EEG-Based Online Regulation of Difficulty in Simulated Flying

Ping-Keng Jao, Ricardo Chavarriaga, *Senior Member, IEEE*, and José del R. Millán, *Fellow, IEEE*

Abstract—Adaptively increasing the difficulty level in learning was shown to be beneficial than increasing the level after some fixed time intervals. To efficiently adapt the level, we aimed at decoding the subjective difficulty level based on EEG signals. We designed a visuomotor learning task that one needed to pilot a simulated drone through a series of waypoints of different sizes, to investigate the effectiveness of the EEG decoder. The EEG decoder was compared with another condition that the subjects decided when to increase the difficulty level. We examined the decoding performance together with behavioral outcomes. The online accuracies were higher than the chance level for 16 out of 26 cases, and the behavioral results, such as task scores, skill curves, and learning patterns, of EEG condition were similar to the condition based on manual regulation of difficulty.

Index Terms—EEG, Real-time decoding of difficulty, Closed-loop adaptation, (Simulated) Flying

1 INTRODUCTION

ENGAGEMENT in the task is a crucial factor for learning and to facilitate positive outcomes [1], [2], [3]. One way to keep the engagement is keeping the task at a proper difficulty level for the person's skills. This can be formalized by the theory of challenge point that outlines that the dependency between task performance or potential learning outcome and the task difficulty follows an inverted-U curve [4]. Easy tasks may disinterest people while high levels of difficulty will lead to frustration; both cases hinder the learning process. As a result, the performance or learning outcome is optimized if one can operate at the best difficulty level, a condition referred to as the sweet spot.

A specific aspect in the theory of challenge point is that the independent variable is functional task difficulty. The term functional indicates that the task difficulty should be subject-dependent, because novices and experts behave differently on a task with the same objective task difficulty. Therefore, a continuous adaptation of the difficulty level is essential to keep a person operating around their sweet spot.

Many applications, such as education and gaming, may benefit from operating around the sweet spot. However, estimating the sweet spot is not trivial because the subjective difficulty level is time-varying and is a function of current skills, cognitive states, and affective states [4]. This assessment usually requires human intervention, either from the

learner herself who should concentrate on learning or from an experienced instructor who is rare. An alternative is to automatically estimate their subjective difficulty.

We hypothesize that physiological signals correlates with the functional task difficulty and utilizing such correlates helps stay in the sweet spot. Therefore, we aim at building a Brain-Machine Interface (BMI) regulator, able to estimate the functional difficulty and, to adapt the difficulty level. We test the feasibility of building such a BMI with a visuomotor task as they are at the core of a wide range of applications. In particular, our task allows subjects to practice their piloting skill with different functional difficulty levels defined by their skills.

During interaction, either the BMI regulators or learners manually decide whether to increase the difficulty level or keep it unchanged. The reason is that, for learning, it is common not to decrease the difficulty level during a session because user's skills are unlikely to decline. Human training literature also supports our choice; the difficulty level in learning should remain unchanged and increase only when considered necessary [5]. Importantly, adaptively increasing the difficulty level has been shown effective in promoting learning, while increasing it at a fixed rate was not [5]. Therefore, the presented functional difficulty level increases only when the current level is considered to be easy.

In this study, we compare the subject's performance when the difficulty level was set by their explicit inputs, or automatically regulated by the BMI. In addition, a critical question is to uncover the neural correlates of the subjective difficulty level. We do so by analyzing the features in the offline session. Regarding the evaluation of the BMI against the manual condition, subject performed the piloting task in two sessions, each session with the two conditions. Then, we compare how well learners performed (task score), how their skills evolved (skill curve), and how the difficulty level changed over time (learning patterns). Our results confirm the feasibility of a BMI regulator to support skill learning as its performance is not significantly different than in the

- Ping-Keng Jao is with the École Polytechnique Fédérale de Lausanne, Geneva, Switzerland (email: ping-keng.jao@alumni.epfl.ch).
- Ricardo Chavarriaga is with the ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland, and also with the École Polytechnique Fédérale de Lausanne (email: ricardo.chavarriaga@zhaw.ch).
- José del R. Millán is with the Department of Electrical and Computer Engineering & the Department of Neurology, University of Texas at Austin and also with École Polytechnique Fédérale de Lausanne. (email: jose.millan@austin.utexas.edu).

Manuscript received Sept. DD, 2020; revised Month DD, 2021. This work has been supported by the Swiss National Centres of Competence in Research (NCCR) Robotics. Digital Object Identifier 00.0000/????.0000.0000000

manual condition.

The article starts a literature survey in Section 2. Collection of data is presented in Section 3. The EEG decoder and methods for analysis of behavioral results are addressed in Section 4 and 5, respectively. Results are reported in Section 6, followed by a discussion and conclusion in Sections 7 and 8, respectively.

2 STATE-OF-THE-ART

Learners' perception of difficulty has a strong correlation to workload level [6]. For decoding workload level, there already exist several studies utilizing measurements from task performance to physiological signals [7]. Physiological signals used to decode workload include heart rate [8], [9], pupil dilation [10], [11], [12], electrodermal activity [13], [14], EEG signals [6], [10], [15], [16], [17], and Functional Near-Infrared Spectroscopy (fNIRS) [18]. Among different signals, we particularly focus on EEG because of their better time resolutions and latency compared to other signals.

Although many researchers have been working on decoding workload/difficulty from EEG signals, only a few studies conducted closed-loop experiments. To the best of our knowledge, Pope *et al.* conducted one of the earliest online experiments using EEG signals and reported workload-relevant EEG frequency features, where the task was Multi-Attribute Task Battery [19]. Their online evaluations mainly concerned whether each index of interest can induce expected oscillations of the difficulty levels. Szafir *et al.* tried to detect a sudden drop of engagement, in order to find the proper timing for a new educational robot to attract the attention of a student [17]. However, the result was not significantly better than a control group. Yuksel *et al.*, although not using EEG signals, combined both fNIRS and task performance to improve learning experience of piano [18]. The beginners in learning piano had benefited more in both self-experience and task performance, but this was not necessarily the case for the experienced players. Their approach, however, can hardly be extended to other cases as the system was extremely tailored. For example, each subject and each objective, instead of subjective, difficulty level has an individual threshold that was defined from many pilot studies and also involved brain data from her own training task. Ewing *et al.* investigated the trends in EEG signals by varying the game demand and mental effort. The trends were then modeled to adapt the level of Tetris [15]. The EEG conditions yielded unstable changes of levels and worse immersive experience than self decisions. A potential way to improve is incorporating advanced decoders because they only compared the θ band (4-8 Hz) at Fz and upper α band (10.5-13 Hz) at P4 with a baseline activity. Aricò *et al.* tried to adaptively trigger an assistive system to reduce workload level when overloaded, but the adaptive system was not significantly better than without the system [16]. Walter *et al.* compared the learning outcomes in octal arithmetics questions between an EEG adaptation system and an error-adaptive system that adjusts the level based on the number of wrong answers [20]. The EEG system was competitive but needed a long observation window of about 80 seconds. Recently, Faller *et al.* provided feedback of arousal from EEG signals to subjects [10]. The arousal

