EEG Correlates of Difficulty Levels in Dynamical Transitions of Simulated Flying and Mapping Tasks

Ping-Keng Jao, Ricardo Chavarriaga, *Senior Member, IEEE*, Fabio Dell'Agnola, *Member, IEEE*, Adriana Arza, *Member, IEEE*, David Atienza, *Fellow, IEEE*, and José del R. Millán, *Fellow, IEEE*

***Abstract*—Decoding the subjective perception of task difficulty may help improve operator performance, *i.e.*, automatically optimize the task difficulty level. Here, we aim to decode a compound of cognitive states that covaries with the task difficulty level. We designed a protocol composed of two different subtasks, flying and visual recognition, to induce different difficulty levels. We first showed that electroencephalography (EEG) signals could be a reliable source for discriminating different compound states. To gain insight into the underlying components in the compound states, we examined the attentional index and engagement index as in our previous study. We showed that (1) attention and engagement are essential components but fail to provide the best accuracy, and (2) our model is consistent with our previous study, which means that lateralized modulations in the α bands are representative of the flying task. We also analyzed a practical issue in the design of adaptive Human-Machine Interaction (HMI) systems, namely, the latency of changes in the user's compound state. We hypothesized that the EEG correlates of the task difficulty level do not instantaneously reflect the changes in the task difficulty. We validated the hypothesis by measuring the time required for our decoders to provide stable accuracy after the task changed. This amount of time, or latency, could be as high as ten seconds. The results suggest that the latency of changes in the user's compound state between different tasks is a factor that should be taken into account when building adaptive HMI systems.**

***Index Terms*—EEG, Cognitive, Workload, Difficulty, Transitions, EEG correlates**

I. Introduction

A user's cognitive state strongly influences his or her performance with a Human-Machine Interaction (HMI). Therefore, decoding the correlates of the user's cognitive state could be exploited to adapt the level of difficulty during the interaction to improve the user's performance and enhance the user's experience. For instance, intelligent systems can change the level of assistance they provide [1], adapt the level of difficulty in gaming [2], filter the amount of information according to the user's workload [3], and decide the level when learning piano [4]. Additionally, providing users with feedback about their current arousal state can help improve task performance [5].

We are interested in "the correlates of the perceived task difficulty" because it is critical in the challenge point theory [6]. The theory states that optimal performance resides at a task difficulty level that is neither too high nor too low. More importantly, the theory notes that employing objective task difficulty (*e.g.*, the target size in a shooting task) can cause inconsistency in the performance curves among people because experts and novices perceive the task difficulty differently for the same objective level. Instead, the task difficulty should be a function of skill. Hereafter, we rephrase the functional task difficulty level in [6] as the perceived (subjective perception of) difficulty level. Several experimental paradigms have identified many cognitive states that can vary with the perceived difficulty level. They include arousal, attention, engagement, and working memory [5], [7]–[9]. Unfortunately, they are hard to disentangle in realistic tasks, as they influence each other but do not always covary. In this study, we do not aim to identify correlates of a specific cognitive state, but instead, we are interested in detecting variations in the compound of cognitive states related to how difficult a task is perceived by a subject.

The user's cognitive states can be inferred from various sources, such as task performance, explicit user feedback, and physiological signals. Examples of the latter are heart rate [10], [11], pupil dilation [5], [12], [13], electrodermal activity [14], [15], and electroencephalography (EEG) signals [2], [5], [16], [17]. In this study, we focus on EEG signals, as they exhibit a higher temporal resolution than other physiological signals, for decoding the compound cognitive state linked to the perceived level of difficulty.

Some existing EEG-based experimental protocols for measuring cognitive states require external stimuli irrelevant to the task. For example, a person highly focused on a visual task may not easily perceive auditory stimuli [18]. As a result, while the main task is not auditory, one can use EEG event-related potentials elicited by oddball sounds to measure the level of concentration [19], [20]. Nonetheless, the additional stimuli may degrade the user experience, *e.g.*, a game designer may not want to add irrelevant sounds in a game designed with good music.

Manuscript received June 20, 2019; revised Mar. 31, 2020, and July 04, 2020; accepted Oct. 04 2020. This work has been supported by the Swiss National Centres of Competence in Research (NCCR) Robotics and by the ONR-G through Award Grant No. N62909-20-1-2063. We also would like to acknowledge Alexander Cherpillod for his help in the implementation of the simulator. (Corresponding author: Ping-Keng Jao.)

Ping-Keng Jao is with the École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland (email: ping-keng.jao@alumni.epfl.ch).

Fabio Dell'Agnola, Adriana Arza, and David Atienza are with the Embedded Systems Laboratory (ESL), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Ricardo Chavarriaga is with the École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland, and with ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland.

José del R. Millán is with the Department of Electrical and Computer Engineering & the Department of Neurology, University of Texas at Austin, Austin, United States of America, and with École Polytechnique Fédérale de Lausanne (EPFL), Geneva, Switzerland.

Digital Object Identifier 00.0000/THMS.0000.0000000

In this paper, we take a different approach that aims to identify the EEG correlates of the compound cognitive state without the burden of additional tasks. Specifically, we designed a realistic protocol where the participant has to pilot a simulated drone, recognize visual targets, or perform both tasks at the same time. We previously showed the feasibility of decoding different levels of difficulty in a piloting task [16]. Here, we extend this approach from a single-task study to a multitask study. Therefore, the *first goal is to investigate whether it is feasible to distinguish differences in the compound of co-occurring cognitive states induced by combinations of multiple tasks from EEG signals*.

In real-life applications, the level of task difficulty typically changes over time and likely induces dynamic cognitive states. An adaptive HMI system can regulate the difficulty level according to the estimated cognitive states, thus establishing a closed-loop system between user and machine. Consequently, successful decoding of cognitive states must continuously capture these temporal variations. However, most HMI studies on decoding cognitive states (*e.g.*, workload or level of attention) focus on setups where the conditions of the task remain constant and assume the user states would as well [2], [5], [16], [17].

In this work, we address the aforementioned limitation by studying the effects of dynamic changes in difficulty on the user's cognitive states. Assuming that there is a perceptible latency in the EEG signals, we denote a *transition period* as the time from the moment when the difficulty changes to the moment we can reliably decode the new cognitive state (*i.e.*, the *stable period)*. Consequently, the *second goal is to study the dynamic changes in the EEG correlates of the level of difficulty*.

The remaining sections are organized as follows. We first present the data collection in Section II. We detail our analytical method in Section III. Then, the results of the data analysis, discussion and conclusion are provided in Sections V, VI, and VII, respectively.

