the above methods, it adds insight into the source of task-related neural activity.

Our results can likewise be used to guide BCI development and future applications. P3-based paradigms remain a key component of the BCI application space, such as the P300 Speller (Krusienski et al., 2006). Presently, these reactive BCIs typically classify the neural response to passively viewed stimuli, such as in a rapid serial visual presentation (Gerson et al., 2005; Touryan et al., 2011; Bigdely-Shamlo et al., 2013). In this passive condition, stimuli are presented to the user who detects the desired target (e.g., target object within an image). In contrast, for targets occurring in natural or ordered environments, a more ecologically valid approach for detection would be through goal-directed visual search (Jangraw et al., 2014; Ušćumlić and Blankertz, 2016). In this case, stimulus presentation is controlled through the user's search strategy, with fixation onset serving as a natural time-locking event. Thus, the growing body of work on single-trial classification of FRPs will support the improved performance of future FRP BCI technology.

CONCLUSION AND FUTURE WORK

In this study, we provide a principled framework for interpreting EEG in the presence of eye movements and concurrent task demands by adapting a recently developed independent source aggregation technique (Bigdely-Shamlo et al., 2013). This approach enabled us to both quantify the discriminant information contained within each cortical region and measure the effect of cognitive load on the evoked response. While these phenomena have been previously observed, our results demonstrate the feasibility and utility of combining synchronous recordings of EEG and eye-tracking to measure both sensory and cognitive processes. Our approach can be extended to tasks that incorporate unconstrained eye movements, however, additional techniques would be needed to account for the overlapping activity from adjacent saccades and fixations.

In this experiment we did not explicitly manipulate top-down (goal-directed) or bottom-up (stimulus driven) components of the visual task beyond increasing the overall working memory load. However, our ROI mapping framework would be well suited for such an assessment. ROI analysis could be applied across conditions to identify what factors bias top-down vs. bottom-up neural activity in a visual search paradigm. Eye movements biased by top-down task influences may have greater pre- or post-saccadic activity in frontal cortices (Nikolaev et al., 2013). Likewise, eye movements driven by bottom-up stimulus influences may have greater post-saccadic activity in occipital cortex. This distinction may become essential for understanding free-viewing search in natural scenes where visual information leading to a detection event can be accumulated across fixations (Jangraw et al., 2014) rather than isolated to a single gaze position.

ETHICS STATEMENT

The voluntary, fully informed consent of the persons used in this research was obtained in written form. The document used to obtain informed consent was approved by the U.S. Army Research Laboratory's Institutional Review Board (IRB) in accordance with 32 CFR 219 and AR 70-25, and also in compliance with the Declaration of Helsinki. The study was reviewed and approved (approval# ARL 14-042) by the U.S. Army Research Laboratory's IRB before the study began.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnhum.2017.00357/full#supplementary-material

REFERENCES

Aston-Jones, G., and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* 28, 403–450. doi: 10.1146/annurev.neuro.28.061604.135709

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* 91, 276–292. doi: 10.1037/0033-2909.91.2.276

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 55, 289–300.

Bigdely-Shamlo, N., Mullen, T., Kreutz-Delgado, K., and Makeig, S. (2013). Measure projection analysis: a probabilistic approach to EEG source comparison and multi-subject inference. *Neuroimage* 72, 287–303. doi: 10.1016/j.neuroimage.2013.01.040

Brouwer, A.-M., Hogervorst, M. A., Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Brouwer, A.-M., Reuderink, B., Vincent, J., van Gerven, M. A. J., and van Erp, J. B. F. (2013). Distinguishing between target and nontarget fixations in a visual search task using fixation-related potentials. *J. Vis.* 13:17. doi: 10.1167/13.3.17

Burns, M. D., Bigdely-Shamlo, N., Smith, N. J., Kreutz-Delgado, K., and Makeig, S. (2013). "Comparison of averaging and regression techniques for estimating Event Related Potentials," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka), 1680–1683.

Cecotti, H., Rivet, B., Congedo, M., Jutten, C., Bertrand, O., Maby, E., et al. (2011). A robust sensor-selection method for P300 brain–computer interfaces. *J. Neural Eng.* 8:016001. doi: 10.1088/1741-2560/8/1/016001

Czigler, I., Balázs, L., and Winkler, I. (2002). Memory-based detection of task-irrelevant visual changes. *Psychophysiology* 39, 869–873. doi: 10.1111/1469-8986.3960869

Dandekar, S., Ding, J., Privitera, C., Carney, T., and Klein, S. A. (2012a). The fixation and saccade P3. *PLoS ONE* 7:e48761. doi: 10.1371/journal.pone.0048761

Dandekar, S., Privitera, C., Carney, T., and Klein, S. A. (2012b). Neural saccadic response estimation during natural viewing. *J. Neurophysiol.* 107, 1776–1790. doi: 10.1152/jn.00237.2011

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Delorme, A., Palmer, J., Onton, J., Oostenveld, R., and Makeig, S. (2012). Independent EEG sources are dipolar. *PLoS ONE* 7:e30135. doi: 10.1371/journal.pone.0030135

Devillez, H., Guyader, N., and Guérin-Dugué, A. (2015). An eye fixation–related potentials analysis of the P300 potential for fixations onto a target object when exploring natural scenes. *J. Vis.* 15, 20–20. doi: 10.1167/15.13.20

Dias, J. C., Sajda, P., Dmochowski, J. P., and Parra, L. C. (2013). EEG precursors of detected and missed targets during free-viewing search. *J. Vis.* 13, 13–13. doi: 10.1167/13.13.13

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., and Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: analyses and review. *J. Exp. Psychol. Gen.* 140, 552–572. doi: 10.1037/a0023885

Engbert, R., and Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7192–7197. doi: 10.1073/pnas.0509557103

Gaarder, K., Krauskopf, J., Graf, V., Kropfl, W., and Armington, J. C. (1964). Averaged brain activity following saccadic eye movement. *Science* 146, 1481–1483.

Gerson, A. D., Parra, L. C., and Sajda, P. (2005). Cortical origins of response time variability during rapid discrimination of visual objects. *Neuroimage* 28, 342–353. doi: 10.1016/j.neuroimage.2005.06.026

Gerson, A. D., Parra, L. C., and Sajda, P. (2006). Cortically coupled computer vision for rapid image search. *IEEE Trans. Neural Syst. Rehabil. Eng.* 14, 174–179. doi: 10.1109/TNSRE.2006.875550

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Hong, L., Walz, J. M., and Sajda, P. (2014). Your eyes give you away: prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. *PLoS ONE* 9:e91321. doi: 10.1371/journal.pone.0091321

Jangraw, D. C., Wang, J., Lance, B. J., Chang, S.-F., and Sajda, P. (2014). Neurally and ocularly informed graph-based models for searching 3D environments. *J. Neural Eng.* 11:046003. doi: 10.1088/1741-2560/11/4/046003

Kahneman, D., and Beatty, J. (1966). Pupil diameter and load on memory. *Science* 154, 1583–1585.

Kamienkowski, J. E., Ison, M. J., Quiroga, R. Q., and Sigman, M. (2012). Fixation-related potentials in visual search: a combined EEG and eye tracking study. *J. Vis.* 12:4. doi: 10.1167/12.7.4

Kaunitz, L. N., Kamienkowski, J. E., Varatharajah, A., Sigman, M., Quiroga, R. Q., and Ison, M. J. (2014). Looking for a face in the crowd: fixation-related potentials in an eye-movement visual search task. *Neuroimage* 89, 297–305. doi: 10.1016/j.neuroimage.2013.12.006

Kazai, K., and Yagi, A. (1999). Integrated effect of stimulation at fixation points on EFRP (eye-fixation related brain potentials). *Int. J. Psychophysiol.* 32, 193–203. doi: 10.1016/S0167-8760(99)00010-0

Kimura, M., Katayama, J., Ohira, H., and Schröger, E. (2009). Visual mismatch negativity: new evidence from the equiprobable paradigm. *Psychophysiology* 46, 402–409. doi: 10.1111/j.1469-8986.2008.00767.x

Kowler, E. (2011). Eye movements: the past 25 years. *Vision Res.* 51, 1457–1483. doi: 10.1016/j.visres.2010.12.014

Kristensen, E., Guerin-Dugué, A., and Rivet, B. (2017). Regularization and a general linear model for event-related potential estimation. *Behav. Res. Methods* doi: 10.3758/s13428-017-0856-z. [Epub ahead of print].

Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., et al. (2006). A comparison of classification techniques for the P300 Speller. *J. Neural Eng.* 3, 299–305. doi: 10.1088/1741-2560/3/4/007

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., et al. (2002). Dynamic brain sources of visual evoked responses. *Science* 295, 690–694. doi: 10.1126/science.1066168

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48, 229–240. doi: 10.1111/j.1469-8986.2010.01061.x

Murphy, P. R., Robertson, I. H., Balsters, J. H., and O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology* 48, 1532–1543. doi: 10.1111/j.1469-8986.2011.01226.x

Nieuwenhuis, S., Aston-Jones, G., and Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychol. Bull.* 131:510. doi: 10.1037/0033-2909.131.4.510

Nieuwenhuis, S., De Geus, E. J., and Aston-Jones, G. (2011). The anatomical and functional relationship between the P3 and autonomic components of the orienting response. *Psychophysiology* 48, 162–175. doi: 10.1111/j.1469-8986.2010.01057.x

Nikolaev, A. R., Jurica, P., Nakatani, C., Plomp, G., and van Leeuwen, C. (2013). Visual encoding and fixation target selection in free viewing: presaccadic brain potentials. *Front. Syst. Neurosci.* 7:26. doi: 10.3389/fnsys.2013.00026

Nikolaev, A. R., Meghanathan, R. N., and van Leeuwen, C. (2016). Combining EEG and eye movement recording in free viewing: pitfalls and possibilities. *Brain Cogn.* 107, 55–83. doi: 10.1016/j.bandc.2016.06.004

Nikolaev, A. R., Nakatani, C., Plomp, G., Jurica, P., and van Leeuwen, C. (2011). Eye fixation-related potentials in free viewing identify encoding failures in change detection. *Neuroimage* 56, 1598–1607. doi: 10.1016/j.neuroimage.2011.03.021

Parra, L. C., Spence, C. D., Gerson, A. D., and Sajda, P. (2005). Recipes for the linear analysis of EEG. *Neuroimage* 28, 326–341. doi: 10.1016/j.neuroimage.2005.05.032

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *Int. J. Psychophysiol.* 18, 49–65. doi: 10.1016/0167-8760(84)90014-X

Plöchl, M., Ossandón, J. P., and König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Front. Hum. Neurosci.* 6:278. doi: 10.3389/fnhum.2012.00278

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Ries, A. J., Touryan, J., Ahrens, B., and Connolly, P. (2016). The impact of task demands on fixation-related brain potentials during guided search. *PLoS ONE* 11:e0157260. doi: 10.1371/journal.pone.0157260

Rivet, B., Souloumiac, A., Attina, V., and Gibert, G. (2009). xDAWN algorithm to enhance evoked potentials: application to brain-computer interface. *IEEE Trans. Biomed. Eng.* 56, 2035–2043. doi: 10.1109/TBME.2009.2012869

Scherg, M. (1990). "Fundamentals of dipole source potential analysis," in *Auditory Evoked Magnetic Fields and Electric Potentials,* eds F. Grandori, M. Hoke, G. L. Romani (Basel: Karger), 40–69.

Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., et al. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* 39, 1064–1080. doi: 10.1016/j.neuroimage.2007.09.031

Smith, N. J., and Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology* 52, 157–168. doi: 10.1111/psyp.12317

Thickbroom, G. W., Knezevič, W., Carroll, W. M., and Mastaglia, F. L. (1991). Saccade onset and offset lambda waves: relation to pattern movement visually evoked potentials. *Brain Res.* 551, 150–156. doi: 10.1016/0006-8993(91)90927-N

Torrence, C., and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* 79, 61–78.

Touryan, J., Gibson, L., Horne, J. H., and Weber, P. (2011). Real-time measurement of face recognition in rapid serial visual presentation. *Front. Psychol.* 2:42. doi: 10.3389/fpsyg.2011.00042

Touryan, J., Lance, B. J., Kerick, S. E., Ries, A. J., and McDowell, K. (2016). Common EEG features for behavioral estimation in disparate, real-world tasks. *Biol. Psychol.* 114, 93–107. doi: 10.1016/j.biopsycho.2015.12.009

Ušćumlić, M., and Blankertz, B. (2016). Active visual search in non-stationary scenes: coping with temporal variability and uncertainty. *J. Neural Eng.* 13:016015. doi: 10.1088/1741-2560/13/1/016015

Wenzel, M. A., Golenia, J.-E., and Blankertz, B. (2016). Classification of eye fixation related potentials for variable stimulus saliency. *Front. Neurosci.* 10:23. doi: 10.3389/fnins.2016.00023

Assessing the Depth of Cognitive Processing as the Basis for Potential User-State Adaptation

Irina-Emilia Nicolae[1,2]*, Laura Acqualagna[2] and Benjamin Blankertz[2]*

[1] Department of Applied Electronics and Information Engineering, Politehnica University of Bucharest, Bucharest, Romania,
[2] Department of Neurotechnology, Technische Universität Berlin, Berlin, Germany

Objective: Decoding neurocognitive processes on a single-trial basis with Brain-Computer Interface (BCI) techniques can reveal the user's internal interpretation of the current situation. Such information can potentially be exploited to make devices and interfaces more user aware. In this line of research, we took a further step by studying neural correlates of different levels of cognitive processes and developing a method that allows to quantify how deeply presented information is processed in the brain.

Methods/Approach: Seventeen participants took part in an EEG study in which we evaluated different levels of cognitive processing (no processing, shallow, and deep processing) within three distinct domains (memory, language, and visual imagination). Our investigations showed gradual differences in the amplitudes of event-related potentials (ERPs) and in the extend and duration of event-related desynchronization (ERD) which both correlate with task difficulty. We performed multi-modal classification to map the measured correlates of neurocognitive processing to the corresponding level of processing.

Results: Successful classification of the neural components was achieved, which reflects the level of cognitive processing performed by the participants. The results show performances above chance level for each participant and a mean performance of 70–90% for all conditions and classification pairs.

Significance: The successful estimation of the level of cognition on a single-trial basis supports the feasibility of user-state adaptation based on ongoing neural activity. There is a variety of potential use cases such as: a user-friendly adaptive design of an interface or the development of assistance systems in safety critical workplaces.

Keywords: encephalography (EEG), event-related potentials (ERPs), oscillatory activity, cognitive processes, single-trial classification

INTRODUCTION

While Brain-Computer Interface (BCI) research primarily targets medical applications (Birbaumer, 2006; Dornhege et al., 2007; van Gerven et al., 2009; Wolpaw and Wolpaw, 2012; Guger et al., 2014; Hassanien and Taher Azar, 2015), more and more perspectives are being explored that go beyond communication and control paradigms (Blankertz et al., 2010, 2015, 2016; Zander and Kothe, 2011; Borghini et al., 2014; Zander et al., 2016). One of those perspectives are systems that take the

ongoing user mental state into account and automatically adapt according to the user's mindset (Müller et al., 2008; van Erp et al., 2015) exploiting implicit information (Gamberini et al., 2015).

Neural correlates of cognitive processes can be found in the time-locked Event-Related Potentials (ERPs), usually composed of different components, and in modulations of spontaneous brain rhythms. The phase of those background rhythms is not time-locked to external stimuli, but the modulation of their amplitudes (resp. their hull curves) can be time-locked. In that case, the effect is called Event-Related Desynchronization or Synchronization (ERD/ERS) (Pfurtscheller and Lopes da Silva, 1999; Lemm et al., 2009), depending on whether it is a decrease or an increase in spectral power of the given frequency band.

ERP components have been widely investigated and previous studies involving cognitive activities in oddball paradigms show that the amplitude and latency of the ERPs are modified according to task difficulty (Donchin et al., 1973; Ullsperger et al., 1987; Polich, 2007; Kim et al., 2008). Specifically, increased P300 amplitude and longer latencies are usually found, relating to more complex processes and stronger attentional demand. These cognitive ERP components are reported to appear between 300–500 ms after the stimuli, in the centro-parietal cortex (Polich, 2007).

Moreover, changes in frequency are also reported, being modulated by the difficulty of cognitive processes. Firstly, the presentation of a stimulus triggers a short synchronization proceeded by a prolonged desynchronization mostly occurring in the α band at the frontal, temporal, central, and parietal locations of the scalp in accordance with the type of cognitive process (Klimesch et al., 1992, 1993, 1997; Klimesch, 1999), and follows the P300 potential (Yordanova et al., 2001). The α band desynchronizes during mental activity and cognitive judgment (Klimesch, 1999), and is reported proportionally increasing with more difficult cognitive processing, change mostly encountered at the centro-parietal sites. The effect is visible in laboratory environments (Gevins et al., 1997) and as well in more realistic scenarios (Venthur et al., 2010).

In addition, β oscillations are also linked to complex cognitive processes (Pesonen et al., 2007; Okazaki et al., 2008; Sheth et al., 2009), whereas decreased β oscillations appearing at the central and parietal sites are correlated to complex reasoning (Basile et al., 2013), decision making (Nakata et al., 2013) and are also related to the transition of cognitive states (Sheth et al., 2009).

Changes in the θ activity are observed as synchronizations in relation to task difficulty (Klimesch, 1999), e.g., increased θ power proportionally with increased memory load (Gundel and Wilson, 1992; Gevins et al., 1997), which also relate to the encoding of new information (Klimesch et al., 1996; Klimesch, 1999). Regarding localization, the θ changes are known to appear at the frontal midline scalp location (Gevins et al., 1997).

A pronounced ERD was found in relation to different types of cognitive processes. For example, in memory processes (Mecklinger et al., 1992; Klimesch et al., 1994; Stipacek et al., 2003; Pesonen et al., 2007), perceptual encoding and attentional processes, a stronger ERD is observed in the α band (Sergeant et al., 1987; Klimesch, 2003; Schack et al., 2005; Polich, 2007).

In addition, the processing of semantic information, e.g., words, shows likewise ERD enhancements (Klimesch et al., 1997).

In a recent study, Naumann et al. (2017), investigated gradual differences in task difficulty by estimating the difficulty level of a video game from the ongoing neural activity of the user. They found likewise significant modulations in the θ (4–7 Hz) and α (8–13 Hz) frequency bands, associated with changes in task difficulty.

In the present work, we have studied the feasibility of quantifying how deeply presented information is processed in the brain by tapping the corresponding components of brain activity. To that end, we analyzed single-trial EEG data with respect to its discriminative value of ERPs and of modulations of brain rhythms. In the ERPs, we found as main effect an increase of the P300 amplitude with task demand and also domain specific modulations in later components, see also (Nicolae et al., 2015a). Spontaneous background oscillations showed a prolonged suppression of the α and β rhythm as a reflection of profound cognitive processing. For extracting the features for classification, we employed spatio-temporal features of the ERPs (Blankertz et al., 2011) and we exploited the ERD effect by combining spatio-spectral decomposition (SSD, Nikulin et al., 2011) in two frequency bands with Common Spatial Patterns (CSP) analysis (Fukunaga, 1990; Koles, 1991). Combining features of ERPs and ERDs in a multimodal classification approach (Fazli et al., 2015) leads to a performance increase compared to using a single modality only (Nicolae et al., 2015b). The current manuscript comprises a detailed investigation of the depth of cognitive processing integrating the analysis over the modulations of brain rhythms (ERDs) and the spectral analysis discriminability (power spectrum, CSPs), extending previous work in Nicolae et al. (2015a), where the temporal and spatial evolution of the ERPs are briefly analyzed and the abstract in Nicolae et al. (2016), where a short overview over the power spectrum is presented. In addition, the present work contributes further with an enhanced multivariate classification approach based on both, the temporal and spectral features extracted from the EEG, as compared to the classification based on only the temporal features presented in (Nicolae et al., 2015b).

MATERIALS AND METHODS

Experimental Setup

Participants

Fifteen healthy participants with no acute or chronic neurological and/or psychiatric disorders and no pregnant women were considered for the study. Eleven participants were right-handed, ten were males and all were aged between 22 and 35 years old. Eleven participants had German as mother tongue, one participant had English and the others had different languages as native tongue, with a required good command of English or German in order to fulfill the task in the language condition. Seventeen participants were initially recorded, but two participant's data were removed for the analysis due to high artifacts probably caused by improper recordings. The experimental procedure was conducted in accordance with the declaration of Helsinki, approved by the ethics committee of the

Department of Psychology and Ergonomics of the Technische Universität Berlin and written informed consent was obtained from each participant. To countenance participant's motivation, they received financial compensation.

Material

The hardware equipment used for the acquisition of the Electroencephalography (EEG) was a BrainAmp amplifier with 64 active electrodes (Brain Products GmbH, Munich, Germany) positioned according to the 10–20 international system. One electrode, named *EOG*, was placed under the left eye and used for eye movements recording. We used unipolar recording at a sampling frequency of 1 kHz with the ground placed on the scalp at position AFz and with reference at left mastoid. The acquisition system was re-referenced to left and right mastoids. When mounting the electrode cap, the impedance was kept below 20 kΩ. The acquired data will be made available from the DepositOnce repository of Technische Universität Berlin (https://depositonce.tu-berlin.de/).

The stimuli were designed as vectorial graphics with the *Inkscape* software (version 0.91.0.1 https://inkscape.org) besides the images from the animals category which were created by *Freepik* and taken from the *Freepik* database (http://www.freepik.com/free-photos-vectors/icon). The stimuli were presented on a 24" display with 60 Hz refresh rate and 1,920 × 1,200 resolution (Dell U2410). For developing the experimental paradigm, the *Processing* software (version 3.0a4 https://processing.org/) was used in conjunction with MATLAB software for signal acquisition (release R2014a, The MathWorks, Inc., Natick, MA, USA). For the offline analysis, the data was processed with the BBCI MATLAB Toolbox (https://github.com/bbci/bbci_public).

Experimental Scenario

We considered two degrees of cognitive processing, namely shallow and deep processing levels (Craik and Lockhart, 1972) in a visual stimuli paradigm. In our scenario, the shallow processing involves a basic information processing revealed by attention (color appearance) and deep processing requires a complex activity related to specific cognitive tasks in the *memory*, *language*, and *visual imagination* conditions (Ganis et al., 2013). Each visual stimulus consisted of a pair of two images, cartoon-like drawings and every image represented an object chosen out of three categories: animals, fruits and mobility and was represented in one of the colors: red, green, blue, or magenta. Both images had same color and same category, where each category consisted of a total of 10 objects, see **Figure 1**. In order to maintain the desired ratio and without increasing too much the difficulty of the tasks, only two out of the three categories were chosen for each run.