state can be linked to task difficulty, because arousal is the independent variable of the Yerkes-Dodson curve that has the same shape as the theory of challenge point [21]. In the study of Faller *et al.*, subjects had arousal feedback when performing an engaging task that is different from usual learning process, because the task does not tolerate an error [10]. They showed that consciously self-regulating arousal toward lower levels improves task performance. Nevertheless, low-skilled subjects were excluded after screening.

Although some studies had positive outcomes, there is still space to improve. For example, although Walter *et al.* yielded promising results [20], it is still unclear whether the decoder can function with a much shorter observation window and whether using an EEG decoder is still competitive in applications that need a short latency for feedback. Although Faller *et al.* used a similar task as our study [10], their task was less realistic; their subjects only had one degree of freedom in piloting, and the task stopped after one mistake – contrarily to usual learning practice. In addition, most works only examined the behavioral outcome without evaluating the decoding accuracy during online interaction. Our study design considers all these issues.

3 MATERIALS

3.1 Participants

Thirteen subjects (eight females; Mean age 22.6; SD 1.04) participated in the study. The protocol was approved by the local ethical committee and all the subjects provided written consent. All subjects had normal or corrected-to-normal vision and reported no history of motor or neurological disease, except one subject who previously experienced vasovagal syncope but not during this study.

3.2 Recording Setup

We used a Biosemi ActiveTwo amplifier to record EEG and Electrooculography (EOG) signals at 2,048 Hz. Sixty-four EEG electrodes were placed according to the international 10-10 standard. Three additional channels were placed in the middle of the eyebrows and on the bulge bones below the outer sides of canthi to measure the EOG.

Subjects sat comfortably in front of a twenty-inch screen showing the protocol with 1680x1050 resolution. They hold a joystick (Logitech Extreme 3D Pro) to provide inputs to the protocol. One female subject was left-handed but reported being comfortable using a right-handed joystick.

3.3 Task

The subjects were instructed to pilot the simulated drone through a series of circular waypoints. The subjects controlled the roll and pitch while the drone had a constant velocity, 11.8 arbitrary unit (A.U.) per second, when flying straightly. Figure 1 is a screen-shot of the protocol. The simulated environment was implemented with Unity (<https://unity3d.com/>). The green circle is the current waypoint to fly through while the next one appears in blue. The purple cross at the center was used to determine whether the drone was inside (hit) or outside (miss) a waypoint when passing by. In either case, the current waypoint disappears and the next waypoint becomes the green one. There were

Figure 1: Protocol screen-shot in the offline session. The subjects tried to pilot the simulated drone with the central cross inside the green waypoint when passing through. The blue waypoint is the waypoint after the green one. The numbers on top, from left to right, are points, the number of hits, the radius of the green waypoint, and the elapsed time.

16 levels of difficulty defined by the radius of the waypoint (details in Section 3.4), and a hit of level k rewarded k points. The purple arrow below was a 3D indicator pointing to the center of the current waypoint. The numbers on top, from left to right, are points, the number of hits, the radius of current waypoint, and the elapsed time.

In the case of online sessions (see Section 3.5), the radius (difficulty level) could decrease (increase) or stay the same after every other four waypoints, forming a decision group. This resulted an average of 10.3 seconds to finish a group, so that subjects had some time to be accustomed to a level. The final one of every four waypoints is called as a decision point, and the circle becomes yellow as an indication for the subjects. The subjects were additionally instructed to press the button when the current level is easy, as a way to collect ground truth for decoding or to proceed with the self-paced learning. During piloting, the numbers of points and hits were hidden to avoid that subjects focused solely on optimizing points. On the other hand, we displayed the current level and radius at where the red number is shown in Figure 1. A ☺ symbol also appears next to the radius if the current level is indicated as easy by the subject.

3.4 Personalizing Difficulty Levels

Personalizing levels of objective difficulty (radii) for operating on the same levels of functional task difficulty is necessary. This allows to properly compare the behavioral data in the online sessions. We conducted the personalization before the start of each offline sessions and the start of each condition in the online sessions (see Section 3.5). The personalization was based on a skill evaluation process that collected hit rates of the tested radii, and fitted them with a sigmoid regression, where the personalized levels were sampled from the targeted hit rates. Details can be found in Section S.I in the supplementary material.

3.5 Experimental Protocol

Each subject participated in one offline and two online recording sessions that took place on different days within three months. The offline session aimed at collecting necessary data to build a subject-specific decoder for the online sessions. The online sessions aimed at both evaluating the decoding accuracy and comparing the behavioral outcome

Figure 2: Offline session is composed of a v-shape design of difficulty levels during the recording. There are in total 16 objective difficulty levels, and a subjective difficulty level was reported after each trajectory as a number (the line) and as a descriptive label (the dots).

between EEG decoder-based and Manual (self-paced decisions) interactions.

For all sessions, waypoints of a trajectory were arranged in a way that the subject needed to either perform a pitch or roll every other two waypoints; the waypoints in between were placed 32 A.U. (at least 2.7 seconds) in front of the previous one for letting the subject to adjust the orientation of the drone. The numbers of required pitch and roll maneuvers were balanced for all the directions. The Euclidean distance between waypoints was 32 and 24 A.U. away for pitch and roll, and at least needed 2.7 and 2.0 seconds, respectively.

3.5.1 Offline Session

During the setup, each subject familiarized with the protocol by at least 10 minutes of piloting with a pre-defined training trajectory of 121 waypoints (c.f. Section S.I), each piloting lasted about five minutes. The result of last piloting would be used to personalize difficulty levels (as presented in Section 3.4), and the subjects were informed.