II. Materials

A. Participants and Setup

Twenty-four subjects (six females; mean age 27.27; SD 4.8) participated in the study. Each subject participated in two recording sessions on different days (min/max elapsed days: 1/36; median: 5; average: 8.42). The protocol was approved by the local ethical committee, and all the subjects provided written consent. Subjects sat comfortably in front of a twenty-four-inch screen showing the protocol with 1920x1200 resolution. They held a game pad with which they could provide inputs to the protocol. The game pad had a joystick on the left side and four colored buttons on the right side. All subjects had normal or corrected-to-normal vision and reported no history of motor or neurological disease.

A Biosemi ActiveTwo amplifier was used to record EEG and Electrooculography (EOG) signals at 2,048 Hz. The placement of sixty-four EEG electrodes followed the international 10-10 standard. Three additional channels were placed on the middle point of both eyebrows and on the bulge bones below the outer sides of the canthi to measure the EOG.

We simultaneously recorded electrocardiography, respiration, skin conductance, skin temperature, rate of blood flow, and impedance cardiography using a Biopac system. A hardware trigger was sent in parallel to both the Biosemi and Biopac systems to synchronize the events. The analysis of these signals is outside of the scope of this study.

B. Protocol

In this work, we aimed to study the compound cognitive states in a task that presented dynamic changes in the level of difficulty by combinations of different subtasks.

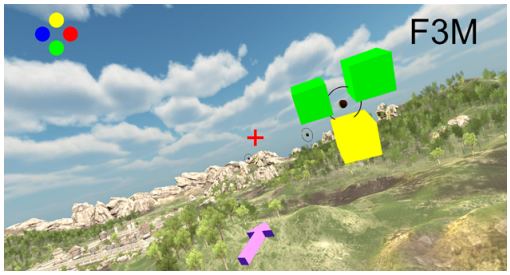
Each recording session was composed of multiple conditions, which included flying a simulated drone, visual recognition, and their combination, the same conditions as in [21]. First, we presented a low-difficulty baseline condition (B), where the subject watched the drone automatically fly through a set of waypoints. The subjects were instructed not to use the game pad. Any attempts to control the drone would have no effect.

The second condition was a flying task where the subject steered the simulated drone (F). In this task, the subject used her left hand to make the drone pass through 122 circular waypoints. The waypoints served as reference points for the subject to correct the orientation. To avoid confusion, only the current and two upcoming waypoints were shown. For the flying task, only roll and pitch were allowed; the subject could not control yaw or throttle.

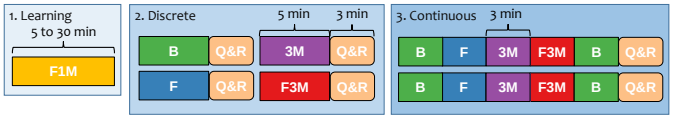
The third condition was a visual recognition task, where the subject was asked to map objects (3M) while the drone was automatically steered by the simulator. The automatic navigation mechanism in 3M was the same as in B. In this task, three colored cubes out of four possible colors would appear close to the next waypoint when the current waypoint was crossed. There were a total of 80 groups of objects over 122 waypoints. Subjects were asked to press the corresponding color-coded buttons on the game pad upon appearance of the objects. If the subject pressed the correct color button (hit), the corresponding object disappeared; otherwise (miss), there was no effect on the visual feedback. The subjects were told that a miss reduced their task performance. The colors of the three cubes could be the same, but one button press would eliminate only one cube. The order of button pressing did not matter.

The fourth condition, expected to have the highest level of difficulty, was to perform flying and mapping (F3M) at the same time.

Fig. 1(a) is a screenshot of the protocol. On the top right of the screen, the current condition is displayed (in this case, F3M). On the top left, the layout of the buttons on the game pad is presented as a reminder to the subjects to prevent them from looking at the game pad and to help them fixate on the screen, reducing possible EEG artifacts from muscle activity and EOG during the recording. On the bottom center, there is an arrow indicating the direction to the next target waypoint. This was particularly helpful in cases in which the target waypoint fell outside the current view. The waypoint has a small sphere in the center. The radius of each waypoint



(a) Protocol screenshot



(b) Experimental design

Fig. 1: Protocol and the design of the experiment. (a) The top right shows the current task. The top left is the layout of the button colors. The purple arrow below points to the current waypoint. The red cross is the center of the drone. Circles are waypoints to be passed. Colored cubes are the objects of the mapping task. (b) Each session consists of three phases, a learning-to-control phase and two phases that present different conditions either discretely or continuously. The presenting order of 3M, F, and F3M in the discrete and continuous phases was pseudorandomized across the two sessions for each subject.

was one arbitrary unit (A.U.). Based on [16], one A.U. was the mean value of the radius distribution for many subjects (for highly skilled subjects, this value was approximately 0.7 A.U.). Therefore, one A.U. was small enough that most subjects needed to focus on the navigation task if they wanted to pass through the waypoint. The subjects were instructed to do that if possible, but they were also informed that the trial would not stop due to a missed waypoint. The red cross in the center represents the center of the drone, center used to determine if the drone was inside a waypoint. The cubes to be mapped appeared around the target waypoint and disappeared when the target waypoint changed, even if the subject failed to eliminate them.

To study the transition period, we designed each recording with three phases, as depicted in Fig. 1(b). The *first phase* was devoted to allowing the subject to become familiar with the task, and there was no recording of neural signals. During the sensor setup, the subject performed a combined flying and mapping task (F1M). In this phase, the mapping task included only a single cube instead of three. One trial of this task lasted 5 minutes. A subject could attempt this task as many times as she wished before starting the experiment. Setting up all sensors normally took between 30 and 60 minutes, so the subject had sufficient learning time.

During the *second phase (discrete)*, each condition was performed separately. After finishing each condition, there was a three-minute questionnaire and a resting period (Q&R). The Q&R period was intended to reduce the effect of an experimental condition on the following one and to allow studying the transition effects from the resting state. During this period, the subject reported her perceived difficulty level from 0 to 100 and her workload level through the NASA-Task Load Index (NASA-TLX) [22]. Both questionnaires provided an assessment of whether the four conditions induced different subjective cognitive states. In this phase, the baseline condition was always recorded first. Each condition lasted 5 minutes, yielding a total effective recording time of 20 minutes for each of the two sessions. With the five-minute constraint, the subjects passed through 93.4 waypoints out of 122 on average.

During the *third phase (continuous)*, the experimental conditions were presented without interleaving the rest periods, allowing us to study the effect of dynamic changes in the level of difficulty on the neurophysiological signals. There were two trials per recording, each composed of 5 segments corresponding to the different conditions (see Fig 1(b)). The baseline was presented twice, always as the first and last condition of each trial. Each of the other three conditions was presented once in a trial that finished with a Q&R period as in the second phase. Due to the continuous nature of this phase, it was impossible to gather the subjects' immediate self-reports for each condition. Each condition lasted three minutes for a total of 30 minutes of effective signals for the two trials.