The experimental study took place in a laboratory environment. Participants performed the experiments seated, and were requested to stay still, relaxed, and focused. They were allowed to freely explore the information presented, but in general, to focus their view in the center of the screen. Each participant was preliminary prepared with a practice test (1–3 runs) in order to become familiar with the tasks.

The experiment paradigm is described below. Before the start of the sequence, and after a short personal current state evaluation, the condition to be performed was displayed along with the target image pair (target color and category). When the sequence started, participants had to distinguish first between color, and subsequently for category. If the color does not match the target, then no processing at all was requested (**Non-Target, NT** case). If only the color matched, perform only mental computation (**Shallow Target, ST** case). If color and category matched the target (**Deep Target, DT** case), then evaluate the requested cognitive task and perform the corresponding mental computations (addition). Following the example sketched in **Figure 2**, the procedure was as follows:

(1) Target pair (cue): remember the target category, target color and the target images;

(2) For each stimulus:

○ First, check if the stimulus color matches the target *color*:

→ if not, do nothing (**NT**—non-target);
→ if yes, count +1 and:

■ check if the category matches the target *category*:

→ if not, do nothing (**ST**—shallow target);
→ if yes, perform the *cognitive task* associated to each cognitive process and memorize the new images for the next trial in case of the *memory condition*:

● if the answer is negative, do nothing additional (**DT**—deep target);
● or in case of positive answer, additionally count +10 (**DT**—deep target);

The resulting final number was inserted at the end of each run, followed by the feedback regarding the correct number.

The cognitive *conditions* were separately conducted in this order: *memory*, *language*, and *visual imagination*, with five runs per condition. We did not alternate between the conditions after each run in order to avoid confusions between the tasks, which have been quite demanding. In each *condition* participants had to "answer" a yes-no question, which is detailed hereunder.

Memory

The first *condition* considers memory retrieval by comparing a previously presented stimulus with the current stimulus (Kirchner, 1958; Chen et al., 2008). Specifically, the question is whether the current stimulus coincides with previous *target pair* (last pair of *target color* and *target category*). The accomplishment implies memorization of the current pair and retrieval of the previous target part. For an example see **Figure 2**, left column. An additional example of the *memory* experimental paradigm is provided in **Supplementary Video 1** of the Supplementary Presentation 1.

Language

The *language* task considered comparisons based on phonemic representations between the words that represent the images. The task was to decide whether the number of syllables of
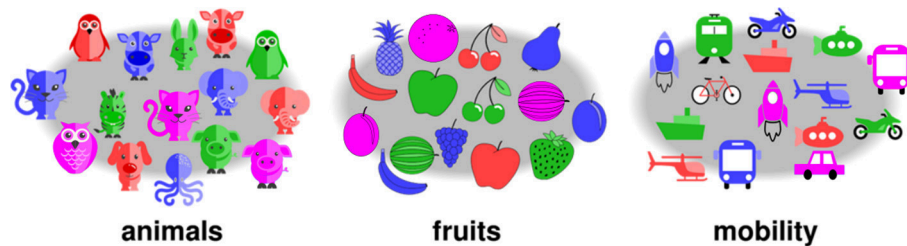
FIGURE 1 | Stimuli categories (animals, fruits, mobility) and objects representations examples. Each object could be presented in one of the four colors (red/green/blue/magenta).
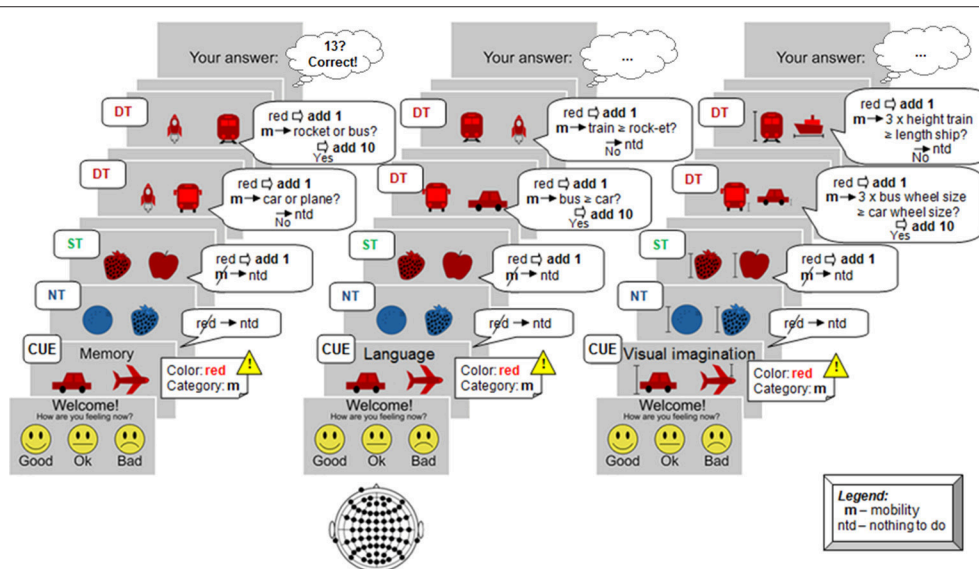
FIGURE 2 | Example of experimental protocol. At each stimulus participants had to perform or not a specific task, involving mental computation as a result. The total number counted was inserted at the end of the run. From left to right, the cognitive processes are illustrated here: *memory* case (recall *memory* and compare with the last target pair), *language* (compare the number of syllables), and *visual imagination* (imagine the objects in reality and perform dimensions comparison).

the left image's word was greater or equal than the number of syllables of the word for the object on the right side. This *condition* considered English or German words based on the participant's native *language*. Examples are shown in **Figure 2**, middle column. Note that the chosen objects (**Figure 1**) have a quite unique mapping to their representing words. In case a test person employed a different word, that could just affect the behavioral data, while the actual performance (DT vs. ST or NT) is not affected. Additional example for the *language* experimental paradigm is encountered in **Supplementary Video 2** of the Supplementary Presentation 1.

Visual imagination

The *visual imagination* task required mental representations in order to perform a comparison based on the size in reality. The differentiations made are accomplished by judging whether three times the dimension of the left object (or a part of the object) is greater or equal to the right object's size, considering average dimensions of the represented objects. The respective part of the object and the dimension type (length, height, or thickness) used
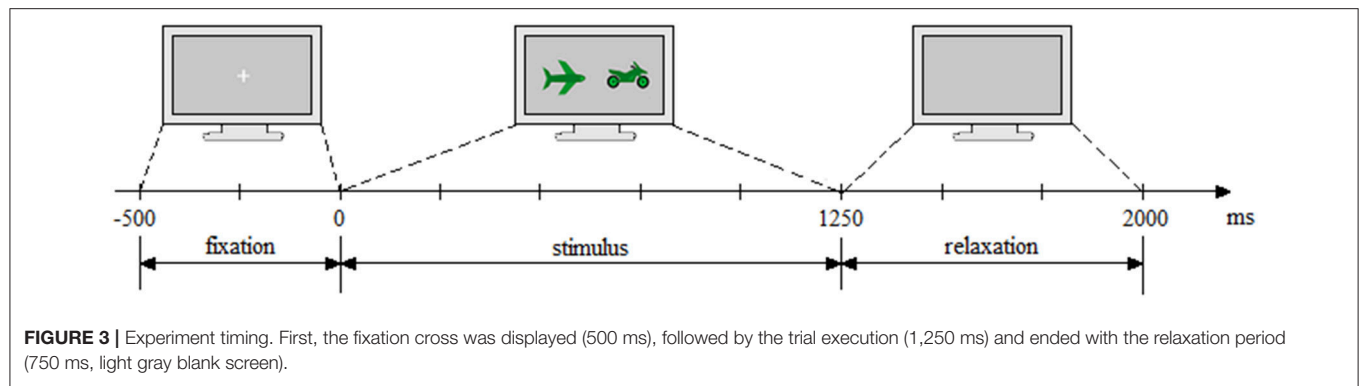
for comparison, was emphasized with a marker on the stimulus image. Further examples are in **Figure 2**, right column.

When generating the stimulus pairs for the *visual imagination condition*, a small constraint was added in order to ensure similar complexity within the *visual imagination* task. This related to the differences of dimensions within a range, more exactly the absolute difference between three times the left object and the right object should be less or equal than the left object dimension. Hence, no big discrepancy between the objects sizes was assured: no large differences implying an easy comparison, and no small ones either, which would result in a hard or ambiguous comparison. Short additional example for the *visual imagination* experimental paradigm is encountered in **Supplementary Video 3** of the Supplementary Presentation 1.

Experimental Design

As a pilot study, four participants were asked to test the application without EEG cap. The results of this pilot study have been used to calibrate the speed and the complexity of the task. The time course of the experiment is presented in **Figure 3**. It

FIGURE 3 | Experiment timing. First, the fixation cross was displayed (500 ms), followed by the trial execution (1,250 ms) and ended with the relaxation period (750 ms, light gray blank screen).

consisted of 2,500 ms Inter-Stimuli Interval and was divided as follows: 500 ms fixation, 1,250 ms stimulus presentation, and 750 ms relaxation period.

Each condition was evaluated in five runs with a total of 600 stimuli. The two images representing the stimulus, were scaled to a common 480 × 480 resolution and presented close to the center of the screen at a distance of 2" from one another on a light gray background.

The same percentage of shallow and deep targets was chosen in order to avoid confounds caused by a different number of occurrences: 75 ± 2% for non-targets (NT) and 12.5 ± 2% for both, shallow targets (ST) and deep targets (DT).

Behavioral Assessment

The subjects behavioral responses were assessed by subjective and objective indicators. The subjective indicators that we took into account by means of a questionnaire are: personal feedback, meaning personal internal state/mood (good/ok/bad) and personal overview for the difficulty of the conditions, which was acquired by scoring (0—easy; 1—medium; 2—hard). The objective indicator that we integrated (subjects answers ratio), is given by the actual responses, assessed by a ratio of the absolute difference between the correct number and user response number, divided by the correct number.

Data Analysis

Hereunder, the signal processing and machine learning methods are described, mainly the multi-modal analysis and multivariate classification which give us complementary and additional information over the neurophysiological effects of the cognitive processes. The multi-modal analysis investigates the neural correlates from the temporal and spectral domain: ERPs and Event-Related (De)Synchronization, ERDs. The brain sources of neural oscillations which substantiate cognitive activity (Varela et al., 2001; Buzsáki and Draguhn, 2004) were extracted by advanced decomposition methods: SSD by Nikulin et al. (2011) and Common Spatial Pattern (CSP) (Fukunaga, 1990). The two different types of neurophysiological information, temporal, and oscillatory activity, are combined with the concatenation approach as described in Dornhege et al. (2004), in order to give better performance as shown for example in other studies (Mühl et al., 2014). Finally, the depth of cognitive processing of

the external information is estimated using multivariate single-trial classification, by Regularized Linear Discriminant Analysis (Friedman, 1989).