Before the piloting task, we recorded one-minute of EEG signals with eye closed and opened as an EEG baseline for calibration (to be addressed in Section 4.1). Then, each subject piloted through thirty-two trajectories, each having a constant difficulty level. As depicted in Figure 2, the level decreased from level 16 to level 1, and then increased from 1 to 16. Each trajectory was composed of thirty-two waypoints and lasted around 90 seconds.

After each trajectory, the subjects reported a numeric level between 0 and 100 for the assessment of perceived difficulty level. The mean values and standard deviations across subjects were depicted in the right side of Figure 2. They also declared whether the trajectory was easy (green dots), hard (blue dots), or extremely hard (red dots). A larger dot indicates more subjects declared for that case. The definition of easy was that the subject felt in good control of the drone, while the opposite falls in one of the other two. The extremely hard was differentiated as a level that the subject feels herself cannot manage the level in reasonable training time.

3.5.2 Online Sessions

During the setup, the subject firstly practiced for at least 10 minutes, and then, piloted the drone once more for personalizing the difficulty without recording signals. Eye-open and eye-close were then recorded for a preliminary calibration of the EEG decoder (see Section 4.1).

Figure 3: Online session. (a) The flow of online sessions, where the personalize and skill evaluation refer to the same task as in Figure S1, with a difference that, skill evaluation itself, does not adjust difficulty levels. Each block is composed of 12 trajectories. The curve below is an example of plotting the difficulty levels during all trajectories against the decision points. The vertical lines indicate the boundaries between trajectories where are the only occasions of decreasing difficulty levels. (b) Example of one trajectory to demonstrate the online interaction in the EEG condition. The blue asterisks are associated to the right y -axis and indicate the events of the increasing level (harder), the subject wanted to increase the level (wants harder), or hit and miss of the decision waypoint. The vertical lines help align the timing of those events, and the lines are extended to the bottom panel in the event of decision waypoints. As shown in the bottom panel, the level is increased if the ratio of Hard class is lower than 0.5, and the ratio is reset at the beginning of a trajectory or at the event of decision waypoints.

Figure 3(a) shows the flow of online sessions. Each session was composed of two personalization processes, one skill evaluation, one EEG block, and one Manual block. The first personalization was done without recording signals as described in the previous paragraph and the skill evaluation is the same task without further tuning the levels. Since skill evaluation is a subset of personalization, please notice that in the later texts, it may also refer to the skill evaluation parts in the personalization processes.

If a subject begins with the EEG condition in the first online session (see Figure 3(a)), The Manual condition will

become the first block in the second online session. We divided the subjects into two groups of similar task scores in the first session. The first group had the EEG condition being the first block in the second session and the other group began with Manual condition. The subjects knew the condition before starting each block.

In the EEG condition, the level was increased if the EEG decoder decides as Easy (see Section 4.3 for details of interaction and Figure 3(b) for an illustration). On the other hand, the Manual condition was based on whether the subject pressed the button before the decision point.

As both cases cannot decrease the level in the case of false positive (hard), we unbiasedly lowered the levels when starting a new trajectory. This avoids a dominate class in ground truth for both conditions. Specifically, each condition was conducted in a block of twelve trajectories (see Figure 3(a)). The beginning level of the 1st trajectory was always level 1 while the other trajectories began four levels (as there were eight decisions) lower than the final level of their previous one. For example, the 2nd trajectory began at level 3 if the final level of the 1st trajectory was 7. The curve is an example of concatenating all decision points in the same block. One can see that the level is either increasing or staying at the same level, except for changing trajectories as indicated by the vertical lines.

Each trajectory consisted of 33 waypoints with eight decision points (4th, 8th, ..., and 32nd waypoints). After the decision point, the level either increased or stayed the same. In the case of EEG condition, the operator examined the behavior of the decoder before the 1st trajectory. This also happened before the 5th and 9th trajectories if the level was not moving between Easy and Hard. During the examination adjustment of the bias term of regression (see Section 4) was based on the feedback of the subject, where the subject additionally piloted one time or a few times of an extra and long trajectory. The adjustment did not require the subject to finish the trajectory and the time was made as short as possible.

4 DECODING PERCEIVED DIFFICULTY

4.1 Signal Pre-processing

EEG and EOG signals were downsampled to 256 Hz and casually band-passed between 1 and 40 Hz by a 14th order Butterworth filter. The vertical EOG component was computed by subtracting the sensor between eyebrows by the average of the other two. The horizontal EOG component was derived from the bipolar signal between the two sensors close to canthi.

Out of 39 recordings, P2 was removed twice from the offline or online sessions due to short-circuit with the CMS or DRL electrode. A 20th order spatial low-pass filter, SPHARA [22], was applied to interpolate signals for the removed electrodes and more importantly to reduce high spatial frequency components, likely corresponding to artifacts [22]. After SPHARA, 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 re-referencing was then applied [23].

Potential EEG artifacts due to eye movements were alleviated by Independent Component Analysis (ICA). We used an iterative method to estimate the assumed number of selected independent components. Specifically, we performed RUNICA [24], starting with 15 components, and iteratively searching for the maximum number of components with a proper ICA solution. A proper solution is that the returned weight matrix has no imaginary number and that the maximum and minimum values in the weight matrix are similar, for which we picked 5 as a threshold. If any of the final independent components had a correlation coefficient higher

than 0.7 with the vertical or horizontal EOG component, the independent component was then dropped out from the future analysis. On average, 15.8 components were returned by ICA and 1.07 components were then removed during the construction of online decoders.

We took $10\log_{10}$ of Power Spectral Density (PSD) as features, using Thomson's multitaper algorithm with the time-half bandwidth product being 2. The features were extracted over a two-second sliding window, with a 500ms shift for evaluating the decoder in offline and 125ms for online decoding. This resulted in a 0.5 Hz resolution, and only the power bands between 2 to 28 Hz were extracted, in order to avoid the movement-related slow cortical potentials [25] and too low signal-to-noise ratio. We rejected windows where any EEG time sample had a peak value larger than $50\mu\text{V}$ after the previous pre-processing. The features were then subtracted by a feature vector from the baseline recording (eye-open and eye-close). The vector was computed by averaging the features across time. In online sessions, we calibrated the features by subtracting the newly recorded baseline.