Given the length of a session (110 to 140 minutes, including the setup), it was impossible to test all condition permutations in each recording session. We therefore presented the conditions in a pseudorandomized way across two recordings for the second and third phases. Namely, we presented the six permutations of three conditions (F, 3M, and F3M) randomly order during the two recordings for each subject.

III. Decoding Cognitive States from the EEG

The decoding of the compound cognitive state was formulated as a classification problem for the four conditions (B, F, 3M, F3M). For each recording, we sorted the self-reported difficulty level averaged from the three trials and assigned a label from 1 (easiest) to 4 (hardest) to a condition. Then, we assessed different sets of conditions by pairs or triads (1-2-4, 1-3-4) or all four classes.

As we were interested in studying the transitions of the EEG correlates, the EEG signals were processed asynchronously using sliding windows. In the following subsections, we first present the architecture for signal processing and classification, including previously reported methods used here for comparison. Afterward, we present the performance metrics followed by the validation methods employed in this study.

A. Signal Preprocessing

The EEG and EOG signals were downsampled to 256 Hz and bandpass filtered between 1 and 40 Hz by a 14^{th}-order Butterworth filter with forward and backward processing. The vertical EOG component was computed by subtracting the signal from the sensor between the eyebrows by the average of the other two sensor signals. The horizontal EOG component was derived from the bipolar signal between the two sensors close to the canthi.

Some EEG electrodes were manually removed when they were short-circuited with the CMS or DRL electrode. Out of 48 recordings, CP1, POz, and PO3 were removed once, PO4 was removed twice, and P2 was removed four times. A 20^{th}-order spatial filter, SPHARA [23], was applied to interpolate the signals for the removed electrodes and, more importantly, to reduce any high spatial frequency components, likely corresponding to artifacts [23]. Peripheral electrodes were left out of the analysis to reduce the likelihood of muscular contamination, yielding twenty-five channels centered at Cz, namely, F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, CP4, P3, P1, Pz, P2 and P4. Common-average rereferencing was then applied [24].

To further remove potential EEG artifacts due to eye movements, Independent Component Analysis (ICA) was utilized. The idea was to remove the components similar to the processed EOG signals and then project the remaining components back to the original sensor (electrode) space. ICA needs a hyperparameter indicating the assumed number of independent (source) components. Since there is no golden rule for setting the hyperparameter, we used an iterative approach. Specifically, we performed RUNICA as our base ICA [25], starting with setting the hyperparameter to 15 components and iteratively adding or reducing one component at a time. The iteration stopped with the largest and most valid hyperparameter. An invalid hyperparameter returns a weight matrix with imaginary numbers or whose maximum and minimum values differ by too large a value, for which we picked five as a threshold.

Based on the largest and valid hyperparameters, we obtained the corresponding independent components with RUNICA. If any of these independent components had a Pearson correlation coefficient above 0.7 with the vertical or horizontal EOG component, the independent component was then dropped from future analysis. The selection of relevant components was only performed for each training fold (see Section III-D, the description of the cross-validation). On average, for both folds, 15.4 components were kept for analysis, and 1.6 components were removed when using both phases.

We computed as features the Power Spectral Density (log-PSD) as features using Thomson's multitaper algorithm [26] over a two-second sliding window with a 500-ms shift. We specified the frequency resolution to be 1 Hz and the frequency range to be [2 28] Hz to avoid movement-related slow cortical potentials [27] and too-low signal-to-noise ratios at high frequencies. We rejected windows where any EEG time sample had a peak value larger than 50μV after the previous preprocessing.

B. Classification Method

We evaluated the decoder as a classification (discrete) problem instead of a regression (continuous) problem because we only had limited samples for the possible output values; after each phase, one or two difficulty levels were reported for one condition. This resulted in at most 4 or 8 unique difficulty levels for the four conditions in a phase, while the level ranged between 0 and 100.

We, however, still utilized regression, as it may better describe the perceived difficulty level than descriptive labels (*e.g.* , easy and hard). We regressed from the log-PSD values to the self-reported difficulty level in that trial with a generalized linear model [28] and sparsity regularization. The chosen distribution and link model were binomial and logistic, respectively, and the regularization parameters were scanned with α = [0.15, 0.5, 1] (1 = LASSO, 0 = ridge regression) and nonnegative values of λ = [0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20] for the penalty term.

Based on the fact that a cognitive state is less likely to change instantaneously, we smoothed the output of the regression by a moving average to obtain a more reliable estimation of the state. In particular, we averaged the output with the previous nine samples to reach a compromise between accuracy and latency; since log-PSD-based features were computed every 500 ms, this resulted in a five-second delay. In the special case corresponding to the samples at the beginning of a trial, we padded the data by replicating the first sample of a trial instead of using zeros.

We then discretized the smoothed output onto the number of desired classes at each time window using the thresholds learned by a Linear Discriminant Analysis (LDA) from the training set during the cross-validation (to be detailed in Section III-D).

C. Engagement and Attentional Indices

Different levels of task difficulty are likely to induce different levels of engagement and attention. Several EEG-based metrics have been proposed in the literature to be modulated by these cognitive processes [7].

We evaluated the possibility of decoding user states in this task using previously reported indexes for these variables: the well-known engagement index, defined as the ratio $\frac{\beta}{\alpha+\theta}$ [8], [29]–[31], and the attentional index, defined as $\frac{\theta}{\beta}$ [32]. In this work, we defined the following ranges for each band for computing the two indices: θ band, 4-7 Hz; α band, 8-12 Hz; and β band, 16-28 Hz.

To test the effectiveness of the two indices in the current paradigm, we applied the same signal preprocessing framework to compute log-PSD. Then, we calculated the average power of each band for computing the indices in each window. Based on the two indices computed over the twenty-five selected channels, we established two methods for comparison. One replaces the log-PSD feature by the two indices and then applies the same classification method as in Section III-B. The second approach is similar to the first, except that we replace the regressor with an LDA classifier and use the predicted posterior probabilities as output for subsequent smoothing and classification.

D. Measurement of Decoder Performance

Decoding performance was assessed using class-balanced accuracy to take into account the effect of imbalanced sample sizes between classes; the accuracy for each class was independently computed, and all were averaged together, which means that an equal weight was assumed for each class. As a

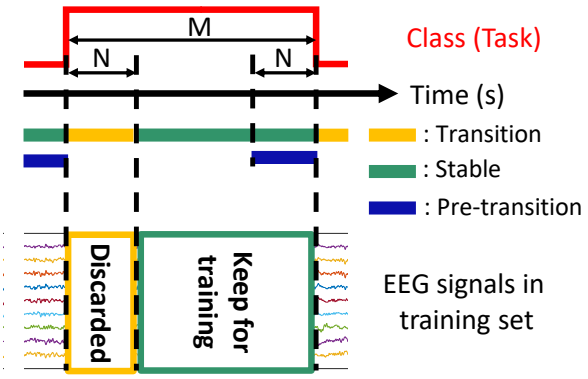


Fig. 2: Analysis of decoding accuracy in the transition period. The three periods are defined using one task as an example. M is the duration of the task, while N is the assumed duration of the transition. For training the decoder, only the signals from the stable period were used. The test accuracies were computed for each period for different values of N with test sets.

result, the theoretical chance level was $1/k$, with k being the number of classes.