Filtering and Epochs Rejection

For dimensionality reduction, the data was downsampled to 100 Hz. Preliminary processing of the data was performed by a sequence of low-pass filtering for anti-aliasing and high-pass filtering for drifts removal. The low-pass filter design for 42 Hz, is a Chebyshev type II of order 10 with 42 Hz pass-band edge frequency and 3 dB ripple, and a 49 Hz stopband with 50 dB attenuation. For high-pass filtering, a 1 Hz FIR filter of order 300 (three times the downsampling frequency) was applied, using least-squares error minimization and reverse digital filtering with zero-phase effect (For future online classification, appropriate causal filters must be considered). Following, the data was segmented into epochs considering the experiment timing detailed in **Figure 3**.

A rough pre-cleaning of the data was additionally performed obviating noisy channels and epochs. A criterion evaluated on band-pass filtered data in a broad frequency band (5–40 Hz) was applied to remove the channels dropping to zero. Particularly, the channels with variance smaller than 0.5 μV^2 in more than 10% of the trials were removed. Moreover, the epochs with muscle artifacts were also removed considering the trials with excessive variance in 20% of the channels. For strong eye movements artifacts, max-min criterion was applied and epochs with more than 150 μV difference between maximum and minimum voltage in channels F9, F10, AF3, and AF4 were removed. For removing the background noise of the remaining epochs, baseline correction was performed trial-wise by subtracting the mean amplitude computed on 100 ms of the pre-stimulus trial period from each trial period time point.

Artifact Removal

As a next step for removing the artifact data, including smaller eye movement artifacts, muscular artifacts and loose electrodes we performed Independent Component Analysis (ICA) with artifactual components selection given by the Multiple Artifact Rejection Algorithm (MARA, Winkler et al., 2011). For further verification, visual inspection was performed over each component considering the power spectral density and its topographic distribution.

Univariate Discriminative Analysis

In addition to the temporal ERP analysis in Nicolae et al. (2015a), we evaluated the differences among levels of cognitive processing in the spectral domain (Nicolae et al., 2016). Moreover, we investigated the power spectrum with a discriminative measure given by the signed and squared point biserial correlation coefficient (signed r^2) which quantifies the discriminability between the two classes and the time course of power modulations with the help of Event-Related De/Synchronization (ERD/ERS) curves (Pfurtscheller and Lopes da Silva, 1999) in selected frequency bands.

Multivariate Classification and Validation

In order to join complementary information about the neural activity and therefore improve single-trial classification, we combined two different types of neurophysiological information. We considered spatio-temporal features reflecting ERPs and oscillatory activity features, which are described hereunder.

Spatio-temporal features

The spatio-temporal features (channels and time) were extracted as in Nicolae et al. (2015a). The method, further described in Blankertz et al. (2011), detects five temporal windows for each participant based on a heuristic selection of the intervals with maximum discriminability and a constant pattern between two classes based on the signed and squared point biserial correlation coefficient (signed r^2). The selected intervals contain the most significant spatio-temporal features, effective as found in other studies (e.g., Acqualagna and Blankertz, 2013).

Spatio-spectral features

As the cognitive processes produce observable modulations in the oscillatory activity, we considered extracting this information from the respective frequency bands. Based on the discriminative analysis, we selected the most significant frequency bands: α (8–14 Hz) and β (16–20 Hz). To enhance the discrimination of activity in the frequency band of interest, we performed SSD (Nikulin et al., 2011) and Common Spatial Pattern (CSP) (Blankertz et al., 2008) for the corresponding frequency bands.

Spatio-spectral decomposition, SSD.

Linear spatial filtering prior to CSP was used in order to reach an efficient differentiation of mental states depicted by ERD/ERS rhythm patterns. Neural activity can be sometimes concealed in the background noise fluctuations, and therefore, for a better discrimination of the neural oscillations regarding cognitive processing, we applied SSD (Nikulin et al., 2011) which enhances the signal-to-noise ratio. In order to extract individual oscillatory sources, SSD finds the optimal spatial filters based on a generalized eigenvalue decomposition, that relates to high band power in the frequency of interest (pass-band filter) and low band power in the noise in the adjacent frequencies. The adjacent noise frequencies reduction is obtained with two pass band filters of a desired width (e.g., 1 or 2 Hz) below and, respectively, above the frequency of interest, before or after a gap (stop-band filter of e.g., 1 Hz) just below, respectively, just above the frequency of interest. Important notice: because SSD requires frequency filtering in advance, we used continuous data to avoid filter edges artifacts.

For the SSD decomposition, we consider only components with eigenvalues higher than 10^{-6} times the highest eigenvalue (see low-rank factorization in Haufe et al., 2014). Typically, between 15 and 35 components per discrimination pair were further selected.

Multi-band Common Spatial Patterns, mCSP.

In order to obtain discriminative information about the cognitive processes based on oscillatory activity, we make use of the widely used CSP method as described by Fukunaga (1990) and Koles (1991), which was successively applied in a similar context (Schultze-Kraft et al., 2016). CSP facilitates the binary discrimination of different brain states by spatial filtering, enhancing the signal of interest while suppressing the background activity, by maximizing the variance for one class whilst minimizing the variance for the other class and vice versa. In our case, it increases the variance of a higher level of cognitive processing while diminishes the variance of a lower level of processing and vice versa. The components reaching this goal were automatically selected (as in Blankertz et al., 2008) up to a maximum of three spatial filters per class.
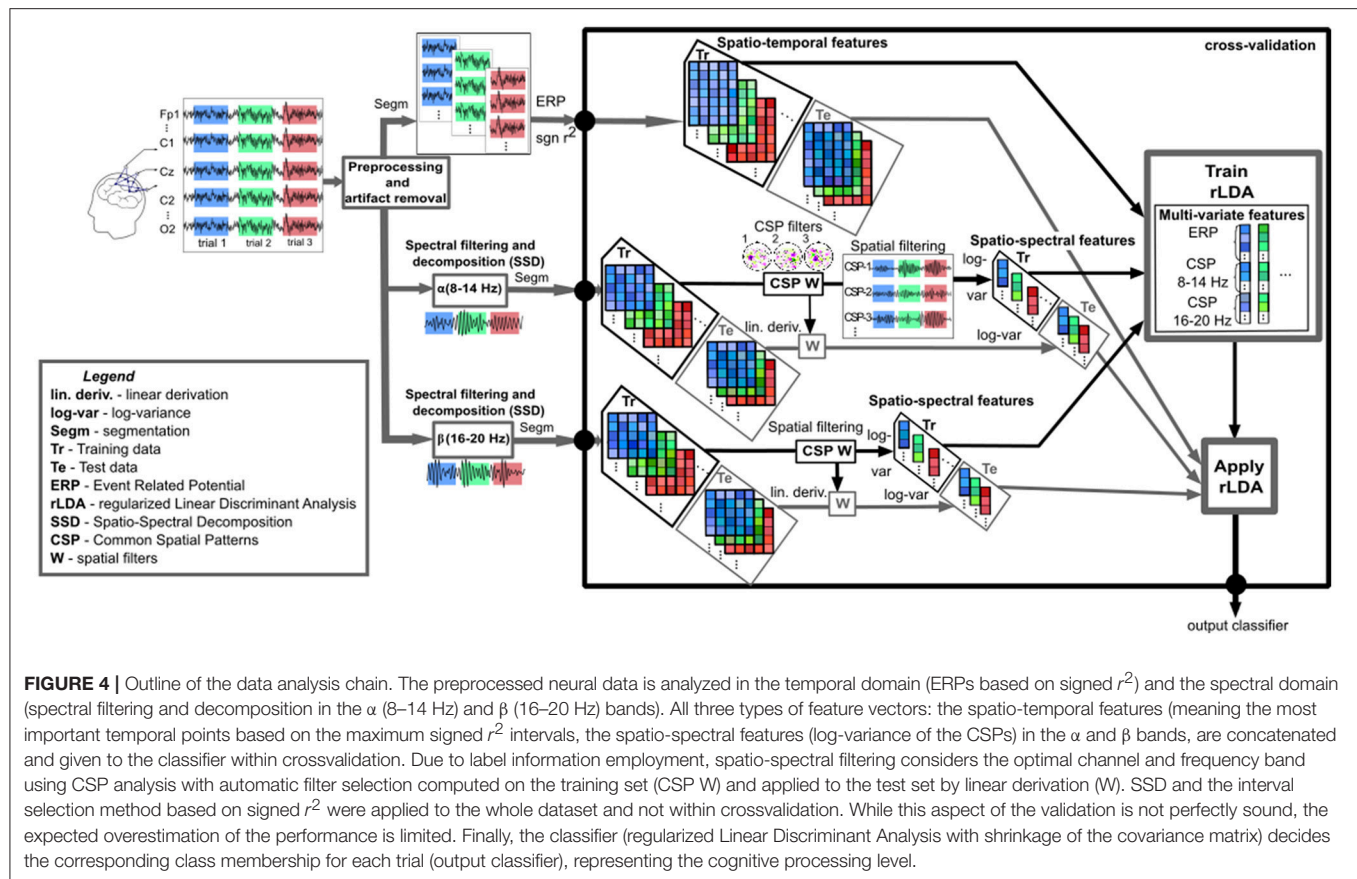
Based on the two relevant frequency bands, α (8–14 Hz) and β (16–20 Hz), we extracted the most discriminative spatial filters within each band and we combined them, such that the neural components referring to both frequency bands could be simultaneously exploited. The entire process was performed on the data after applying band-pass and SSD filters. The time interval considered was selected from the ERD/ERS phenomena, starting from 350 ms after the stimuli, which corresponds roughly to the peak time point of the P300, from which point the cognitive process should generally begin (Nicolae et al., 2015a).

Combined spatio-temporal and spatio-spectral features and evaluation scheme

Targeting the estimation of user's cognitive processing, we followed the processing pipeline described in **Figure 4**. After appropriate preprocessing for each feature type (time or power), the spatio-temporal features were combined with the spatio-spectral features given by the mCSP process, which were then classified and evaluated by cross-validation. More specifically, considering the band-power domain, the relevant spectral (log-variance) and spatial features were detected from the training data and used to spatially filter the testing data with the corresponding CSPs and were applied repeatedly in cross-validation manner (10 folds with 10 repetitions). Further, the separation was performed by a regularized Linear Discriminant Analysis (Friedman, 1989; Lemm et al., 2011) with shrinkage of the covariance matrix (Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005; Vidaurre et al., 2009; Blankertz et al., 2011), which proved to be successful for this type of analysis (Müller et al., 2003; Bartz and Müller, 2013; Farquhar and Hill, 2013). The classification performance, as the amount of correct estimated trials, was measured by the area under the Receiver Operating Characteristic (ROC) curve (Hanley and McNeil, 1982).

Initially, no normalization was performed on the feature vectors, because in our case, the features are roughly on the same scale. However, z-score normalization of the feature vectors was also applied as comparison, by subtracting the mean and dividing by the standard deviation on each feature type.