4.2 Classification Method

Given that the perception of difficulty level is more naturally represented by a continuous value, we first regressed from the log-PSD values to the reported difficulty level (0 to 100, but normalized to 0 and 1) with generalized linear models [26] and elastic net regularization. The chosen distribution and link model were binomial and logistic, respectively, and a grid search was performed for the regularization parameters $\alpha = [0.15, 0.5, 1]$ and $\lambda = [0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20]$.

The regression model includes a bias term that was shifted in online decoding when the decoder was not operating in a normal range. Since the perceived difficulty was less likely to change instantaneously, we smoothed the output of the regression by a moving average filter. In particular, we averaged the output within the latest 6 seconds to compromise between accuracy and latency. We then classified the smoothed output as Easy (Increase the Level) or Hard (Keep the Level), where the Hard class from now on refers to the collection of both reported Hard and Extremely hard labels unless specified. The classification was performed at each time window using a threshold learned by a Linear Discriminant Analysis (LDA). The LDA was implemented by assuming equal distributions from both classes.

The regression and LDA were trained based on the data of offline session for the online sessions. The regularization parameters of regression were those yielded the best performance during the cross-validation (to be detailed in Section 4.4).

4.3 Online Interaction

Figure 3(b) depicts the principle of the EEG based online interaction in a trajectory, where the blue asterisks are linked to the y -axis on the right and indicate the event of decision waypoints, decisions made (harder), and ground truth given by the subjects (wants harder). The decoder was giving a decision in an 8 Hz rate during online sessions (see the panel with y -axis labeled as Estimated difficulty level). However,

the level can only be increased at a decision waypoint (those vertical lines extended to the bottom panel). Whether to increase depends on the amount of each class within the four waypoints (see the bottom panel with y -axis labeled as Ratio of Hard Class). If the decoder output was dominated by the Easy class, the level increased (harder). Otherwise, the level was kept. The counter of each class was reset to zero when a decision point was met or at the beginning of a trajectory.

4.4 Performance Validation

Both the offline and online decoding performances were assessed by class-balanced accuracy. The α and λ hyperparameters of the regression model for each subject in their online sessions were decided by the best decoding performance in an offline validation. The best performance was decided by maximizing $0.5 \times \text{mean}$ (across test sets) of class-balanced accuracy - $0.5 \times \text{mean}$ (across classes) of standard deviations (across test sets).

The offline validation strategy ideally is four times of leave-one-pair-out cross-validation, where each test set holds a pair of trajectories; one trajectory labeled as Easy while the other as Hard or Extremely Hard, and this allows the assessment of the class-balanced accuracy. However, it is only equivalent to four times of leave-one-pair-out cross-validation if the total amount of Easy trajectories is equal to the total amount of Hard plus Extremely Hard trajectories. Otherwise, it is sixty-four (4×16 pairs) times of leave-one-pair-out validation which ensured the highest priority on the least picked trajectories for testing. In this case, some Easy trajectories were chosen less or more times than the Hard plus Extremely Hard trajectories.

For the EEG condition of online sessions, which is referred as closed-loop, two class-balanced accuracies were provided, one evaluated on the entire block and the other is divided into three groups of four trajectories, where the four trajectories of each group always share the same bias term in regression. When evaluating on the entire block, the accuracy was computed based on 96 samples (8 decision points per trajectory $\times$ 12 trajectories), and 32 samples with a group of four trajectories. The ground truth of a decision point was obtained from the button pressing. Once the button is pressed, the current decision point is considered as Easy, and as Hard if not pressed. If a button is pressed but the level stays the same after a decision point, the ground truth of the next decision point was also considered as Easy.

On the other hand, we also evaluated the open-loop cases in the online sessions, where no feedback was provided to the subject. Specifically, the open-loop cases refer to the Manual condition, the skill evaluation part right before the second block, and the skill evaluation in the flow. The case before the first block was not evaluated because it was done during the setup without recording the physiological signals. The purpose was to evaluate the best decoding performance that could be achieved by shifting the bias term in the regression. The evaluation was done by a post-hoc simulation that used the same decoding procedure as in the closed-loop case. We scanned different shifts (add or subtract a value) from the bias term, where the shift ranged from -0.5 to 0.5 with 0.005 as a step.

5 ANALYSIS OF ONLINE BEHAVIORAL DATA

5.1 Task Scores

One question of interest is whether subjects could achieve similar task performance in both conditions. As a result, two-tailed paired t-tests over task scores were conducted between both conditions in each session for a group level analysis ($n=13$ subjects). We further applied a two-tailed paired t-test ($n=12$ trajectories) for each subject in each session. We then applied a Bonferroni-Holm correction to reduce the false positive rate [27]. A perfect decoder should yield similar task scores as the Manual condition. Although the task scores in the personalization or skill evaluation can be measured, they were not compared. The comparison otherwise would not be fair as the functional task difficulty levels would not be the same, because the used eleven levels were the same in all the online sessions; the personalized levels were the sixteen levels used in the twelve trajectories of a condition.

5.2 Skill Curves

Apart from the task scores, skill improvement is also worthy of examination. The sigmoid regression between hit rates and radii (details in Section 5.1) can represent a subject's skill curve. The skill curves are characterized by several parameters, the two important ones are x_{50} (the x -value has 50% of the range in y response) and slope. Generally speaking, a larger (smaller) x_{50} (slope) represents a worse skill. The comparison was conducted by a two-tailed paired t-test across subjects ($n = 13$) either over conditions or between the first and second blocks (to see time effect). The correlation between decoding accuracy and the two parameters were also checked.

The data being used for the skill curves are the hit rates of all levels in each block, where the x -axis represents the 16 levels instead of radii for each block. This is based on the assumption that the same level is mapped to the same point on the functional task difficulty level. The regression was bounded between 0 and 1 for the hit rate (y -axis), but the lower bound was relaxed if there is a convergence issue.

In addition to the x_{50} and slope, another index is the area beneath the skill curve. The area was computed by an integral over level 1 to 16 which leads to overall hit rates across all the levels. The larger the area is the better the skill. With the overall hit rates, their differences between both conditions were computed for each subject. The differences were further t-tested ($n = 13$) against 0. A significant result indicates that the overall hit rates between both conditions are different. The correlation with decoding accuracy was also computed.