To distinguish different conditions and characterize the transition period, we performed two-phase cross-validation in each recording. In other words, either the discrete or continuous phase served as the training data, while the other phase was treated as the test data. It is important for this type of study not to use a cross-validation method that performs random partitioning since it may result in producing training and test samples that are close in time, which would violate the principle of independent training and test sets [33] and thus yield optimistic performance estimations. To ensure rigorousness, we also estimated the model related to ICA preprocessing in the training folds.

IV. Transition Analysis

A. Transition Period

We assume that N seconds are needed for the EEG to reach a *stable* state (*i.e.*, a state where the cognitive state can be reliably decoded) from the onset of the changing condition. If so, then the neural activity in the first N seconds will not reflect the new state. Hence, a decoder would benefit from not using the data collected during the transition period. This hypothesis was tested by analyzing the class-balanced accuracy with different transition lengths (N).

For the analysis, three periods were defined around the moment the task changed, as drawn in Fig. 2. The transition period was defined as the first N seconds after the task change. A second stable period spans from the end of the transition period until the next task change (M-N seconds, where M is the length of the condition). The last period is a pretransition period, which is a subset of the stable period and corresponds to the last N seconds before the task change. The pretransition period was defined as a control term where the accuracy should be higher than in the transition period in case the hypothesized transition effect exists.

Given the assumption that there are N seconds of transition, the information content of the signals during this period is not obvious: it can correlate to the previous state, the next state, or some in-between state. Therefore, only the data in the stable period were used for training; *c.f.*, Fig. 2. This approach ensures that the training samples are rather reliable and can yield an optimal model. Testing performance is reported as the class-balanced accuracy for each period and different N values.

According to the hypothesis, the EEG correlates will only reliably reflect the current state after the transition period. Thus, it is expected that the accuracy curve of the transition period should improve as N increases. One reason is that the decoder may have been trained with a larger portion of reliable samples. Another reason is that, as the tested N is larger than the real N, more reliable samples from the true stable period are being tested in the assumed transition period, which is larger than the true transition period. As the real N is unknown, the two reasons are inseparable. If the underlying hypothesis does not hold, the accuracy curves should be flat for all three periods. This means that there is no need to discard samples immediately after the task change to train the decoder models. In other words, the neural activity will instantaneously change to reflect the new user state.

B. Estimating the Latency

Once the transition period is confirmed, the next objective is to estimate the latency. The idea is to choose a sufficiently large N (60 s) such that the training data only belong to the stable period. Then, a two-second sliding window is used to compute the class-balanced accuracy for the transition periods with the test set.[1] The issue involving the use of all the data to assess the class-balanced accuracy is that one cannot discern when the decoder output becomes stable. As a result, computing the class-balanced accuracy inside a smaller window can better reflect the exact time point: if a window contains only stable data, the accuracy should be higher than that of windows containing any data from the transition period.

However, the sliding-window approach still cannot exclude the latency from signal processing, especially the latency from postprocessing.[2] To exclude this latency unrelated to user behavior, forward-backward, instead of just forward, postprocessing was conducted. This theoretically neutralizes the latency in signal processing, although it is not feasible for online decoding.

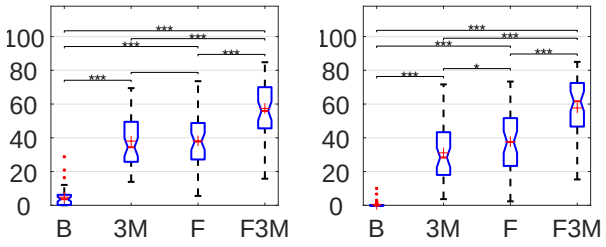
V. Results

A. Self-Reporting Questionnaires

Fig. 3 shows the reported averaged workload and perceived difficulty levels for the different conditions in the two recording sessions. Each box corresponds to 48 data points; within-session values were averaged across the three trials. The workload and perceived difficulty levels increase

[1]The two-second window includes any log-PSD window that starts inside the two-second window.

[2]Delay from the spectral filter can be neglected since forward-backward processing was used.



(a) Workload (b) Difficulty

Fig. 3: Subjective assessment averaged across trials. Two-tailed t-tests for 48 samples were applied (*: $p < 0.05$, and ***: $p < 0.001$). The red crosses indicate mean values. The red horizontal lines indicate median values, and the confidence intervals are represented by the notches. Red dots are outliers.

TABLE I: Median values of the engagement and attentional indices, where values in boldface highlights a significant difference from the baseline condition.

		Best Subject				Median Subject			
		B	F	3M	F3M	B	F	3M	F3M
Eng. Idx	Pz	0.27	**0.27**	0.27	0.27	0.14	**0.10**	**0.11**	**0.12**
	Fz	0.27	**0.29**	**0.29**	**0.30**	0.15	**0.14**	**0.12**	0.14
	C4	0.28	**0.29**	**0.30**	**0.31**	0.15	**0.11**	**0.09**	0.14
	CP4	0.29	**0.30**	**0.30**	**0.31**	0.17	**0.13**	**0.12**	**0.13**
Atten Idx	Pz	1.80	**2.06**	**1.97**	**2.04**	3.93	5.35	4.93	5.03
	Fz	1.78	**1.96**	**1.85**	**1.84**	3.75	**4.40**	5.35	4.40
	C4	1.54	**1.92**	**1.67**	**1.79**	3.04	4.24	5.24	**4.32**
	CP4	1.46	**1.78**	**1.64**	**1.77**	2.93	4.15	4.51	4.44

in the order of B, 3M, F, and F3M. One-way ANOVA was conducted for the four conditions on workload levels with $F(3, 188) = 111.63, p = 1.5 \times 10^{-41}$ and perceived difficulty levels with $F(3, 188) = 107.83, p = 1.2 \times 10^{-40}$. Two-tailed t-tests yielded significant differences in most cases ($p < 0.001$, n = 48). The only exception was the workload level between 3M and F, but this was not the case for difficulty level ($p < 0.05$).

The Pearson correlation coefficient between workload level and difficulty level for each condition had a p-value < 0.001 ($n = 48$). The correlation coefficients for trials 1, 2, and 3 were, respectively, 0.53, 0.57, and 0.56 for B, 0.81, 0.83, and 0.82 for F, 0.88, 0.83, and 0.88 for 3M, and 0.77, 0.86, and 0.83 for F3M. The high correlations suggest that both the reported workload level and difficulty level are highly intertwined, and using either of them as the ground truth should not result in large differences regarding the decoding approach.