FIGURE 4 | Outline of the data analysis chain. The preprocessed neural data is analyzed in the temporal domain (ERPs based on signed r^2) and the spectral domain (spectral filtering and decomposition in the α (8–14 Hz) and β (16–20 Hz) bands). All three types of feature vectors: the spatio-temporal features (meaning the most important temporal points based on the maximum signed r^2 intervals, the spatio-spectral features (log-variance of the CSPs) in the α and β bands, are concatenated and given to the classifier within crossvalidation. Due to label information employment, spatio-spectral filtering considers the optimal channel and frequency band using CSP analysis with automatic filter selection computed on the training set (CSP W) and applied to the test set by linear derivation (W). SSD and the interval selection method based on signed r^2 were applied to the whole dataset and not within crossvalidation. While this aspect of the validation is not perfectly sound, the expected overestimation of the performance is limited. Finally, the classifier (regularized Linear Discriminant Analysis with shrinkage of the covariance matrix) decides the corresponding class membership for each trial (output classifier), representing the cognitive processing level.

RESULTS

Behavioral Data

Regarding the subjective indicators about the difficulty of the conditions (**Figure 5**, right), the lowest score was attributed to the *language* condition, the 25% percentile shows 0 score difficulty and the 75% shows a medium difficulty with score 1, meaning that the subjects mostly considered *language* as the easiest condition. The *memory* and *visual imagination* conditions were considered equally difficult by the subjects (25% rated a medium difficulty with score 1 and around 75% rated a high difficulty with score 2). For the personal user mood evaluation, it was reported a good mood in 53% of all the experiment runs and subjects, and in 47% it was specified as "ok." No bad mood was reported by the subjects during or after the experiment.

Considering the objective method given by answers ratio, we observe more accurate answers for the *language* condition with a ratio closer to zero (**Figure 5**, left—the answers ratio for all participants averaged over the runs for each condition). No improvement or decrease in performance was observed over time considering the answers ratio, showing insignificant correlations by the Spearman rank-order correlation (*memory*: $p = 0.3367$; *language* $p = 0.3982$; *visual imagination* $p = 0.1211$; and in total over all 15 runs: $p = 0.6192$). However, most of the participants showed engagement and enthusiasm throughout the experiment.

Neurophysiological Data

The temporal and spatial distribution of the neural activity represented by the ERPs is shown in **Figure 6** (with more details in Nicolae et al., 2015a).

Looking at the spatial and temporal distribution of the ERPs in **Figure 6**, a gradual difference between the levels of cognitive processing is observed in the centro-parietal area reflecting a positive peak about 400 ms after the stimulus (∼P300). Earlier, a peak around 250 ms is observed, with a negative component more pronounced in the right-occipital cortex, discriminating between no processing and processing, and similarly in the *visual imagination condition*, but discriminating also between shallow and deep processing.

Second, we visualized the modulations of the signals' power spectrum computed for the entire trial timing (2 s) in a spatial-spectral representation (topographic maps). In order to focus on the discriminative aspects, the visualization was based on signed r^2-values.

For the distinction between shallow and deep processing (**Figure 7**, the bottom graphs), we investigated the average spectrum (mean over trials and participants) given by the signed r^2 over the parietal site (Pz) in the frequency range from 3 to 40 Hz. We observed a higher discriminative difference in the α band (8–14 Hz) and smaller in the β band (16–20 Hz). Due to their prominent difference, also in the scalp maps at frontal and parietal sites, both frequency bands were selected for
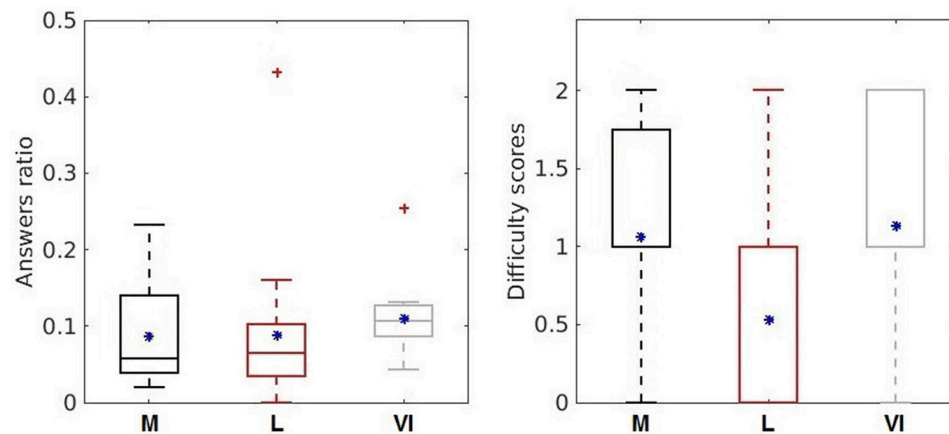
FIGURE 5 | Behavioral assessment indicators: answers ratio **(left)** and difficulty scores **(right)** for the three cognitive processes (memory—dark gray, language—light red, visual imagination–light gray). The upper and bottom whiskers of each box-plot corresponds to the maximum and minimum values over all participants. The horizontal sides of the rectangular boxes represent the 25 and 75% percentiles of the data. The mean values are represented by the blue asterisk (*) and the outliers are indicated by the red crosses. Left plot of answers ratio is taken from Nicolae et al. (2015b) with permission from Springer.
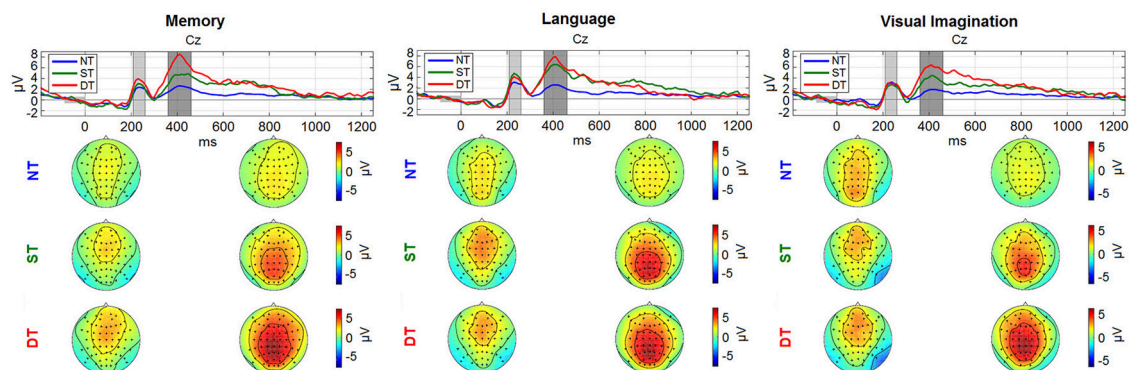
FIGURE 6 | Grand average ERPs: (a) *Memory condition*; (b) *Language*; (c) *Visual imagination*. The upper plots represent the time evolution at representative electrode Cz for the non-processing (blue line), shallow processing (green), deep processing (red). The scalp plots underneath show the topographies (-2 to 8 μV^2) referring to the shaded areas (210–260 ms; 360–460 ms) highlighted in the time plots. The figure (with scaling adjustments) was taken from Nicolae et al. (2015a), with permission from the IEEE Proceedings.

the analysis in a multi-band approach. A modulation appears also in the θ band (5–7 Hz), visible when comparing shallow with no-processing and deep with shallow processing (upper and bottom graphs). This effect was less marked compared to the other frequency bands and it was not encountered in all levels of processing, therefore it was not considered for the analysis.

Figures 8, **9** depict the grand average desynchronization and synchronization effects in the 8–14 Hz and 16–20 Hz frequency band which start about 300 ms after stimulus. The time evolution is initially similar for all levels and *conditions* until 300 ms, representing the same amount of evaluating the stimulus information. After this point, the effect of the desynchronization appears, climaxing around 500 ms and it follows a synchronization around 800–1,800 ms. For visualization, we chose the central parietal electrode Pz (for other electrode patterns, see Supplementary Figures S1, S2 in Supplementary Presentation 1). It can be noticed that shallow or

deep processes (DT) elicits an attenuation of brain rhythms in comparison to the reference of no-processing (NT). While the ERDs in the ST and DT levels start similarly, they are markedly more sustained in the DT *level*. The effect is more pronounced for the α band as -1 to 0.5 μV and less for the β band, as -0.4 to 0.4 μV. Looking over the scalp distributions, we clearly see higher synchronization (ERS) for the shallow processing (0.1–0.35 μV) compared to the reference no-processing and more pronounced desynchronization (ERD) for a more complex processing (-0.3–0.1 μV). Comparing between processes, a higher ERS in amplitude and spatial distribution is encountered for the *memory* process, contrasting to a more pronounced ERD in the *language* and *visual imagination* case.

Next, the CSP analysis provides a deeper view of the neural oscillations and activity related to cognitive processes. The patterns provide information about the presumed sources of the neural activity which are then optimally projected on the surface of the scalp.
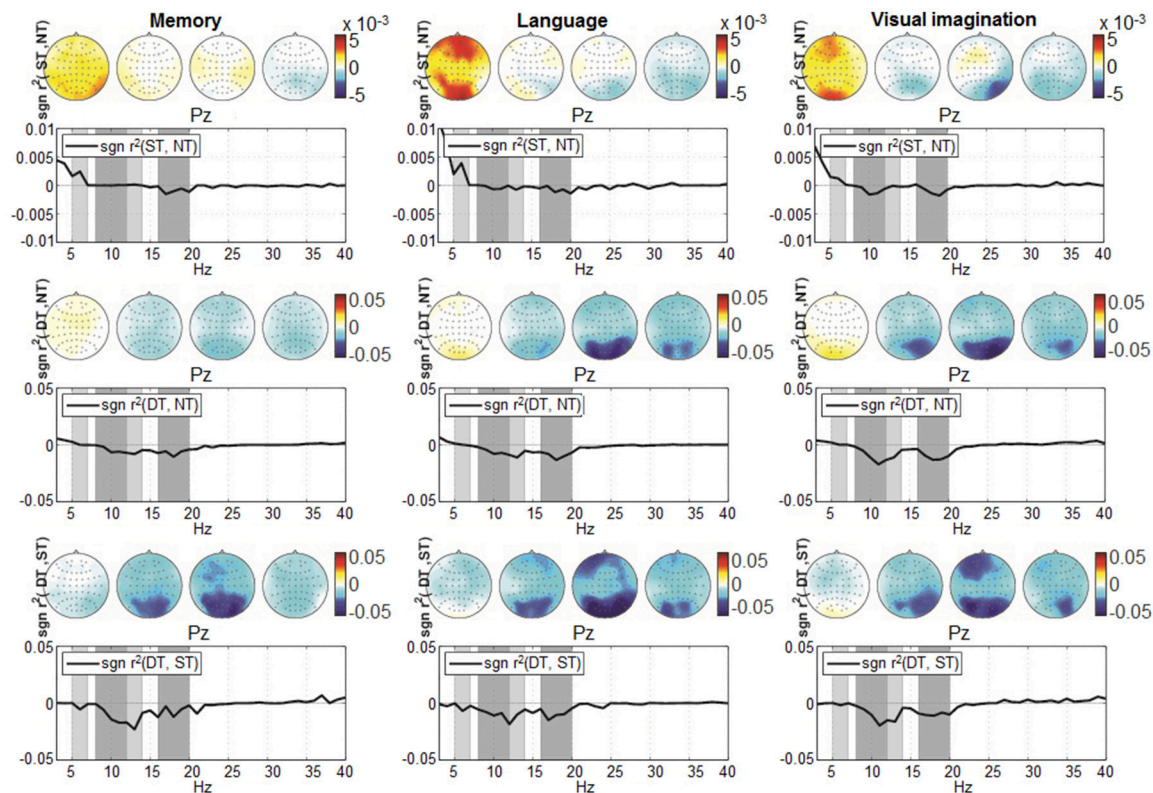
FIGURE 7 | Grand-average spectrum discrimination given by *sgn r²* at location Pz, computed over the entire trial timing (0–2,000 ms) in the frequency range from 3 to 40 Hz. The discrimination pairs ST-NT, DT-NT, and DT-ST are represented from top to bottom, while from left to right are represented according to the three *conditions*: *memory*, *language,* and *visual imagination*. The four scalp plots refer to discriminative signed *r²*-values, corresponding to θ (5–7 Hz), α (8–12 Hz; 12–14 Hz), β (16–20 Hz) frequency bands (shaded in gray). Note the upper graphs scale −0.01 to 0.01, compared to the other scales −0.05 to 0.05.
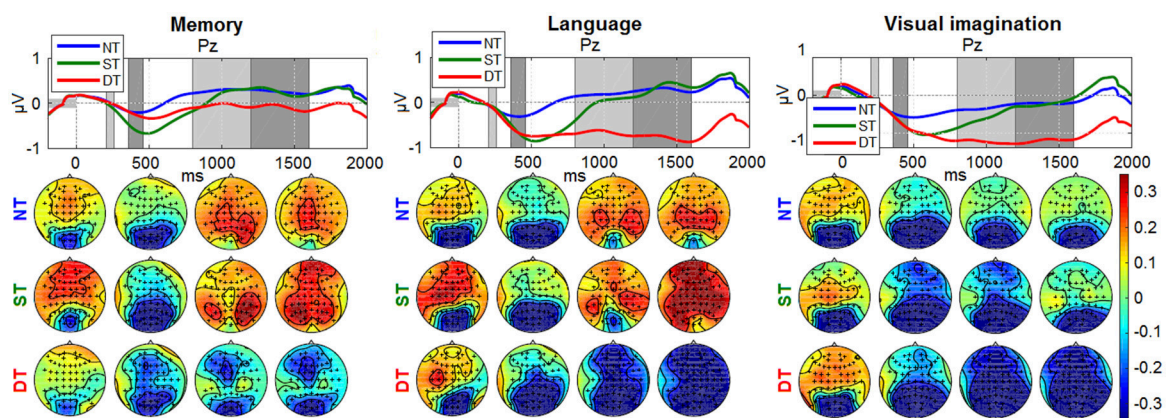
FIGURE 8 | Grand-average ERDs on 8–14 Hz for all processing levels (NT, ST, DT) at electrode Pz. The baseline interval is represented with a gray horizontal bar, equivalent to 200 ms of pre-stimulus interval. The amplitude range was chosen the same for all graphs, as −1 to 1 μV for the time evolution and −0.35 to 0.35 μV for the scalp distributions. The timing intervals for the scalp plots are: 210–260, 360–460, 800–1,200, 1,200–1,600 ms.