5.3 Final Levels

Assuming the subject's skill didn't change drastically and the personalization process worked perfectly, the curve of final levels v.s. trajectories should also be similar between the EEG and Manual conditions, especially, the radii were re-mapped to the same set of functional difficulty levels. In order to compare dissimilarity of the curves, three different indices were defined:

Figure 4: Online decoding accuracy in EEG condition at decision-point level, in terms of each class and class-balanced accuracy. A upward bar indicates the accuracy is higher than the theoretical chance level, 0.5, and *vice versa*.

- Pearson’s correlation: this gives the similarity of a co-varying trend between two patterns (vectors, $n = 12$), and dissimilarity can be computed as $(1 - r)/2$, where r is the correlation coefficient. Although correlation is a good indicator of similarity, it does not tolerate overshooting but tolerates any magnitude of shifting. We consider $r > 0.5$ with significance as two patterns are highly similar.
- Mean of difference (MD): it indicates the difference of averaged level between both conditions. It tolerates symmetrical overshoots and undershoots without considering the amount of shooting. We consider MD lower than 1.5 level as two patterns are highly similar.
- Mean of absolute difference (MAD): a typical index to compare two curves, it does not tolerate any kind of overshoots or undershoots with a high magnitude. We consider MAD lower than 1.5 level² as two patterns are highly similar.

Each index has its pros and cons. Therefore, having at least two indices showing high similarity are convincing to conclude that both curves are similar.

6 RESULTS

6.1 Offline Behavioral Result

The plot of subjective difficulty level in Figure 2 summarizes the reported numerical and descriptive difficulty levels. The curve is the average of all subjects with the shaded area represents the standard deviation. The green, blue, and red dots represent the total amount of Easy, Hard, and Ex. Hard labels, respectively. A larger dot stands for more subjects labeling that class. It can be seen that the designed protocol nicely induced subjective difficulty level in the intended v-shape. For subject-wise result, please refer to Figure S2 in the supplementary material.

6.2 Offline Accuracy Validation

The mean class-balanced accuracy in offline validation across subjects was 76.7% with the standard deviation being 5.1%. These values indicate plausible decoders are available for the online sessions. For detailed result, please refer to the Figure S4 in the supplementary material.

6.3 Online – Decoding Accuracy

6.3.1 Closed-loop

Figure 4 illustrates the online decoding accuracy of the EEG condition for each session and subject. Blue bars show the class-balanced accuracy while green and red bars are class-accuracy for Easy and Hard, respectively. An upward (downward) bar means the accuracy is higher (lower) than the theoretical chance level of a binary problem (50%). The ‘x’ markers are the ratio of Easy samples in the ground truth. If the ratio is 0 or 1, the class-balanced accuracy corresponds to accuracy.

Each sub-figure (session) shows the per-subject result computed with 96 samples (8 decision points $\times$ 12 trajectories) of the entire EEG block. For both sessions, in total 16 out of 26 (62%) recordings have a class-balanced accuracy higher than the theoretically chance level. The mean class-balanced accuracy in Session 2 (3) across subjects was 56.2% (54.7%) with the standard deviation being 8.6% (11.0%). For accuracies evaluated on four-trajectory groups and the used shifts of bias term during the online decoding, please refer to Section S.II.4 in the supplementary material. The online decoding accuracies were lower than in the offline validation. This can be a typical case for EEG decoders.

6.3.2 Open-loop

Figure 5 plots the highest class-balanced accuracy with the best bias term in the Manual condition (see Figure 5(a)) and the skill evaluations (see Figure 5(b)). In Figure 5(b), each skill evaluation is drawn in a specific color. The upper

Figure 5: Open-loop class-balanced accuracy at decision-point level in the Manual condition and the last two skill evaluations.

panels of Figure 5 report the best accuracy per subject. Similar to Figure 4, an upward bar of accuracy is higher than the chance level, and the one without a bar (s4 in session 2) performed as a chance-level decoder. The lower panels of Figure 5 show the corresponding best shift of bias term. A longer bar regardless of the pointing direction indicates that the decoder is less stable across the offline and the online sessions, while the difference between the two bars in Session 2 and Session 3 indicate the stableness across the online sessions.

Nearly half of the subject had over 60% of accuracy in the Manual condition, and the ranking of accuracies was consistent across sessions (Spearman’s correlation, $r = 0.92$, p -value = 0, $n = 13$). The accuracy was therefore highly dependent on the subject. From the panels of shifts, the stableness across the offline and the online sessions was clearly low. Between the online sessions, the average difference of the best shifts was 0.05 with a standard deviation of 0.04, where a shift of 0 between sessions means that the decoder does not need to be tuned again. Given that an averaged inter-session difference is 0.05, this indicates that the best bias term can be relatively stable across online sessions.

The skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, p -value = 0.12, $n = 13$ for the 2nd evaluation and $r = 0.09$, p -value = 0.71 for the 3rd one). The average differences of the best shift between session were 0.12 and 0.01 for the 2nd and 3rd evaluations, where the standard deviations were 0.27 and 0.06. The within session differences of the best shifts were higher in the second session than the third session. The mean values were 0.12 and 0.00 for the second and third sessions, where standard deviations were 0.26 and 0.04.

The skill evaluation part has much higher accuracies than the Manual cases. The higher accuracy may be explained by four possible reasons. First, a smaller number of decision points were easier to optimize. There were 96 decision points in the Manual condition but only 10 in each skill evaluation. Second, interruptions between each trajectory in the Manual condition changed the subjects’ background states. Third, a longer time between decision points made the estimation more stable and reliable; Skill evaluation has one decision point after every eleven waypoints while in the

Manual condition, the spacing was four waypoints. Fourth, the skill evaluation went through all the 11 levels. The more extreme the levels are, the easier to decode (see section S.II.3 in the supplementary material).

6.4 Online – Task Scores

Figure 6 shows the comparison of task scores between the Manual and EEG conditions as boxplots for each online session. Each subject has a red box representing the Manual condition and a blue one for the EEG condition. The horizontal line inside a box is the median and the cross (+) represents the mean value. The colored circle dots are outliers.

Group level analysis using a two-tailed paired t-test ($n=13$) on each session shows there is a large effect of the conditions on the scores,¹ where Manual is significantly better. Session 2 yielded a p -value of 0.01 and an effect size of 0.8. Session 3 yielded a p -value of 0.005 and an effect size of 0.95.