B. Engagement and Attentional Indices

TABLE I lists the engagement and attentional indices averaged from all the computed windows for the four different conditions. Each quadrant provides one index over four representative channels for the best or median recording (in terms of accuracy, *c.f.*, Indices + LDA in Section V-C). A number written in boldface indicates a significant difference between B and the condition (two-tailed t-test, all data, $n > 1,000$). Significant differences were obtained for nearly all the cases for both indices for the best recording. However, for the attentional index of the median recording, significant differences were found only for two cases. This suggests that α power seems to be more important because the main difference between both indices is the presence of α in the denominator.

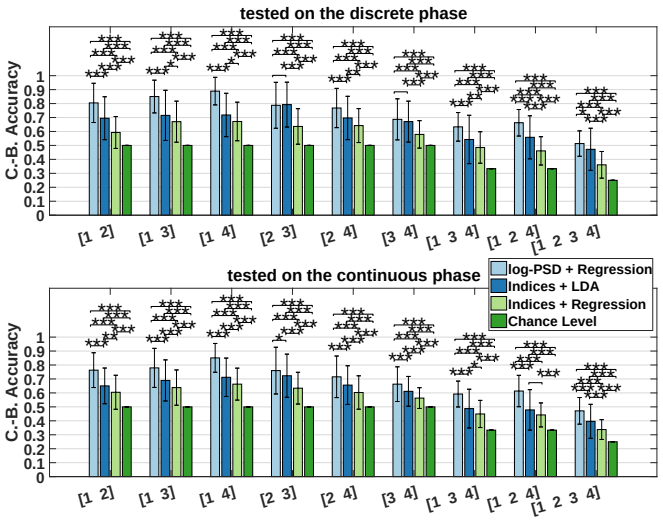


Fig. 4: Accuracies for the two-phase cross-validation. Error bars indicate standard deviations over 48 samples. A one-sample t-test (n=48) was applied to each pair of methods (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

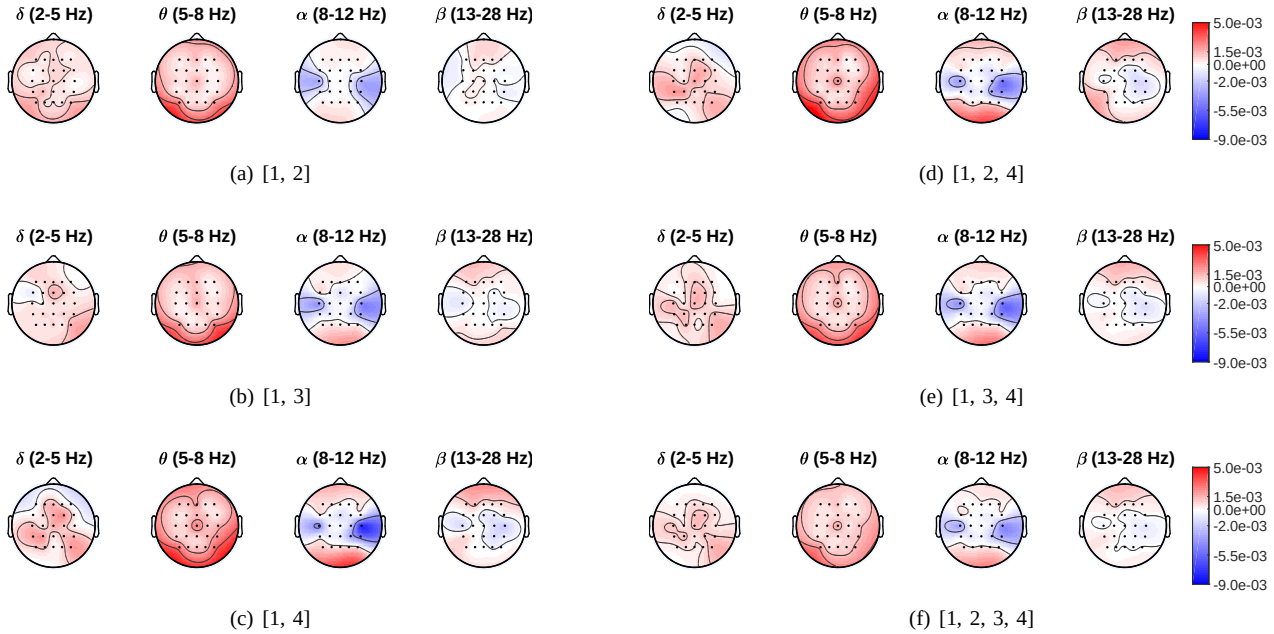
We computed the indices for all twenty-five channels, as has been done in the literature for all available channels without any specific channels being specified or referred to as prominent [8], [29]–[32]. However, we only reported four channels in TABLE I for the sake of space. The channels Pz, Fz, C4, and CP4 were selected based on previous studies of working memory (related to the mapping task) [9] and visuomotor workload (flying a drone with different sizes of waypoints) [16], where more difficult tasks yielded higher engagement and attentional indices in general.

C. Decoding Conditions Based on Difficulty Levels

Fig. 4 compares the class-balanced accuracies when decoding the tested classes using the four methods:

- log-PSD + Regression: the method in Section III-B.
- Indices + LDA: LDA classification using the engagement and attentional indices as features.
- Indices + Regression: the method in Section III-B based on the two indices.
- Chance Level: the theoretical accuracy from random guessing.

The top chart shows the results obtained in the discrete phase (trial 1), and the bottom plot shows the performance in the continuous phase (trials 2 and 3 combined). One-way ANOVA showed the effect of different methods on the class-balanced accuracy for testing on the discrete phase with $F(3, 1724) = 293.1, p = 9.9 \times 10^{-154}$ and for testing on the continuous phase with $F(3, 1724) = 219.7, p = 1.1 \times 10^{-120}$. Pairwise one-sample t-tests (n = 48) were performed between different methods for all the explored combinations of conditions. In most cases, decoding based on the engagement and attentional indices performed worse than the log-PSD-based framework.

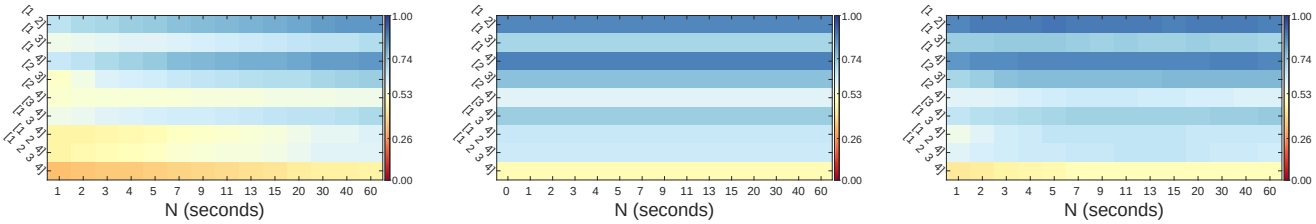


(a) [1, 2] (d) [1, 2, 4]

(b) [1, 3] (e) [1, 3, 4]

(c) [1, 4] (f) [1, 2, 3, 4]

Fig. 5: Regression coefficients averaged across 48 recordings. Red (blue) means that a low (high) log-PSD power favors the easiest condition below each subfigure. The values range between 5×10^{-3} (red) and -9×10^{-3} (blue).