Considering the selected frequency bands, the CSP for participant P4 are shown in the Supplementary Presentation 1 as scalp topographies, computed for all pairs of class combinations (Supplementary Figures S3, S4).

Classification

The evaluation of the binary multivariate classification based on the combined spatio-temporal (ERP) and multi-band CSP (mCSP with SSD) features, are presented hereunder. The results
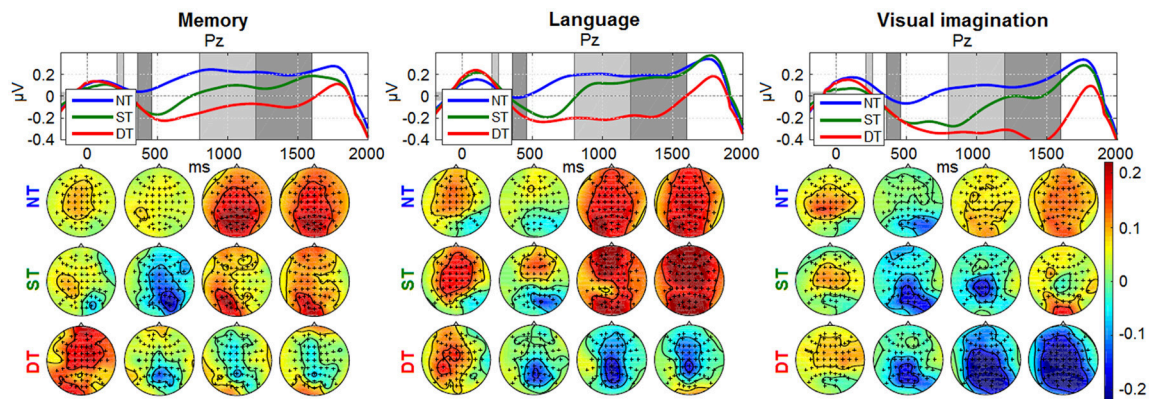
FIGURE 9 | Grand-average ERDs on 16–20 Hz at electrode Pz for all processing levels (NT, ST, DT) with 200 ms baseline. Same amplitude range for all *conditions*, as −0.4 to 0.4 µV for time evolution and −0.22 to 0.22 µV for scalp distributions.

of classification based on ERP only were presented in Nicolae et al. (2015b).

The general classification performance across participants given by the area under the ROC curve is presented as box-plots in **Figure 10**. Here we observe good performance which are on average above 70% for ST-DT discrimination, around 75–80% for NT-ST pair and the highest performance for NT-DT discrimination, around 85–90%. All performances for all participants are significantly above chance level (indicated by t-test with alpha = 0.01). A two-way repeated measures ANOVA was performed over the AUC values with the factors: *condition* and classification pairs, which provide a statistically significant difference between the classification pairs ($p < 0.001$, $F = 64.99$). Based on the *condition* factor, the results in **Figure 10** expose the highest average AUC for the *language condition*, but this observation was not statistically significant ($p = 0.2112$). The distribution of the data was verified using the one-sample Kolmogorov–Smirnov test, supposing the null hypothesis of standard normal distribution samples. The null hypothesis was rejected below the 1% significance level.

With standardization of the feature vectors, the performance is actually increased by 1–3%, compared to the results in **Figure 10** with no normalization, although this effect is not statistically significant (n-way ANOVA: $p = 0.0601$).

Now relating the classification results to the signed r^2 discrimination for the ERD/ERS curves, it is important to notice that even when no substantial difference was encountered for the NT-ST pair, the combined classification still performed considerably, which was due to the ensemble features approach that integrated also the temporal features. Moreover, when discriminating the deep processing, the classification is significantly improved in the presented combined approach compared to the separate temporal classification results (Nicolae et al., 2015b): increasing from 83 to 87% on average for the NT-DT pair ($p = 0.0027$; $F = 9.57$) and from 68 to about 72% for the ST-DT pair ($p = 0.0234$; $F = 5.33$). No statistical significant difference was obtained for the NT-ST pair. Additionally, while considering only the spectral features, using the SSD method
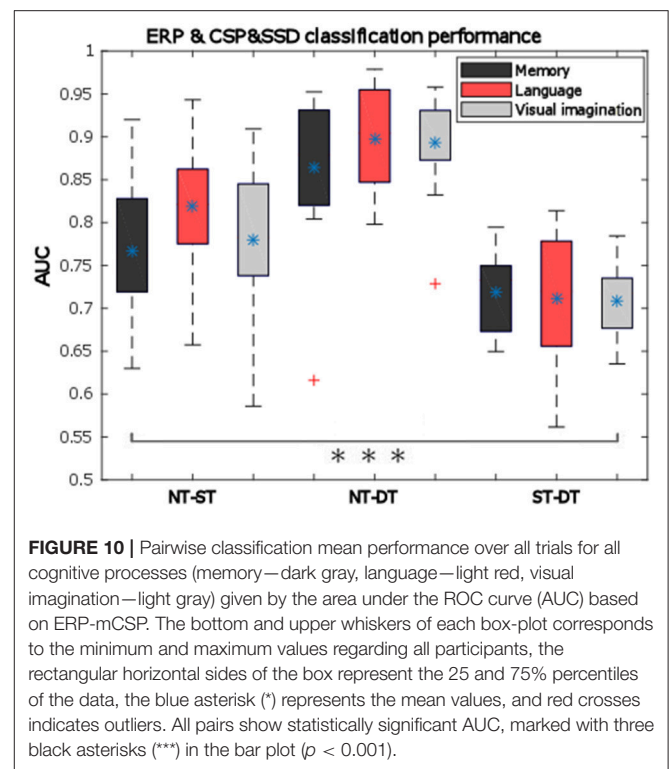
FIGURE 10 | Pairwise classification mean performance over all trials for all cognitive processes (memory—dark gray, language—light red, visual imagination—light gray) given by the area under the ROC curve (AUC) based on ERP-mCSP. The bottom and upper whiskers of each box-plot corresponds to the minimum and maximum values regarding all participants, the rectangular horizontal sides of the box represent the 25 and 75% percentiles of the data, the blue asterisk (*) represents the mean values, and red crosses indicates outliers. All pairs show statistically significant AUC, marked with three black asterisks (***) in the bar plot ($p < 0.001$).

before CSP boosts the classification performance significantly for all cases: for example, from 61 to 74% on average for the ST-DT pair considering the *visual imagination condition* (statistically significant: $p = 0.0001$ with the two-way ANOVA test).

DISCUSSION AND CONCLUSION

We investigated the neural correlates of cognitive processing that indicate the level of profoundness. Effects have been found in ERPs as well as in brain rhythms.

In the ERPs, two peaks arise: one at 250 ms and one prolonged at 400 ms. The first peak refers to the first decision on the type of the stimulus (NT, ST, or DT), based on appearance (color and category) which is similar between different levels of processing. the second peak relates to no processing (NT), mild (ST), or intense (DT) processing, and is graded in amplitude in relation to the corresponding level of processing. The latencies are similar because the levels of processing involve to the same decision task, mental computation. Further, in the deep processing task another decision is involved, deciding the task fulfillment or not, and hence different latencies will be observed trial to trial, which are not clearly visible on grand averages.

The power spectrum analysis, over a wide frequency range from 3 to 40 Hz, provided an overview of the spectral components which were pronounced for the α (8–14 Hz) and the β band (16–20 Hz) and reduced for the θ band (5–7 Hz). Analyzing the components specific to each frequency band, we observed the present activity in the θ band being more pronounced for the *memory* and *language* condition, which reflects the *memory* retrieval activity (Meyer et al., 2015), cognitive processes (Klimesch, 1999), and sustained attention (Huang et al., 2007). However, the signed r^2 differences occurred in the θ band (see **Figure 7**) are substantially smaller than the differences occurred in the α and the β band. Moreover, when testing the classification considering the θ band only, the performances were around chance level for all discrimination pairs, as expected (e.g., 0.51, 0.56, 0.51 mean AUCs for the *language* condition) which do not show significance with *t*-test at alpha = 0.0056 with Bonferroni correction for multiple comparisons, given $p > 0.0195$ for the NT-ST and ST-DT classification pairs and show barely significance for the NT-DT discrimination with $p = 0.0052$. Therefore, the decision was made to discard this band from further analysis, since in our case it did not produce significant differences. The ERD/ERS curves of the remaining frequency bands display a marked ERD with a peak at a latency of about 500 ms. The duration of the ERD is modulated by the degree of cognitive processing. In addition, strong discrimination in the α band elucidates mental coordination (Palva and Palva, 2007), alert states (Klimesch, 1999), cognitive processing, access to stored information (Klimesch, 2012), and is completed with strong β band differentiation that corresponds to complex mental process and analyzing the presented information (Lachaux et al., 2005).