Two-tailed paired t-tests ($n=12$) with Bonferroni-Holm correction ($m=13$) on each session were also conducted between both conditions as shown in Figure 6 (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). Although the scores for Manual condition are generally higher than the EEG condition and have a smaller variation, they are not necessarily significantly better. A few cases show that the average score (+ sign) for EEG condition was higher than the Manual one. Specifically, s3 and s10 in session 2 as well as s2 and s5 in session 3 were higher. This implies that using an EEG decoder still has the potential of having a better score.

6.5 Online – Skill Curves

Following Section 5.2, Table 1 lists the results of t-tests for x50 and slope. The condition columns mean that the test was conducted between the Manual and EEG conditions, while the Time columns refer to the result between the first and second presented blocks. Row p is the p -value and Power is the t-test statistical power, where a positive value means that Manual (1st block) is higher than the EEG condition

1. Equivalent to ANOVA since there are only two variables.

Figure 6: Task scores and t-tests ($n=12$) with Bonferroni-Holm correction ($m=13$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

Table 1: T-tests ($n=13$) on the parameters of skill curves between the conditions and the order of presentation.

		x50		slope	
		condition	Time	condition	Time
Session 2	p	0.89	0.42	0.36	0.35
	power	-0.15	-0.84	0.95	-0.97
Session 3	p	0.34	0.66	0.82	0.17
	power	-1.00	0.45	-0.24	1.47

Table 2: Correlations ($n=13$) between the parameters of skill curves and the decoding accuracy in the EEG condition.

	Session 2		Session 3	
	x50	slope	x50	slope
r	-0.56	0.11	0.27	0.48
p	0.05	0.72	0.38	0.10

Table 3: Differences of overall hit rates and the correlation ($n=13$) between the differences and decoding accuracies.

Session	Manual - EEG				1st block - 2nd block			
	Hit rate difference		Correlation		Hit rate difference		Correlation	
	Avg	Std	r	p	Avg	Std	r	p
2	1.56	2.73	-0.26	0.391	-1.69	2.65	0.21	0.501
3	0.40	2.17	0.46	0.111	0.39	2.17	0.07	0.832

(2nd block) in the condition (Time) column. No significant result was found suggesting there is no strong effect on the condition or time. That means both conditions are similar in terms of the skill curve and the improvement in each block is not biased by the time being executed.

Table 2 shows the correlation between the parameters and the decoding accuracy for the EEG condition. Only the second session has x50 as negatively correlated, which is encouraging as a higher accuracy leads to a lower (better) x50.

On the other hand, Table 3 lists the result of overall hit rates in different sessions for comparing the two conditions and the effect across time. Positive values in the Avg columns indicate that Manual (1st block) is better than the EEG condition (2nd block). No significant correlation was found suggesting that the decoding accuracy did not correlate to either of the checked factors.

6.6 Online – Final Levels

Figure 7 reports the final level of each trajectory for all subjects within each session as histograms. The blue bars stand for the EEG condition with the red ones for the Manual. Their mean values are very close and the majority of the

Figure 7: Distributions of the final levels.

final levels located between level 6 to level 9 regardless of the condition. EEG condition, as expected to be less stable than the Manual condition, shows a higher standard deviation and higher counts on the tails of the distributions.

Figure 8 illustrates the curves representing the final level of each trajectory for three exemplar subjects in both sessions. The red curves stand for the Manual condition while the blue ones are for the EEG condition. Based on all subjects (see Figure S8), the patterns of Manual condition generally show trends that the subjects preferred to stay around a certain level, usually around 8. This behavior is consistent with the behavioral condition of a previous EEG study that adapts the levels of Tetris [15]; as reported by the authors, the subjects preferred to stay around a certain level and as soon as possible. However, after a few more trajectories, our subjects increased a bit the level, probably because subjects felt more confident or more familiar with the control. On the side for the EEG condition, around half of the recordings had similar patterns as the Manual cases.

Figure 8: Final levels across trajectories of three subjects.

The adjustments of the bias term probably favored similar patterns, but the adjustments were made at most twice and some subjects did not need. In total 9 out of 26 sessions did not require an adjustment after starting, and s10 kept the bias term from the offline session (see Figure S7).

The correlation, MD (mean difference), and MAD (mean of absolute difference) are also indicated for each subject as texts, where boldfaced values stands for high similarity (see Section 5.3). In total, five and six subjects in session 2 and session 3, respectively, have two indices indicating high similarities.

6.7 Correlation with Decoding Accuracy

A question of interest is the role of decoding accuracy in the interaction, given that the behavioral results are similar between the two conditions even if the online decoding accuracy was not very high. In addition to the correlation tests conducted in Table 2 and 3, other additional correlation tests were performed. Between the task scores and the online accuracy, the correlations were 0.15 ($p = 0.62$, $n=13$) in session 2 and 0.13 ($p = 0.67$, $n=13$) in session 3. For the indices of dissimilarity, the correlations with r were -0.28 ($p = 0.34$, $n=13$) in session 2 and 0.01 ($p = 0.98$, $n=13$) in session 3. The correlations to MAD were -0.01 ($p = 0.98$, $n=13$) in session 2 and -0.67 ($p = 0.01$, $n=13$) in session 3. The significant negative coefficient indicates that the higher the accuracy is, the lower the dissimilarity. The correlations with MD were 0.46 ($p = 0.11$, $n=13$) in session 2 and 0.27 ($p = 0.37$, $n=13$) in session 3. Although there was one significant result, it is probably by chance.

7 DISCUSSION

We have conducted a comprehensive analysis, from offline validation to online accuracy and behaviors. The designed protocol was able to elicit different levels of difficulty in both numerical and descriptive labels. Even if the difficulty levels were personalized with the same standard, the distributions across subjects are not necessarily the same. (see Figure S2 in the supplementary material). This indicates that either (1) hit rates did not consistently reflect the subjective feeling, (2) some subjects largely improved their skills, but not others, during the 32 trajectories, or (3) the estimated hit rates were not robust enough as based only on 11 data points. It was

not easy to conclude the main factors, but in any case, the data was sufficient to build subject-specific decoders and perform offline validation.

The offline validation yielded promising decoding accuracy. Although the accuracy for some subjects was biased towards one class (see Figure S4 in the supplementary material), the tuning of the bias in regression term during the online sessions seems to help alleviating such situation. Indeed, as long as the best shift of the bias term can be found, the decoding accuracy could still be high.

Many behavioral results in online sessions suggest that the EEG and Manual conditions were similar, even though task scores of Manual condition were significantly better than the EEG condition at the group level. For a subject, the task scores of both conditions could be similar; about two-third statistical tests between conditions did not yield a significant difference in the task scores. The parameters of skill curves also nearly showed no statistical difference. Besides, about half of the cases, the curves of final levels were very similar between both conditions, and nearly every subject had one session with a close pattern. These results suggest that the EEG decoder was able to yield similar results as the self-paced condition. Similar to the study of learning a new arithmetic system [20], the authors also reported that EEG condition yielded similar behavioral results compared to a condition based on user's behavior, namely the error made. Therefore, the feasibility of using EEG decoder in the interaction loop holds promise.

Another similar, but not comparable, study was made by Faller *et al.* which was also using a simulated drone and have two conditions [10]. There are several differences. In their study, the adaptive process did not change the piloting task, while the difficulty level of our piloting task was changed. They focused on whether down-regulating the arousal state is useful in a high-demand and non-learning task, while we aimed at whether users' cognitive states can replace self-paced decisions in a learning-like setting. Faller *et al.* instructed the subjects to actively down-regulate the arousal state while piloting, and their non-EEG condition utilized sham decoders that were irrelevant to the ground truth. On the contrary, the subjects in our work were not instructed to pay attention to the decoder, and our non-EEG condition was ground truth. In both of our conditions, the subjects should be consistent with the definition of the descriptive label when pressing the button. That means, pressing the button in the online sessions was expected to be equivalent to labeling Easy in the offline session.

An interesting finding in our study is that in rare cases the EEG condition performed better than the Manual one in task scores. One case was s10 in session 2 which had high online accuracy. However, s11 in session 3 also had a high online accuracy but did not surpass the ceiling. One possible explanation is that s10 had lower accuracy in the Hard class than s11. In other words, the decoder tended to make the level harder, forcing the subject to challenge himself by leaving his comfort zone, or s10 was rather conservative on when to increase the level in the Manual condition.

Although the offline validation accuracy was high, it did not necessarily guarantee high online decoding accuracy tested on different days. This cross-day variation might come from fatigue, different background states, different

setups, etc. Some approaches based on robust principal component analysis have been proposed to alleviate this cross-day variations [28], [29]. Nonetheless, their performance in these tasks is yet to be assessed. Moreover, these approaches have high computational requirements that may hinder their use in online settings. We, therefore, chose the baseline subtraction approach, because it can be combined with the adjustment of the bias term to save computational power and worked well.

The role of decoding accuracy in the interaction is unclear. Ideally, a perfect decoder should give the same or similar result as the Manual condition, and a decoder biased toward one class would lead to a result dominated by level 1 or 16. The correlation to the accuracy in Table 2 and 3, however, only yielded one significant result which supports a higher decoding accuracy leads to a better skill curve. Other correlations tests in Section 6.7 also yielded limited evidence to show that a higher accuracy leads to a better behavioral outcome, scores, skill curve, and high similarity to the Manual condition. As a result, it is hard to draw any conclusion between the decoding accuracy and the behavioral results. One issue was that the average online accuracy was about 20% lower than the offline validation. This means that the online decoding accuracies were generally not much higher than the chance level, and one standard deviation away from the mean was not even higher than 66%. The rather close-to-chance-level accuracy might hinder showing significant results with a small number of samples. Apart from the relatively low accuracy, this indecisive conclusion might also be largely influenced by the personalization of difficulty without many samples. A small number of samples may not give a reliable estimate, but a rather large sample size would take too much time to collect and, more importantly, improved too much the skills of subjects such that one cannot easily observe any learning effect. We believe that building a robust decoder in the closed-loop condition has a higher priority for better answering the role of decoding accuracy in the future experiment. One working direction is to solve better the cross-session variations. For example, apart from the baseline calibration, we can perform the scan of the best bias term before the actual condition at the cost of much more preparation time and effort that may be less useful for daily usage and lead to lower generalizability. At least, it would be useful to better answer the role of accuracy.

It is worth noting that neural correlates of the subjective difficulty level reported in this paper (see Section 5.2) are similar to those we found in a previous study where subjects piloted the same drone, but with their left hand (instead of right hand as here) and the difficulty level was modulated by the degree of multi-tasking [30]. Based on common features, the neural correlate of the subjective difficulty level seems to lie in the α band in the right centroparietal hemisphere (C4 and CP4 in particular) and the θ band around Cz.

Our protocol is closer to a realistic task as compared to Faller *et al.* [10] and we evaluated the decoder with a much shorter latency than Walter *et al.* [20]. We anticipate that our method may apply to scenarios such as education and gaming. Possible benefits are a shorter learning time and dynamically adapting difficulty levels for action games that

are normally fast pace.

Given the competitive behavioral outcomes and accuracy for some subjects, we believe that a certain population may directly benefit from a BMI regulator. Furthermore, as in other BMI applications, in particular related to gaming [31], users may learn to generate better neural correlates of their subjective difficulty level with practice once they experience the benefit of the BMI regulator.

8 CONCLUSION

The theory of challenge point suggests that leaving the easy zone benefits learning during practice. Following this, the designed online interaction demanded the subjects to leave their comfort (Easy) zone. The closed-loop experiment demonstrated that the behavior results could be similar between using an EEG decoder and self-paced decisions. Subjects might even perform better with an EEG decoder if the decoder is accurate enough and slightly biased toward increasing the difficulty level. However, more positive samples are still needed to draw a decisive conclusion. One potential reason for having lower accuracy might be the duration of a certain level. In the offline validation, each trajectory lasted for around 90 seconds while a level in the online session could be as short as twelve seconds. Apart from improving the accuracy with a scan for the best bias term, another future research direction is examining the applicability of the decoding framework in other tasks with online evaluations, such as in-class education and other kinds of visuomotor responses, e.g. using another hand or feet to control [30]. This will probe whether the decoded cognitive state is task-specific and robust in real interactions.

ACKNOWLEDGMENTS

We would like to thank Alexander Cherpillod for his help in the implementation of the simulator and Ruslan Aydarkhanov for his suggestions in designing the protocol. Some figures were drawn with the Gramm MATLAB toolbox [32].

REFERENCES

- [1] J. D. Finn and K. S. Zimmer, "Student engagement: What is it? Why does it matter?" in *Handbook of Research on Student Engagement*. Springer, 2012, pp. 97–131.
- [2] P. D. Umbach and M. R. Wawrzynski, "Faculty do matter: The role of college faculty in student learning and engagement," *Research in Higher Education*, vol. 46, no. 2, pp. 153–184, 2005.
- [3] R. A. Green, L. Y. Whitburn, A. Zacharias, G. Byrne, and D. L. Hughes, "The relationship between student engagement with online content and achievement in a blended learning anatomy course," *Anatomical Sciences Education*, vol. 11, no. 5, pp. 471–477, 2018.
- [4] 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.
- [5] C. D. Wickens, S. Hutchins, T. Carolan, and J. Cumming, "Effectiveness of part-task training and increasing-difficulty training strategies: a meta-analysis approach," *Human Factors*, vol. 55, no. 2, pp. 461–470, 2013.
- [6] P.-K. Jao, R. Chavarriaga, and J.d.R. Millán, "Analysis of EEG correlates of perceived difficulty in dynamically changing flying tasks," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2018.

- [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] 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.
- [9] P. Jorna, "Heart rate and workload variations in actual and simulated flight," *Ergonomics*, vol. 36, no. 9, pp. 1043–1054, 1993.
- [10] 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.
- [11] 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.
- [12] J. Klingner, R. Kumar, and P. Hanrahan, "Measuring the task-evoked pupillary response with a remote eye tracker," in *Symposium on Eye Tracking Research & Applications*, 2008, pp. 69–72.
- [13] 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.
- [14] 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.
- [15] 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, no. 223, 2016.
- [16] 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 *Proc. 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2015, pp. 7242–7245.
- [17] D. Szafir and B. Mutlu, "Pay attention!: Designing adaptive agents that monitor and improve user engagement," in *SIGCHI Conf. on Human Factors in Computing Systems*, 2012, pp. 11–20.
- [18] 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 *2016 CHI Conf on Human Factors in Computing Systems*, 2016, pp. 5372–5384.
- [19] 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. 1-2, pp. 187–195, 1995.
- [20] C. Walter, W. Rosenstiel, M. Bogdan, P. Gerjets, and M. Spüler, "Online EEG-based workload adaptation of an arithmetic learning environment," *Frontiers in Human Neuroscience*, vol. 11, no. 286, 2017.
- [21] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Journal of Comparative Neurology and Psychology*, vol. 18, no. 5, pp. 459–482, 1908.
- [22] 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.
- [23] 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.
- [24] 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
- [25] G. Garipelli, R. Chavarriaga, and J. d. R. Millán, "Single trial analysis of slow cortical potentials: A study on anticipation related potentials," *Journal of Neural Engineering*, vol. 10, no. 3, p. 036014, 2013.
- [26] A. J. Dobson and A. Barnett, *An Introduction to Generalized Linear Models*. CRC press, 2008.
- [27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, pp. 65–70, 1979.
- [28] Y.-P. Lin, P.-K. Jao, and Y.-H. Yang, "Improving cross-day EEG-based emotion classification using robust principal component analysis," *Frontiers in Computational Neuroscience*, vol. 11, no. 64, 2017.
- [29] P.-K. Jao, R. Chavarriaga, and J. d. R. Millán, "Using robust principal component analysis to reduce EEG intra-trial variability," in *40th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2018, pp. 1956–1959.
- [30] P.-K. Jao, R. Chavarriaga, F. DellAgnola, A. Arza, D. Atienza, and J. d. R. Millán, "EEG correlates of difficulty levels in dynamical transitions of simulated flying and mapping tasks," *IEEE Transactions on Human-Machine Systems*, 2020.
- [31] S. Perdakis, L. Tonin, S. Saeedi, C. Schneider, and J. d. R. Millán, "The Cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users," *PLoS Biology*, vol. 16, no. 5, p. e2003787, 2018.
- [32] P. Morel, "Gramm: Grammar of graphics plotting in Matlab." *J. Open Source Software*, vol. 3, no. 23, pp. 568–571, 2018.

Ping-Keng Jao received his PhD in electrical engineering from the Swiss Federal Institute of Technology in Lausanne (EPFL) in March 2020. For the four years at EPFL, he was studying decoding algorithms and closed-loop applications for Brain-Machine Interface with able-bodied people. He received an Outstanding Reviewer Award in 2017 for the Journal of Neural Engineering, IOP Publishing. From 2013 to July 2015, he was as a research assistant in Academia Sinica, Taiwan. There, he mainly worked on music technology based on sparse representation for musical genre classification and sound source separation. He also conducted small projects of Brain-Machine Interface for the attention to musical source and emotion recognition.

Ricardo Chavarriaga has more than 12 years of experience in research on human-machine interaction, brain-machine interfaces, and artificial intelligence. His work is focused on responsible development of technologies that promote beneficial, humane interaction between human and intelligent machines.

Ricardo is the head of the Swiss office of the CLAIRE (Confederation of Laboratories for AI Research in Europe); Senior researcher at Zrich University of Applied Sciences (ZHAW) and associate fellow at the Geneva Center for Security Policy (GCSP). Dr. Chavarriaga is also chair of the IEEE Standards Association industry connection group on Neurotechnologies for Brain-Machine Interfacing.

José del R. Millán is a Professor in the Dept. of Electrical & Computer Engineering and in the Dept. of Neurology of the University of Texas at Austin since September 2019, where he holds the Carol Cockrell Curran Endowed Chair. He received a PhD in computer science from the Technical University of Catalonia, Barcelona, in 1992. Previously, he was a research scientist at the Joint Research Centre of the European Commission in Ispra (Italy) and a senior researcher at the Idiap Research Institute in Martigny (Switzerland). Most recently, he was the Defitech Foundation Chair in Brain-Machine Interface at the École Polytechnique Fédérale de Lausanne in Switzerland (EPFL) from 2009 to 2019, where he helped establish the Center for Neuroprosthetics. Dr. Millán has made several seminal contributions to the field of brain-machine interfaces (BMI). He has received several recognitions for these seminal and pioneering achievements, notably the IEEE-SMC Norbert Wiener Award in 2011, elevation to IEEE Fellow in 2017, elected Fellow of the International Academy of Medical and Biological Engineering in 2020.