(a) Transition period of the discrete phase (b) Stable period of the discrete phase (c) Pretransition period of the discrete phase

Fig. 6: Cross-validation accuracies averaged across 48 recordings. Only the transition periods have increasing trends (dark red to bright yellow to dark blue).

The proposed method as well as Indices + LDA and Indices + Regression generally had a higher accuracy when tested on the discrete phase than the continuous phase, probably because a shorter task time can have a more stable cognitive state [16]. For the proposed method, one-way ANOVA showed the effect of different phases on the class-balanced accuracy with $F(1, 862) = 14.5, p = 2.0 \times 10^{-4}$. Significant differences (one-sample t-test, n = 48) were found between both phases in most condition sets, except sets [2 3] and [3 4].

D. Neural Correlates

Fig. 5 shows the regression coefficients of the classifier over log-PSD averaged across all recordings, where the coefficients of a recording were trained from all trials. In each subfigure, a topographic plot shows the coefficients averaged for one frequency band, where the β band was defined as [13 28] Hz to include all frequency bins in the analysis. Negative coefficients (shown in blue) mean that low log-PSD values favor the most difficult condition indicated below the subfigure, while positive values (in red) embrace the baseline condition. White represents a zero coefficient (*i.e.*, the feature carries little or no information about the target output).

As seen from Fig. 5(a) and 5(b), while using the easiest condition (B, except in one recording) as the reference, the second easiest condition (3M in 29 out of 48 recordings) and the third easiest condition (F in 30 out of 48 recordings) both have a common negative pattern over the left and right sensorimotor regions within the α band, in particular C3, C4, and CP4. The parietal side is also similar but with smaller coefficients. Both conditions also share a common positive pattern within the δ band at FCz. For the hardest condition (F3M) (see Fig. 5(c)), the patterns are similar to the other two conditions with higher absolute coefficients. In addition, within the θ band, Cz is stronger, and the β band in the right sensorimotor region is more pronounced. In the case of three- and four-class classification (Fig. 5(d), 5(e), and 5(f)), the patterns are similar to [1 4] but with smaller coefficients. In short, the most useful patterns are the α bands at C3, CP4, and C4, the δ band at FCz, and the θ band at Cz.

E. Transition Analysis

Fig. 6 provides the cross-validation results for the transition analysis. From left to right, the results in the subfigures correspond to the transition, stable, and pretransition periods. Each subfigure reports the class-balanced accuracies for the nine sets of conditions for $N = 0, 1, 2, 3, ..., 60$ seconds. Whenever we tested the discrete phase, we considered transitions from rest to each condition, B, 3M, F, or F3M. When we tested the continuous phase, a transition was either between rest and baseline or between two conditions.

In the transition period, we can observe that the accuracy improved by more than 10% as N increased. In contrast, this pattern appears in neither the stable nor the pretransition period. The case N = 0 means that there was no assumed transition period, and therefore decoders were trained without discarding any data.

Fig. 7 provides more precise information regarding the latency by computing accuracies through a two-second sliding window. The red and blue curves represent the defined pretransition and transition periods, respectively. The curves show the mean cross-validation accuracies over the 48 recordings, and the shaded areas plot the 95% confidence interval of the mean values with the assumption of a Gaussian distribution [34]. Therefore, if two shaded areas do not overlap, the difference between the two periods is statistically significant.

Fig. 7(a) provides information on total latency, including the latency contributed by the decoder.[3] The transition curves reached a rather stable state between 6 and 10 seconds, depending on the targeted classes. Significant differences between the pretransition and transition periods are also evident before 8 seconds in most cases.

Fig. 7(b), on the other hand, excludes the latency from the decoder by using forward-backward postprocessing. Increasing trends are observed in each set, and the latencies are between 4 and 8 seconds. Significant differences between the two curves are evident before 3 seconds in most cases.

VI. Discussion

We tested the EEG decoding of different levels of perceived difficulty in a dynamically changing task. Subjects consistently reported significantly different levels for the conditions, supporting the suitability of the experimental design for our study.

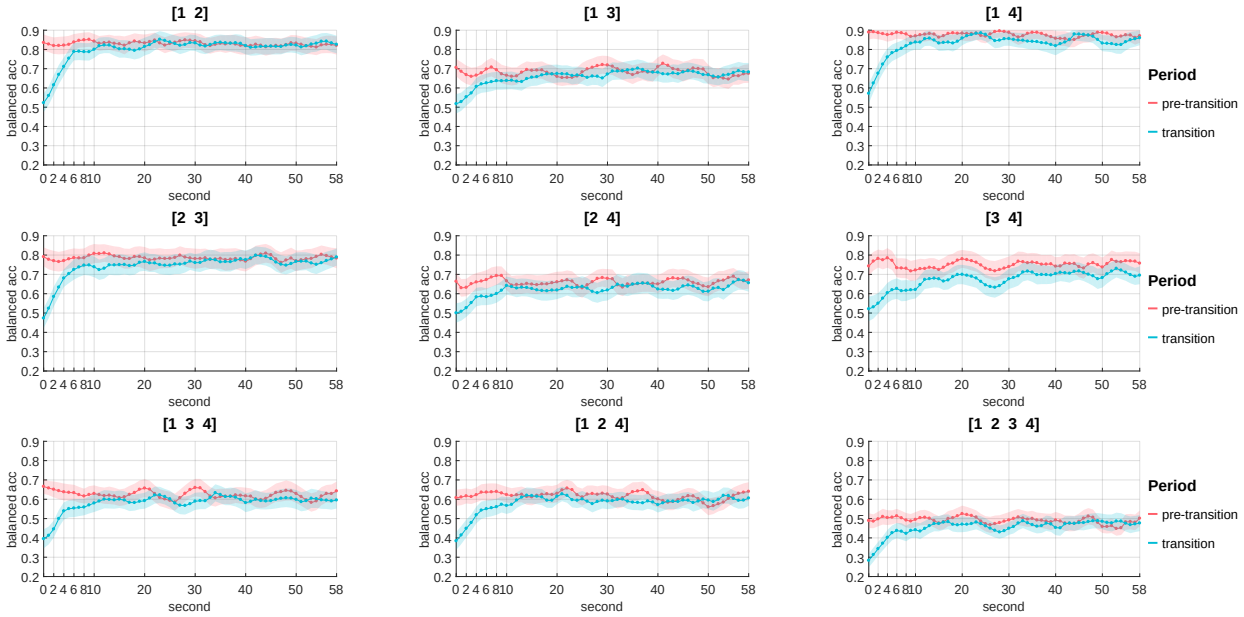
The compound of cognitive states is distinguishable even if the states were induced from different tasks. In contrast, the engagement and attentional indices were sufficient for building a preliminary decoder, but the validation accuracies, in general, were significantly worse than those of the proposed method. This trend supports the hypothesis that realistic tasks induce a compound of cognitive states that need to be analyzed together. Future research, nevertheless, could benefit from experimental designs that selectively influence a specific type of cognitive state at particular times to better characterize their unique physiological correlates.