Comparing across the *conditions*, the activity for *memory* is focused in the temporal and frontal area. More lateralized activity is observed for *language* and more accentuated in parietal and temporal area for *visual imagination* depicting *memory* access and interpretation (Ganis et al., 2004). Comparing the ERD/ERS effects between conditions, a more increased synchronization (higher DT in amplitude and power) is visible for the *memory condition* in comparison with the others, representing an easier process. This effect, contrasts with the behavioral point of view, where more participants stated, on average, the *language* as the easiest method (difficulty scores in **Figure 5**), effect found statistically significant with $p = 0.0451$ given by a one-way ANOVA statistical test considering one factor (subjects) and three levels (conditions).

In order to improve the oscillatory discrimination, CSP filtering was employed to obtain spatial filters that distinguish the areas where the activity is differentiated between two processing levels (Supplementary Figures S3, S4 in the Supplementary Presentation 1).

For classification, we employed an approach which combines spatio-temporal features with spatio-spectral features. The average AUC is over 76% for the NT-ST pair, more than 86% for the NT-DT pair and 70% for the ST-DT pair. Note, that classification was challenged by the fact that participants moved their eyes freely between the two objects of each stimulus. In addition, it seems probably that the eye movements might have different dynamics between the tasks (e.g., DT might induce more alternations of the gaze between the two objects of a stimulus). For that reason, appropriate artifact removal strategies (ICA with MARA) and careful verification was performed. The resulting ERD and CSP patterns (cf. **Figures 8**, **9** and Supplementary Figures S3, S4 in Supplementary Presentation 1) do not suggest influences from eye movements, like strong lateralized activity in the frontal area which would be expected to result from horizontal movements. This gives a strong indication that the classification performance was not based on eye movement dynamics but on neural correlates of the cognitive processing. Furthermore, the evaluation of potential contamination by artifacts becomes more robust when analyzing the level of decoding based on the EOG activity alone. Specifically, the classification was performed considering two feature channels (EOG and Fp1; F10 and F9 channels difference) corresponding to the vertical and horizontal eye movements. The results show significant values at chance level for all conditions and classification pairs (e.g., mean AUCs for the *language* condition considering the classification pairs: 0.51, 0.55, 0.46) with $p > 0.0198$ and Bonferroni corrected for the nine comparisons with alpha = 0.0056.

In our experimental paradigm, different cognitive levels were externally imposed using task instructions. The graduated differences observed in the ERPs between shallow and deep processing correlate with different levels of processing, evidenced also in the ERDs/ERSs and are not generated by the targets occurrence, namely the odd-ball effect, which was controlled by imposing the same percentage of the stimuli between ST and DT (12.5%). When comparing the shallow and deep processing with no processing, it cannot be disentangled which differences in the ERPs are generated by the rarity of the occurrence (oddball effect) and which by the additional processing demands of the task. In particular for the ERDs in **Figure 8** the time course (differences extending far past 500 ms) seems to suggest that the main difference is due to the additional cognitive processing in ST and DT. This design was chosen in order to have better control of the true level of cognitive processing. For the application perspective, however, we strive to estimate the momentary level of cognitive processing within the natural fluctuations.

In conclusion, the performed investigation of the depth of cognitive processing brings us closer to real scenarios. Compared to standard BCI research our study induced different levels of cognitive processing by tasks that go beyond a simple target/non-target discrimination. Moreover, the visual stimuli used had

a higher variability, including the need of eye movements in the exploration of the complex stimuli, which were composed of two objects side-by-side. Our work extends also previous investigation of the effect of task complexity on ERPs and brain oscillations. Again, the set of stimuli used in our study is richer and more complex and importantly, methods from machine learning have been employed for single-trial classification, in a binary approach.

Overall, the present study is a step forward toward applications that estimate the level of cognitive processing in realistic settings of human-computer interaction (Gamberini et al., 2015) and in safety critical workplaces (Venthur et al., 2010). A further step in order to discriminate the ongoing level of cognitive processing is to apply a regression approach, as in Naumann et al. (2017). A different perspective for the current study is the development of techniques suitable for adaptive learning environments based on user state monitoring regarding the depth of cognitive processing. In this regard, it would be interesting to augment the approach with predicting remembered vs. not remembered items (Klimesch et al., 1996) in *memory* tasks.

DATA AVAILABILITY STATEMENT

The EEG signals and behavioral data acquired in this study will be made available from the DepositOnce repository of Technische Universität Berlin (https://depositonce.tu-berlin.de/).

AUTHOR CONTRIBUTIONS

Design of the study by BB. Implementation and data acquisition by IN. IN performed data analysis and LA and BB reviewed and revised the analysis. IN wrote the manuscript draft which was reviewed and revised by LA and BB.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2017.00548/full#supplementary-material

Supplementary Video 1 | Short example with 30 trials for the experimental paradigm considering the *memory* condition. First, the visual questionnaire is presented evaluating participant's current mental state (good, ok, bad). Next, the target pair (pear, banana) specifying the target category (fruits) and the color target (red) is presented along with the current condition that needs to be performed (*memory*). When the participant memorized the target, he/she press a keyboard button and the sequence starts after a short countdown. At the end of the run, the participant enters the number counted and receives feedback considering the correct number. In total, there were 4 DT (from which two fulfilled the *memory* question task), 7 ST, and 19 NT stimuli presented, accumulating a final number counted of 31. Note that the number of trials is higher in the actual run of the experiment (120 stimuli) and the ratio of the stimuli is also different (75:12.5:12.5 ± 2% for NT:ST:DT), while this example may seem harder (with a ratio of 63.3:23.3:13.3 (%) for NT:ST:DT).

Supplementary Video 2 | Short example with 30 trials for the experimental paradigm considering the *language* condition. The target pair here (submarine, submarine) refer to the mobility category and shows the color target (green). The current condition that needs to be performed is also specified (*language*). In total, 4 DT (from which two fulfilled the *language* question task), 7 ST, and 19 NT stimuli were presented, resulting in a final number counted of 31. Same notice as in the memory example, that the number of trials and the ratio are different here compared with the actual experiment.

Supplementary Video 3 | Short example with 30 trials for the experimental paradigm considering the *visual imagination* condition. The target pair here (cow, elephant) and the color target (red) are presented along with the specification of the current condition that needs to be performed (*visual imagination*). In total, 4 DT (from which two fulfilled the *visual imagination* question task), 5 ST, and 21 NT stimuli were presented, resulting in a final number counted of 29. Same notice as in the memory and language example, that the number of trials and the ratio (70:16.67:13.3 (%) for NT:ST:DT) are different compared with the actual experiment.

REFERENCES

Acqualagna, L., and Blankertz, B. (2013). Gaze-independent BCI-spelling using rapid serial visual presentation (RSVP). *Clin. Neurophysiol.* 124, 901–908. doi: 10.1016/j.clinph.2012.12.050

Bartz, D. M., and Müller, K. R. (2013). "Generalizing analytic shrinkage for arbitrary covariance structures," in *Advances in Neural Information Processing Systems*, Vol. 26 (Lake Tahoe), 1869–1877.

Basile, L. F., Sato, J. R., Alvarenga, M. Y., Henrique, N. Jr., Pasquini, H. A., Alfenas, W., et al. (2013). Lack of systematic topographic difference between attention and reasoning beta correlates. *PLoS ONE* 8:e59595. doi: 10.1371/journal.pone.0059595

Birbaumer, N. (2006). Breaking the silence: brain–computer interfaces (BCI) for communication and motor control. *Psychophysiology* 43, 517–532. doi: 10.1111/j.1469-8986.2006.00456.x

Blankertz, B., Acqualagna, L., Dähne, S., Haufe, S., Schultze-Kraft, M., Sturm, I., et al. (2016). The Berlin brain-computer interface: progress beyond communication and control. *Front. Neurosci.* 10:530. doi: 10.3389/fnins.2016.00530

Blankertz, B., Jacucci, G., Gamberini, L., Spagnolli, A., and Freeman, J. (eds.). (2015). "Symbiotic interaction," in *4th International Workshop, Symbiotic 2015, Proceedings* (Cham: Springer International Publishing), 9359.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K. R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *Neuroimage* 56, 814–825. doi: 10.1016/j.neuroimage.2010.06.048

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., et al. (2010). The Berlin brain-computer interface: non-medical uses of BCI technology. *Front Neurosci.* 4:198. doi: 10.3389/fnins.2010.00198

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Müller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* 25, 41–56. doi: 10.1109/MSP.2008.4408441

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Buzsáki, G., and Draguhn, A. (2004). Neuronal oscillations in cortical networks *Science* 304, 1926–1929. doi: 10.1126/science.1099745

Chen, Y. N., Mitra, S., and Schlaghecken, F. (2008). Sub-processes of working memory in the N-back task: an investigation using ERPs. *Clin. Neurophysiol.* 119, 1546–1559. doi: 10.1016/j.clinph.2008.03.003

Craik, F. I. M., and Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *J. Verbal Learn. Verbal Behav.* 11, 671–684. doi: 10.1016/S0022-5371(72)80001-X

Donchin, E., Kubovy, M., Kutas, M., Johnson, R. Jr., and Herning, R. I. (1973). Graded changes in evoked response (P300) amplitude as a function of cognitive activity. *Percept. Psychophys.* 14, 319–324. doi: 10.3758/BF032 12398

Dornhege, G., Blankertz, B., Curio, G., and Müller, K. R. (2004). Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Trans. Biomed. Eng.* 51, 993–1002. doi: 10.1109/TBME.2004.827088

Dornhege, G., Millán, J. d. R., Hinterberger, T., McFarland, D., and Müller, K. R., (eds.). (2007). *Toward Brain-Computer Interfacing.* Cambridge MA: MIT Press.

Farquhar, J., and Hill, N. J. (2013). Interactions between pre-processing and classification methods for event-related-potential classification - best-practice guidelines for brain-computer interfacing. *Neuroinformatics* 11, 175–192. doi: 10.1007/s12021-012-9171-0

Fazli, S., Dähne, S., Samek, W., Bießmann, F., and Müller, K. R. (2015). Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based brain–computer interfaces. *Proc. IEEE.* 103, 891–906. doi: 10.1109/JPROC.2015.2413993

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Stat. Assoc.* 84, 165–175. doi: 10.1080/01621459.1989.10478752

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition, 2nd Edn.* Boston, MA: Academic Press.

Gamberini, L., Spagnolli, A., Blankertz, B., Kaski, S., Freeman, J., Acqualagna, L., et al. (2015). "Developing a symbiotic system for scientific information seeking: the mindsee project," in *Symbiotic Interaction, Lecture Notes in Computer Science,* Vol. 9359, eds B. Blankertz, G. Jacucci, L. Gamberini, A. Spagnolli, and J. Freeman (Cham, Switzerland: Springer International Publishing), 68–80.

Ganis, G., Lacey, S., and Lawson, R. (eds.). (2013). "Multisensory imagery," in *Visual Mental Imagery*, eds S. Lacey and R. Lawson (New York, NY: Springer Science Business Media LLC), 9–28.