One critical question is whether the EEG decoder truly decodes the compound cognitive state or motor activities. First, the analysis of our regression model (*c.f.*, Fig. 5) reveals stronger α power modulations in the sensorimotor region when both hands are used. The signals from right-hand side, from C4 and CP4 in particular, are consistent with a previous study in which subjects performed the flying task with their right hand, as opposed to this study, where the left hand was used [16]. This implies that the activity in the right sensorimotor region may convey information about difficulty level in the navigation process, but a more rigorous study is needed to identify the underlying cognitive component. One of the possible cognitive states being captured by the EEG signals is the attentional level. The attentional level has been reported in other studies of visuomotor tasks as correlated with the α band in the parietal region and in frontal θ activity [35]. Our regression model also somehow captured them, in particular the parietal α band. In addition, working memory may also be considered in this study due to the use of the mapping task. The θ band has been frequently associated with working memory [9], [36], [37]; this is coherent with the pronounced θ at Cz in our case. Therefore, we believe our decoders captured a sufficient amount of cognitive information and perhaps motor-related activity to a certain degree, which is modulated by the complexity of the task [38]; furthermore, studies on brain-damaged patients suggest that the right motor hemisphere is specialized to position control, which varies with task complexity [39].

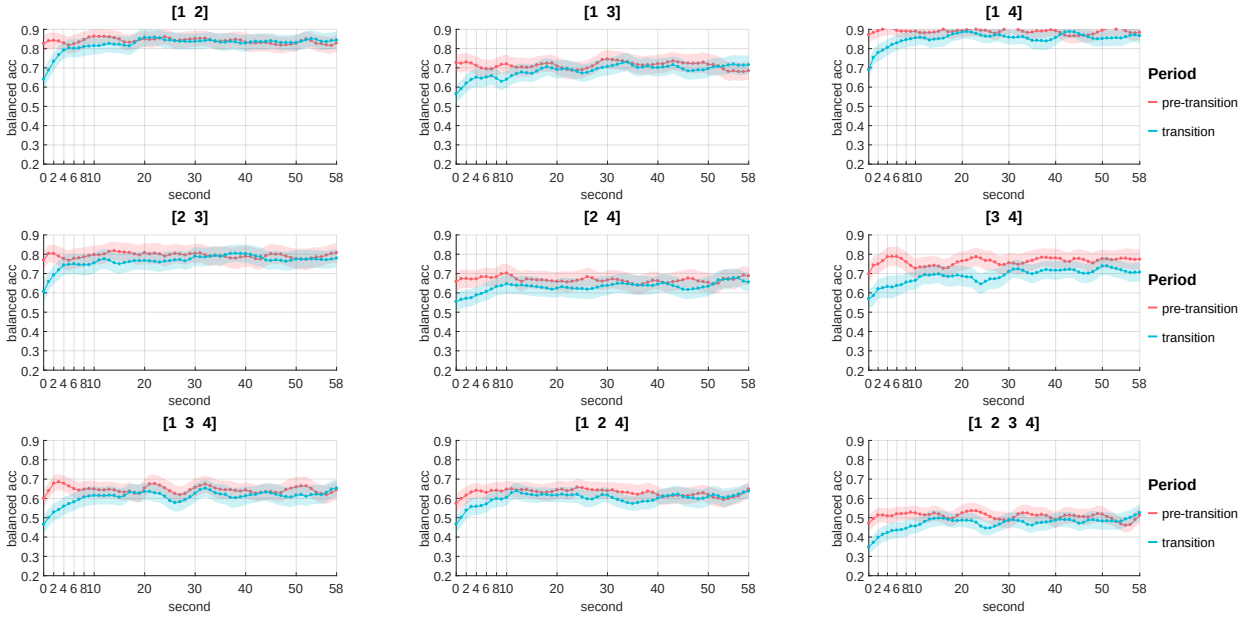
Our results strongly support the existence of a transition period in the EEG correlates after a task change. This transition, together with potential latency from the decoder, is likely to last approximately 6 to 10 seconds before the decoder can reliably provide accurate information about the cognitive state in the evaluated tasks. The latency after minimizing the effect of the decoder is approximately 4 to 8 seconds. It is worth noting that the stable period has very similar results regardless of N. One may expect the test accuracy with N = 0 in the stable period to be lower because the first few samples would actually come from the transition period and be more likely to be misclassified. Indeed, this situation could occur, but 11 seconds out of 180 or 300 seconds is a rather small proportion. This small amount of data from the transition period was unlikely to largely affect the developed model and its accuracies. Hence, the high accuracy with N = 0 does not contradict the hypothesis.

The transition time can reflect the fact that the neural signature and, in turn, the decoder need time to reliably reflect the cognitive state, even if the subject could immediately adapt to a new task. In addition, it is also plausible that the subject needs time to gather information about the new task and its difficulty, effectively delaying changes in her cognitive state. At this point, it is not straightforward to estimate how much each factor contributes to the observed latency. Nonetheless, irrespective of the specific sources of the delay, this transition period should be carefully considered in the design of HMI systems, in particular those envisioning adaptation of the interaction dynamics based on the user's state estimation.

[3]One missing latency is that corresponding to the spectral filtering, as we used forward and backward processing.



(a) Forward processing



(b) Forward and Backward processing

Fig. 7: Estimation of overall and user-related latencies by windowed analysis of accuracy. (a) The estimation is based on forward processing and thus represents the overall latency. (b) The estimation uses forward and backward processing, which excludes the latency caused by signal processing. Nonoverlapping shaded areas indicate significant differences between both periods.

VII. Conclusion

It is possible to decode the perception of difficulty level and, in turn, changes in the underlying cognitive processes. The proposed method performed significantly better than the two traditional indices. However, the applicability of the proposed method in real applications remains to be validated with a closed-loop experiment.

We further confirmed that a transition period exists in the EEG signals and provides a rough estimation of the required time to reach a stable state, either with or without the latency from the decoder. Depending on the applications, designers of HMI systems should take this information into account.

Author Contributions

P.-K.J. contributed to the writing of the manuscript, generation of the figures, data analysis, design of the experiments, and collection of the data. R.C. supervised the findings of this work and contributed to the design of the experiments as well

as revisions of the manuscript. F.D. contributed to the design of the experiments, collection of the data, and revisions of the manuscript. A.A. contributed to the design of the experiments and revisions of the manuscript. D.A. supervised the design of the experiments, collection of the data, and revisions of the manuscript. J.d.R.M. supervised the findings of this work as well as revisions of the manuscript.

REFERENCES

[1] J. C. De Winter, R. Happee, M. H. Martens, and N. A. Stanton, "Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 27, pp. 196–217, 2014.

[2] K. C. Ewing, S. H. Fairclough, and K. Gilleade, "Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop," *Frontiers in Human Neuroscience*, vol. 10, p. 223, 2016.

[3] S. R. Wolfe, "Supporting air traffic flow management with agents." in *AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants*, 2007, pp. 137–138.

[4] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. K. Jacob, "Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 5372–5384.

[5] J. Faller, J. Cummings, S. Saproo, and P. Sajda, "Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task," *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 6482–6490, 2019.

[6] M. A. Guadagnoli and T. D. Lee, "Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning," *Journal of Motor Behavior*, vol. 36, no. 2, pp. 212–224, 2004.

[7] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, 2018.

[8] A. T. Pope, E. H. Bogart, and D. S. Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," *Biological Psychology*, vol. 40, no. 12, pp. 187 – 195, 1995.

[9] A.-M. Brouwer, M. A. Hogervorst, J. B. F. van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *Journal of Neural Engineering*, vol. 9, no. 4, p. 45008, 2012.

[10] P. G. Jorna, "Spectral analysis of heart rate and psychological state: A review of its validity as a workload index," *Biological psychology*, vol. 34, no. 2-3, pp. 237–257, 1992.

[11] P. Jorna, "Heart rate and workload variations in actual and simulated flight," *Ergonomics*, vol. 36, no. 9, pp. 1043–1054, 1993.

[12] U. Ahlstrom and F. J. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," *International Journal of Industrial Ergonomics*, vol. 36, no. 7, pp. 623–636, 2006.

[13] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, 2008, pp. 69–72.

[14] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, 2002.

[15] C. Collet, C. Petit, A. Priez, and A. Dittmar, "Stroop color–word test, arousal, electrodermal activity and performance in a critical driving situation," *Biological psychology*, vol. 69, no. 2, pp. 195–203, 2005.

[16] P.-K. Jao, R. Chavarriaga, and J.d.R. Millán, "Analysis of EEG correlates of perceived difficulty in dynamically changing flying tasks," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2018.

[17] P. Aricò, G. Borghini, G. D. Flumeri, A. Colosimo, I. Graziani, J. Imbert, G. Granger, R. Benhacene, M. Terenzi, S. Pozzi, and F. Babiloni, "Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks," in *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 7242–7245.

[18] J. S. Macdonald and N. Lavie, "Visual perceptual load induces inattentional deafness," *Attention, Perception, & Psychophysics*, vol. 73, no. 6, pp. 1780–1789, 2011.

[19] B. Fowler, "P300 as a measure of workload during a simulated aircraft landing task," *Human Factors*, vol. 36, no. 4, pp. 670–683, 1994.

[20] M. W. Miller, J. C. Rietschel, C. G. McDonald, and B. D. Hatfield, "A novel approach to the physiological measurement of mental workload," *International Journal of Psychophysiology*, vol. 80, no. 1, pp. 75–78, 2011.

[21] F. Dell'Agnola, L. Cammoun, and D. Atienza, "Physiological characterization of need for assistance in rescue missions with drones," in *IEEE International Conference on Consumer Electronics*. IEEE, 2018, pp. 1–6.

[22] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in Psychology*, vol. 52, pp. 139–183, 1988.

[23] U. Graichen, R. Eichardt, P. Fiedler, D. Strohmeier, F. Zanow, and J. Haueisen, "SPHARA-a generalized spatial fourier analysis for multi-sensor systems with non-uniformly arranged sensors: Application to EEG," *PloS one*, vol. 10, no. 4, p. e0121741, 2015.

[24] O. Bertrand, F. Perrin, and J. Pernier, "A theoretical justification of the average reference in topographic evoked potential studies," *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, vol. 62, no. 6, pp. 462–464, 1985.

[25] S. Makeig, T. Bell, T. Lee, T. Jung, S. Enghoff *et al.*, "EEGLAB: ICA toolbox for psychophysiological research," *Swartz Center for Computational Neuroscience, Institute of Neural Computation, University of San Diego California*, 2000. [Online]. Available: www.sccn.ucsd.edu/eeglab

[26] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[27] G. Garipelli, R. Chavarriaga, and J. d. R. Millán, "Single trial recognition of anticipatory slow cortical potentials: The role of spatio-spectral filtering," in *Proceedings of the 5th International Conference on Neural Engineering*, 2011.

[28] A. J. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*. CRC press, 2008.

[29] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 6, pp. 1052–1063, Nov 2011.

[30] D. Szafir and B. Mutlu, "Pay attention!: Designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 11–20.

[31] M. Andujar and J. E. Gilbert, "Let's learn!: Enhancing user's engagement levels through passive brain-computer interfaces," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 703–708.

[32] P. Putman, B. Verkuil, E. Arias-Garcia, I. Pantazi, and C. van Schie, "EEG theta/beta ratio as a potential biomarker for attentional control and resilience against deleterious effects of stress on attention," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 14, no. 2, pp. 782–791, 2014.

[33] G. Varoquaux, P. R. Raamana, D. A. Engemann, A. Hoyos-Idrobo, Y. Schwartz, and B. Thirion, "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, 2017.

[34] P. Morel, "Gramm: Grammar of graphics plotting in Matlab." *Journal of Open Source Software*, vol. 3, no. 23, p. 568, 2018.

[35] Y.-K. Wang, T.-P. Jung, and C.-T. Lin, "Theta and alpha oscillations in attentional interaction during distracted driving," *Frontiers in Behavioral Neuroscience*, vol. 12, p. 3, 2018.

[36] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, vol. 29, no. 23, pp. 169–195, 1999.

[37] S. Raghavachari, J. E. Lisman, M. Tully, J. R. Madsen, E. Bromfield, and M. J. Kahana, "Theta oscillations in human cortex during a working-memory task: evidence for local generators," *Journal of Neurophysiology*, vol. 95, no. 3, pp. 1630–1638, 2006.

[38] P. Manganotti, C. Gerloff, C. Toro, H. Katsuta, N. Sadato, P. Zhuang, L. Leocani, and M. Hallett, "Task-related coherence and task-related spectral power changes during sequential finger movements," *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control*, vol. 109, no. 1, pp. 50–62, 1998.

[39] K. Y. Haaland, J. L. Prestopnik, R. T. Knight, and R. R. Lee, "Hemispheric asymmetries for kinematic and positional aspects of reaching," *Brain*, vol. 127, no. 5, pp. 1145–1158, 2004.