Ganis, G., Thompson, W. L., and Kosslyna, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cogn. Brain Res.* 20, 226–241. doi: 10.1016/j.cogbrainres.2004.02.012

Gevins, A., Smith, M. E., McEvoy, L., and Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cereb. Cortex* 7, 374–385. doi: 10.1093/cercor/7.4.374

Guger, C., Sorger, B., Noirhomme, Q., Naci, L., Monti, M. M., Real, R., et al. (2014). "Brain-computer interfaces for assessment and communication in disorders of consciousness," in *Emerging Theory and Practice in Neuroprosthetics*, ed G. R. Naik (Hershey, PA: IGI Global), 181–214.

Gundel, A., and Wilson, G. F. (1992). Topographical changes in the ongoing EEG related to the difficulty of mental tasks. *Brain Topogr.* 5, 17–25. doi: 10.1007/BF01129966

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Hassanien, A. E., and Taher Azar, A. (eds.) (2015). *Brain-Computer Interfaces: Current Trends and Applications,* Vol. 74. Springer.

Haufe, S., Dähne, S., and Nikulin, V. V. (2014). Dimensionality reduction for the analysis of brain oscillations. *Neuroimage* 101, 583–597. doi: 10.1016/j.neuroimage.2014.06.073

Huang, R. S., Jung, T. P., and Makeig, S. (2007). "Multi-scale EEG brain dynamics during sustained attention tasks," in *Proceedings of the 32th International Conference on Acoustics, Speech and Signal Processing* (Honolulu, HI), 1173–1176.

Kim, K. H., Kim, J. H., Yoona, J., and Jung, K. Y. (2008). Influence of task difficulty on the features of event-related potential during visual oddball task. *Neurosci. Lett.* 445, 179–183. doi: 10.1016/j.neulet.2008.09.004

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Klimesch, W. (2003). "Interindividual differences in oscillatory EEG activity and cognitive performance," in *The Cognitive Neuroscience of Individual Differences,* eds I. Reinvang, M. W. Greenlee, and M. Herrmann (Oldenburg: BIS), 87–99.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends Cogn Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Klimesch, W., Doppelmayr, M., Pachinger, T., and Russegger, H. (1997). Event-related desynchronization in the alpha band and the processing of semantic information. *Cogn. Brain Res.* 6, 83–94. doi: 10.1016/S0926-6410(97)00018-9

Klimesch, W., Doppelmayr, M., Russegger, H., and Pachinger, T. (1996). Theta band power in the human scalp EEG and the encoding of new information. *Neuroreport* 7, 1235–1240. doi: 10.1097/00001756-199605170-00002

Klimesch, W., Pfurtscheller, G., and Schimke, H. (1992). Pre- and post-stimulus processes in category judgement tasks as measured by event-related desynchronization (ERD). *J. Psychophysiol.* 6, 185–203.

Klimesch, W., Schimke, H., and Pfurtscheller, G. (1993). Alpha frequency, cognitive load and memory performance. *Brain Topogr.* 5, 241–251. doi: 10.1007/BF01128991

Klimesch, W., Schimke, H., and Schwaiger, J. (1994). Episodic and semantic memory: an analysis in the EEG theta and alpha band. *Electroenceph. Clin. Neurophysiol.* 91, 428–441. doi: 10.1016/0013-4694(94)90164-3

Koles, Z. J. (1991). The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG Electroencephalogr. *Clin. Neurophysiol.* 79, 440–447. doi: 10.1016/0013-4694(91)90163-X

Lachaux, J. P., George, N., Tallon-Baudry, C., Martinerie, J., Hugueville, L., Minotti, L., et al. (2005). The many faces of the gamma band response to complex visual stimuli. *Neuroimage* 25, 491–501. doi: 10.1016/j.neuroimage.2004.11.052

Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large dimensional covariance matrices. *J. Multivar. Anal.* 88, 365–411. doi: 10.1016/S0047-259X(03)00096-4

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K. R. (2011). Introduction to machine learning for brain imaging. *Neuroimage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004

Lemm, S., Müller, K. R., and Curio, G. (2009). A generalized framework for quantifying the dynamics of Eeg event-related desynchronization. *PLoS Comput. Biol.* 5:e1000453. doi: 10.1371/journal.pcbi.1000453

Mecklinger, A., Kramer, A. F., and Strayer, D. L. (1992). Event-related potentials and EEG components in a semantic memory search task. *Psychophysiology* 29, 104–119. doi: 10.1111/j.1469-8986.1992.tb02021.x

Meyer, L., Grigutsch, M., Schmuck, N., Gaston, P., and Friederici, A. D. (2015). Frontal-posterior theta oscillations reflect memory retrieval during sentence comprehension. *Cortex* 71, 205–218. doi: 10.1016/j.cortex.2015.06.027

Mühl, C., Jeunet, C., and Lotte, F. (2014). EEG-based workload estimation across affective contexts. *Front. Neurosci.* 8:114. doi: 10.3389/fnins.2014.00114

Müller, K. R., Anderson, C. W., and Birch, G. E. (2003). Linear and nonlinear methods for brain-computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11, 165–169. doi: 10.1109/TNSRE.2003.814484

Müller, K. R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: from brain–computer interfacing to mental state monitoring. *J. Neurosci. Methods* 167, 82–90. doi: 10.1016/j.jneumeth.2007.09.022

Nakata, H., Sakamoto, K., Otsuka, A., Yumoto, M., and Kakigi, R. (2013). Cortical rhythm of No-go processing in humans: an MEG study. *Clin. Neurophysiol.* 124, 273–282. doi: 10.1016/j.clinph.2012.06.019

Naumann, L., Schultze-Kraft, M., Dähne, S., and Blankertz, B. (2017). "Prediction of difficulty levels in video games from ongoing EEG," in *Symbiotic Interaction,* Vol. 9961, eds L. Gamberini, A. Spagnolli, G. Jacucci, B. Blankertz, and J. Freeman (Cham: Springer International Publishing), 125–136.

Nicolae, I. E., Acqualagna, L., and Blankertz, B. (2015a). "Neural indicators of the depth of cognitive processing for user-adaptive neurotechnological

applications," in *Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society* (Milano), 1484–1487.

Nicolae, I. E., Acqualagna, L., and Blankertz, B. (2015b). "Tapping neural correlates of the depth of cognitive processing for improving human computer interaction," in *Symbiotic Interaction*, Vol. 9359, *Lecture Notes in Computer Science*, eds B. Blankertz, G. Jacucci, L. Gamberini, A. Spagnolli, and J. Freeman (Cham: Springer International Publishing), 126–131. doi: 10.1007/978-3-319-24917-9_13

Nicolae, I. E., Acqualagna, L., and Blankertz, B. (2016). "Investigating depth of cognitive processing in the brain dynamics of oscillations," in *Proceedings of 6th International BCI Meeting* (Pacific Grove, CA: Verlag der TU Graz), 186.

Nikulin, V. V., Nolte, G., and Curio, G. (2011). A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. *Neuroimage* 55, 1528–1535. doi: 10.1016/j.neuroimage.2011.01.057

Okazaki, M., Kaneko, Y., Yumoto, M., and Arima, K. (2008). Perceptual change in response to a bistable picture increases neuromagnetic beta-band activities. *Neurosci. Res.* 61, 319–328. doi: 10.1016/j.neures.2008.03.010

Palva, S., and Palva, M. (2007). New vistas for α-frequency band oscillations. *Trends Neurosci.* 30, 150–158. doi: 10.1016/j.tins.2007.02.001

Pesonen, M., Hämäläinen, H., and Krause, C. M. (2007). Brain oscillatory 4–30 Hz responses during a visual n-back memory task with varying memory load. *Brain Res.* 1138, 171–177. doi: 10.1016/j.brainres.2006.12.076

Pfurtscheller, G., and Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* 110, 1842–1857. doi: 10.1016/S1388-2457(99)00141-8

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Schack, B., Klimesch, W., and Sauseng, P. (2005). Phase synchronization between theta and upper alpha oscillations in a working memory task. *Int. J. Psychophysiol.* 57, 105–114. doi: 10.1016/j.ijpsycho.2005.03.016

Schäfer, J., and Strimmer, K. (2005). A shrinkage approach to largescale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4:32. doi: 10.2202/1544-6115.1175

Schultze-Kraft, M., Dähne, S., Gugler, M., Curio, G., and Blankertz, B. (2016). Unsupervised classification of operator workload from brain signals. *J. Neural Eng.* 13:036008. doi: 10.1088/1741-2560/13/3/036008

Sergeant, J., Geuze, R., and van Winsum, W. (1987). Event-related desynchronization and P300. *Psychophysiology* 24, 272–277. doi: 10.1111/j.1469-8986.1987.tb00294.x

Sheth, B. R., Sandkühler, S., and Bhattacharya, J. (2009). Posterior beta and anterior gamma oscillations predict cognitive insight. *J. Cogn. Neurosci.* 21, 1269–1279. doi: 10.1162/jocn.2009.21069

Stipacek, A., Grabner, R. H., Neuper, C., Fink, A., and Neubauer, A. C. (2003). Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load. *Neurosci. Lett.* 353, 193–196. doi: 10.1016/j.neulet.2003.09.044

Ullsperger, P., Neumann, U., Gille, H. G., and Pietschmann, M. (1987). P300 and anticipated task difficulty. *Int. J. Psychophysiol.* 5, 145–149. doi: 10.1016/0167-8760(87)90018-3

van Erp, J. B., Brouwer, A. M., and Zander, T. O. (2015). Editorial: Using neurophysiological signals that reflect cognitive or affective state. *Front. Neurosci.* 9:36. doi: 10.3389/fnins.2015.00136

van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., et al. (2009). The brain-computer interface cycle. *J. Neural Eng.* 6:041001. doi: 10.1088/1741-2560/6/4/041001

Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J. (2001). The brain web: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

Venthur, B., Blankertz, B., Gugler, M. F., and Curio, G. (2010). "Novel applications of BCI technology: psychophysiological optimization of working conditions in industry," in *IEEE International Conference on Systems Man and Cybernetics (SMC)* (Istanbul), 417–421.

Vidaurre, C., Krämer, N., Blankertz, B., and Schlögl, A. (2009). Time domain parameters as a feature for EEG-based brain–computer interfaces. *Neural Netw.* 22, 1313–1319. doi: 10.1016/j.neunet.2009.07.020

Winkler, I., Haufe, S., and Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behav. Brain Funct.* 7:30. doi: 10.1186/1744-9081-7-30

Wolpaw, J. R., and Wolpaw, E. W. (eds.). (2012). *Brain-Computer Interfaces: Principles and Practice.* New York, NY: Oxford University Press.

Yordanova, J., Kolev, V., and Polich, J. (2001). P300 and alpha event-related desynchronization (ERD). *Soc. Psychophysiol. Res.* 38, 143–152. doi: 10.1111/1469-8986.3810143

Zander, T. O., and Kothe, C. (2011). Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *J. Neural Eng.* 8:025005. doi: 10.1088/1741-2560/8/2/025005

Zander, T. O., Krol, L. R., Birbaumer, N. P., and Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proc. Natl. Acad. Sci. U.S.A.* 113, 14898–14903. doi: 10.1073/pnas.1605155114

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages
of publishing
in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

**REPRODUCIBILITY OF
RESEARCH**
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership