EPFL

Thèse n° 7696

Decoding Cognitive States under Varying Difficulty Levels

Présentée le 27 mars 2020

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de traitement d'images médicales
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Ping-Keng JAO

Acceptée sur proposition du jury

Dr J.-M. Vesin, président du jury
Prof. D. N. A. Van De Ville, Prof. J. D. R. Millán-Ruiz, directeurs de thèse
Prof. D. Bavelier, rapporteuse
Prof. F. Babiloni, rapporteur
Prof. S. Micera, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

Only those who dare to fail greatly, can ever achieve greatly
— Robert F. Kennedy, Day of Affirmation Address (1966).

In memory of my grandparents and to my parents. . .

Acknowledgments

The journey of the Ph.D. is coming to the end. It is the best moment to recall those helping hands. Undoubtedly, the first appreciation goes to **Prof. José del R. Millán** who offered me this academic adventure. I am also grateful for the freedom he gave me to explore some side projects of my personal interest, and not to mention every awesome CNBI parties hosted by him and **Nancy**. Another thank you for **Prof. Dimitri Van de Ville** who took the last baton of this academic relay race and advised on how to finish it. Without him, I probably will not pass the finishing line.

I would never grow up so much in this journey without our fantastic and active CNBI members in the research activities. I would like to first give a special thank you to **Dr. Ricardo Chavarriaga** for our discussions and seemed-to-be-endless revisions, which always led to a much better article as if operated by the best plastic surgeon. I also enjoyed many valuable discussions with and feedbacks from **Simis**, **Luca T.**, **Marija**, **Kyuhwa**, **Iñaki**, **Sareh**, **Michael**, **Stéphanie**, **Chris**, **Ruslan**, **Bastien**, **Julien**, **Fumiaki**, **Luca R.**, **Huaijian**, **Tiffany**, and **Dong**. Additionally, I need to express my gratitude to **Simis**, **Luca T.**, and **Julien** for their prompt hacking of the mysterious cnbiloop which works most of the time but still needs some maintenance. During my studies, I received much assistance from **Frédérique** and **Manuela**. Therefore, I convey my appreciation.

Another best aspect of my Ph.D. life is enjoying the awesome nature of Switzerland. I have to thank **Simis** and **Chris** for organizing the hikes, **Stéphanie** for her ski equipment to let me enjoy the joyfulness in winter, and **Kyuhwa** for the ski weekend in Verbier and the hike to the Guggihütte which was about-100m-to-succeed and T4-but-claimed-as-easy! The experiences were really unforgettable.

Of course, I must thank my juries, **Prof. Fabio Babiloni**, **Prof. Daphné Bavelier**, **Prof. Silvestro Micera**, and **Prof. Jean-Marc Vesin**, who are willing to provide their valuable time on justifying my effort on this thesis. Besides, I have to acknowledge **Alexander Cherpillod** who developed the first prototype of the protocols in my studies, and also **Fabio Dell'Agnola** who developed the main part of the protocol in Chapter 5 and worked together for revising the protocol and collecting data for Chapter 5.

My lovely accommodation in Geneva had been mainly in Le Centre Universitaire Protestant. I have to show my gratefulness to **Christelle** and **Giuliana** for managing the building that I can comfortably dwell as my second home, and to **Liliana**, **Edilma** and **José** for providing a pleasant environment to live with. Obviously, I have to thank my countless flatmates coming from so many corners of the world. Without you, my life in Geneva would never ever be as

interesting as it is. Sorry for not listing the names, there are just too many of you.
The people who will never be forgotten is my family. Without any of you, I would not be here today. Without any of you, I would not be able to chase what I want. I have to use this occasion to deeply express my thankfulness.
Last but not the least, I believe that there are many people missing from this acknowledgment. In order not to miss anyone, let me express my appreciation to everyone and everything since the Big Bang. If missing any part since then, today might be completely different. I might not be here, or became even better, who knows? At least, I am satisfied with who I am now.
There is no endless journey. However, in this limited pathway, I felt unlimited hope because of the presence of all of you. No one knows whether we will connect again in the future; The crossings of our footprints will nevertheless be preserved in this short and easily neglected acknowledgments (and maybe also my thesis). For whoever (Bastien, are you writing your thesis now?) reads this, I wish you can also create your unique and unforgettable adventure in your life. It may be difficult, it may be disappointing, but hard time will pass, what we need is a strong will to devote. Then, we can pile toward the infinite possibility with our finite contributions and time.

Geneva, 28 January 2020 P.-K. Jao

Abstract

Understanding cognitive states of human under different difficulty levels is useful in improving human-human and human-machine interactions. For example, a crucial factor in designing games is maintaining the engagement of players. An ideal scenario would be decoding the cognitive states of players to dynamically adjust the difficulty level of the game. The same idea can also be applied in learning, where an easy level does not provide a sense of achievement while a too difficult level frustrates students. Therefore, the main idea of improving the interactions is reaching and maintaining at an optimal difficulty level.

The optimal difficulty level is, however, not simple to define and is also a function of skills which varies with time. An automatic approach is to build the interaction loop by the employment of a cognitive state decoder which helps dynamically adjusting the difficulty level. Ideally, the level will converge to the optimal level after a certain time. Even if the skills are improved afterward, the cognitive states reflecting the same objective difficulty level will be different as the person is more skilled and more confident.

Based on this idea, two online experiments were conducted in this thesis for demonstrating the feasibility. The behavioral outcomes using cognitive state decoders were compared with those based on behaviors or decisions of subjects, *i.e.*, ground-truth-like baselines. The outcomes are promising; several subjects had similar results between the two conditions. In some rare cases, decoding cognitive states even outperformed the ground-truth-like condition. These pieces of evidence support the idea of decoding cognitive state to improve the interactions. The backbone of this approach is decoding the cognitive states of interest from physiological signals. Electroencephalography (EEG) is the selected physiological signal for its non-invasiveness and quick response. Based on the defined protocols, this thesis proposed a two-stage processing method and compared it with previously highlighted engagement and attention indices as well as some state-of-the-arts classifiers. The open-loop validation supports the proposed decoding framework, which is further validated by one experiment with both open-loop and closed-loop settings, as well as another open-loop experiment involving a second task.

In order to improve the interaction, the temporal dynamics of cognitive states should be well captured. Previous literature of cognitive state computing mostly focuses on setups where the cognitive states should remain constant over a certain period. This thesis further probes the dynamics of cognitive states around the onset where the cognitive states are supposed to change. The analysis suggests that the latencies introduced by either the decoder or subjects are on a scale of seconds for the designed protocol. This is not a negligible scale if the targeted

application requires a quick reaction or a precise timing from the decoder.
In sum, the closed-loop experiments support the main idea of improving interactions through decoding cognitive states from EEG signals. The proposed decoder has been shown effective in decoding difficulty-related cognitive states, for which one should also be careful about the latencies.
Keywords: Electroencephalography (EEG), Brain-Computer Interface (BCI), Human-Machine Interaction (HMI), difficulty, workload, latency, closed-loop, dynamics.

Résumé

Comprendre les états cognitifs de l'homme à différents niveaux de difficulté est utile pour améliorer les interactions homme-homme et homme-machine. Par exemple, un facteur crucial de la conception du jeu est le maintien de l'engagement des joueurs. Un scénario idéal consisterait à décoder les états cognitifs des joueurs pour ajuster dynamiquement le niveau de difficulté du jeu. Cette même idée peut également être appliquée à l'apprentissage, où un niveau facile ne procure pas le sens de la réussite tandis qu'un niveau trop difficile frustre les élèves. Par conséquent, l'idée principale d'améliorer les interactions est d'atteindre et de maintenir un niveau de difficulté optimal.

Le niveau de difficulté optimal n'est toutefois pas simple à définir et dépend également de compétences qui varient avec le temps. Une approche automatique consiste à construire la boucle d'interaction en utilisant un décodeur d'état cognitif qui aide à dynamiquement ajuster le niveau de difficulté. Idéalement, le niveau convergera vers le niveau optimal après un certain temps. Même si les compétences sont améliorées par la suite, les états cognitifs reflétant le même niveau de difficulté objectif seront différents, car la personne est plus compétente et plus confiante.

Sur la base de cette idée, deux expériences en ligne ont été menées dans cette thèse pour démontrer la faisabilité. Les résultats comportementaux qui utilisent des décodeurs d'état cognitif ont été comparés à ceux basés sur les comportements ou les décisions des sujets, *i.e.*, une condition analogue à la vraie réponse. Les résultats sont prometteurs, plusieurs sujets ont eu des résultats similaires entre les deux conditions. Dans de rares cas, la classification d'états cognitifs était supérieur même à la condition semblable aux véritables états cognitifs. Ces évidences soutiennent l'idée de décoder l'état cognitif pour améliorer les interactions.

L'élément clef de cette approche est le décodage des états cognitifs d'intérêt à partir de signaux physiologiques. L'Electroencéphalographie (EEG) est le signal physiologique choisi pour son caractère non invasif et sa réponse rapide. Sur la base de protocoles définis, cette thèse a proposé une méthode de traitement en deux étapes et comparé aux indices d'engagement et d'attention précédemment soulignés, ainsi qu'à des classificateurs à la pointe de la technologie. La validation en boucle ouverte prend en charge le cadre de décodage proposé, qui est également validé par une expérience avec des paramètres à la fois en boucle ouverte et en boucle fermée, ainsi qu'une autre expérience en boucle ouverte impliquant une seconde tâche.

Afin d'améliorer l'interaction, la dynamique temporelle des états cognitifs doit être bien capturée. Les précédentes études computationnelles de l'état cognitif se concentrent princi-

palement sur des configurations dans lesquelles les états cognitifs devraient rester constants sur une certaine période. Cette thèse approfondit la dynamique des états cognitifs au début, où les états cognitifs sont supposés changer. L'analyse suggère que les retards introduits par le décodeur ou les sujets correspondent à une échelle en secondes pour le protocole conçu. Cette échelle n'est pas négligeable si l'application ciblée nécessite une réaction rapide ou un minutage précis du décodeur.
En résumé, les expériences en boucle fermée supportent l'idée principale d'améliorer les interactions en décodant des états cognitifs à partir de signaux EEG. Le décodeur proposé s'est avéré efficace pour décoder les états cognitifs liés à la difficulté, pour lesquels il faut également faire attention aux retards.
Mots clés : Electroencéphalographie (EEG), Interface Cerveau-Ordinateur (ICO), Interaction homme-machine (IHM), difficulté, charge de travail, retard, boucle fermée, dynamique.

Contents

List of Figures

List of Tables

Acronyms

PSD Power Spectral Density

log-PSD Power Spectral Density in logarithm

EOG Electrooculography

EMG Electromyography

EEG Electroencephalography

fMRI functional Magnetic Resonance Imaging

fNIRS functional Near-Infrared Spectroscopy

ECoG Electrocorticography

MEG Magnetoencephalography

BCI Brain-Computer Interface

HMI Human-Machine Interaction

A.U. Arbitrary Unit

NASA-TLX NASA-Task Load Index

SNR Signal-to-Noise Ratio

ERP Event-Related Potential

SC Shared Control

IIR Infinite Impulse Response

FIR Finite Impulse Response

ICA Independent Component Analysis

PCA Principal Component Analysis

LDA Linear Discriminant Analysis

SPHARA Spatial Harmonic Analysis

CSP Common Spatial Filtering

kNN k-Nearest Neighbours

SVM Support Vector Machine

ANN Artificial Neural Network

GLM Generalized Linear Model

CAR Common-Average Re-referencing

ANOVA ANalysis Of VAriance

1 Introduction

Interactions are ubiquitous in many forms such as human-human, human-machine, and human-animal. In order to achieve a successful interaction, it is important to understand the state of each participant. For example, a game designer may maximize the engagement of players if the designer can change the level of difficulty according to the player's perception of difficulty level. Frustration and disengagement can happen if the level is too high or low. This thesis will focus on improving user experience in Human-Machine Interaction (HMI) by decoding the cognitive states of human.

User experience generally can be linked to engagement or task performance. It was found that task performance is associated to arousal. As illustrated in Figure 1.1, the higher the arousal the better the performance, but the performance decreases after a certain threshold. This results in an inverted-U shape which is well-known as the Yerkes-Dodson curve [YD08].

Although Yerkes-Dodson curve was found on mice to form habits, similar phenomena are also observed on human behavior. Wright proposed a Motivation Intensity model and showed that the effort or arousal of participants dropped when the task difficulty was high enough [Wri08]. Faller *et al.* recently also showed that down-regulating a high arousal state improved the performance in their task [FCSS19]. From the angle of challenge point [GL04], the curve is not limited to arousal but also holds in other axes for learning tasks. In the learning tasks, the

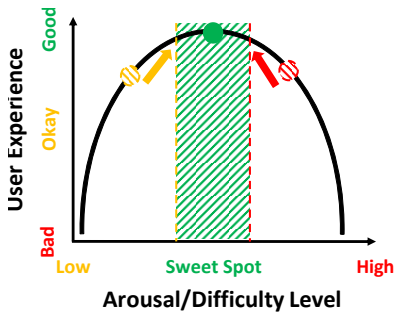


Figure 1.1 – Principle of improving user experience.

y-axis becomes potential learning benefit and x-axis becomes (functional) task difficulty level (to be further addressed in Chapter 4) that may link to different cognitive states other than arousal (see Chapter 1.2). As a result, a way to improve user experience is keeping the user inside the sweet operating region by adapting the interaction according to the user's cognitive states; this thesis, therefore, will decode relevant cognitive states and evaluate the feasibility of utilizing decoders by online experiments.

The following sections discuss how the concept was utilized in different applications of previous studies, and followed by the implementation of cognitive state decoders.

1.1 Applications

The idea of cognitive-based interaction can be implemented in either the user side or task (machine) side. At the task side, it can be implemented by directly tuning the task complexity or difficulty level, and then, the user's cognitive states should normally change afterward as in Figure 1.1. At the user side, provision of feedback about the current cognitive state can help the user self-regulate it, although this process results in an extra mental task [Mor08, RSdM$^+$16, FCSS19]. This thesis focuses on the task side because it does not introduce an extra task for the user.

This idea has been tested in different applications on the task side, but all have their own limitations. Szafir *et al.* designed an educational robot that can draw attention from students to improve their learning experience [SM12]. The idea was to detect a sudden drop of engagement index based on Electroencephalography (EEG) signals [PBB95, CRBP11, SM12, AG13] and had the robot either perform some movements or changing the tone of speaking when telling a story. The students then answered some questions of the story under three different conditions, minimal and pre-designed interaction, random interaction, or EEG-based interaction. Although on average students achieved the highest number of correct answers in the EEG-based interaction, there was no statistical significance compared to the random interaction. One potential explanation is that the engagement index could not properly reflect the cognitive state. As will be seen in Chapter 3, using this index will not lead to a high decoding accuracy. Szafir *et al.* should have noticed this issue and proposed dynamical thresholds to detect the loss of engagement.

For self-learning, adapting the difficulty level based on EEG signals was also found to be as effective as an error-adaptive system in learning a new arithmetic system [WRB$^+$17]. The subjects first performed arithmetic tasks in the decimal system in order to build a non-subject-specific model for the second study. In the second study, the subjects performed arithmetic tasks in the octal system where the difficulty level was adaptively determined by either the EEG system or the errors made by the subjects. Based on the adaptive system, the subjects experienced several difficulty levels during exercises. There were pre- and post-tests to evaluate the effects of adaptive learning. The results showed that using EEG signals can yield similar results as the error-adaptive system. Although the error-adaptive system was claimed to be

state-of-the-art for learning [WRB+17], it lacks a baseline or random condition. This can be crucial as there were 180 trials leading to 45-minute exercises. Another limitation is a long latency from the EEG decoder which used signals spanned around 80 seconds. As will be seen in Chapter 3, a long observation window can significantly improve the decoding accuracy at a cost of long latency. This can constrain potential applicability on other tasks and therefore deserves more investigation.

Another application is in the design of video games, where engagement is an important factor [Dic05]. A desired strategy is maintaining the difficulty level around the sweet spot of players, not too hard to frustrate the players and neither too easy to disappoint the players. Ewing *et al.* conducted an experiment with Tetris to dynamically control the level and compared to the self-paced decision [EFG16]. Their results suggested that the self-paced condition had the highest immersive experience and had many stable changes of levels than three other conditions using EEG signals. As reported by the authors, the EEG decoder was probably too simple. The decoder only compared the signal power in θ band at Fz and α band at P4 to a baseline recording where the subjects watched a relaxing video.

Professional activities can also benefit from having a neuro-feedback. Air-traffic control managers can avoid overloading their workload with an adaptive automation system that triggers at the right moment [ABD+16a]; hiding non-critical traffic information when overloaded for example. The automation system was defined together with the experts in air-traffic control which were therefore effective in reducing workload but being task-specific. In the presented studies, EEG signals have been shown useful in online experiments, but it was unclear whether the decoder could be transferred to other tasks. Moreover, the decoder also requires 30 seconds of data to make a decision, and thus, further confines any potential applications in need of a short latency.

On the other hand, although the interaction is made on the user side, pilots may avoid catastrophic consequences resulted from pilot-induced oscillations if they have feedback about their arousal state [SSJS16, FCSS19]. The subjects were asked to control a simulated drone with one degree of freedom in the altitude while advancing with a constant velocity. Along the flying path, there were some waypoints where the drone should get inside without hitting their boundaries, where a hit equals to crash in the protocol. If the arousal is high, a subject will likely induce high-frequency oscillations in the altitude and crash. The authors provided the estimated arousal from EEG signals as the volume of heartbeat from a headset, and meanwhile, the subjects had to minimize their arousal by any effective mental tasks. Their study showed that, being aware of the arousal, one can minimize the crash rate. The study, however, has restrictions on the decoding side. The online decoding accuracy was unknown as the ground truth was not accessible. Therefore, it is unclear if the subjects were really regulating their arousal. Besides, the generalizability of the decoder is also a question because the setting of only one-dimensional control is far from real-world situations and the decoder may be task-specific instead of really decoding arousal.

This thesis will focus on an application involving visuomotor activities, which can happen in many cases such as driving, eating, gaming, and sports. The common thing among them is the correction of error on-the-fly with visual information. For example, a driver aligning to the lane should constantly minimize the error between the heading direction and the line of lane, where the error is visually observed. The cognitive states can, for example, be utilized to tune the assistance level of smart driving systems. It is therefore of great interest to study the underlying cognitive states and their dynamics during visuomotor activities. Specifically, I will use piloting a simulated drone with two degrees of freedom as the main study platform which demands the abilities of orientation, visual tracking, and precise motor control. On top of it, a visual recognition task will be added to verify the generalizability which is missing in most of the literature.

1.2 Selection of Cognitive States

One challenge is targeting meaningful cognitive states for the defined application. This can be tricky as relevant cognitive states are not easy to disentangle. For example, by increasing the difficulty level of a task, one may not only have a higher arousal, but may also experience different levels of cognitive load [BPL03, APGvG10] or workload [HHA18], be aware of making more errors [CSM14, VPV11], lose the awareness of background situations [MRMH11, CPF16, BAV+14], regulate the level of attention [MMM14, KOL15, GMSW13], stay at a high vigilance [BLL+07], increase or decrease engagement [AG13, SM12], or induce different emotions [CRBP11]. Therefore, the inverted-U can still be observed with a different cognitive state as x-axis. These cognitive states may not always co-exist and can be subject or task-specific. As a result, focusing on a specific type of cognitive states may not be the best.

Given the Yerkes-Dodson curve, one may think that using arousal is sufficient. However, it has certain limitations. In many cases, arousal is linked or measured through cardiovascular activity, heart rates for example. However, utilizing cardiovascular activity has certain disadvantages. Heart rate may take a long time to reflect the current situation; after a running exercise, the heart recovery time is roughly two minutes [SDH82]. After one minute of recovery, the rate probably just decreased by 38% from the peak rate [JZBJ02]. Besides, the time resolution is generally as low as 30 seconds [RWAH16], which is coincidentally the same as the duration of moving average in the air-traffic study [ABD+16a]. Apart from the characteristics of time, not necessarily all the tasks can result in a great difference in cardiovascular activities. For example, in a simulated task of drone piloting, the normal-to-normal interval failed to show a statistical significance between the two conditions where significantly different workload levels were observed [DCA18].

As a result, this thesis will not focus on a specific type of cognitive state. I will try to elicit different cognitive states by tuning the objective difficulty level of the task to find physiological correlates. However, it is inevitable that a person can have a different skill level from others or a different background state on a different day. Using objective difficulty levels as the ground

truth of cognitive states, therefore, may not be suitable for the studies. On the contrary, subjectively reported values may be more useful to capture such inter- or intra-subject variations. I will ask for the perceived difficulty level from the subjects as a relative ground truth. Meanwhile, as NASA-Task Load Index (NASA-TLX) [HS88] is a widely adopted questionnaire for estimating cognitive load or workload, I will also include NASA-TLX in the following chapters.

The main physiological signals of interest are EEG signals for its potential. First, as pointed, peripheral signals such as cardiovascular activity are less sensitive and have worse temporal characteristics. This leads to the use of signals from the central nervous system, which generally refers to brain imaging techniques. Brain imaging technologies nowadays are either structural or functional imaging. Structural imaging only retrieves geometrical information and therefore is unlikely to indicate cognitive states. On the other hand, functional imaging measures signals from properties of blood, electricity, or magneticity associated with neuron activities that are likely to reflect cognitive states. Frequently used technologies are functional Magnetic Resonance Imaging (fMRI), functional Near-Infrared Spectroscopy (fNIRS), EEG Electrocorticography (ECoG), Magnetoencephalography (MEG), and so on. Each technology has its own range of spatial and temporal resolutions. Since a high temporal resolution is preferred, blood-based technologies (fMRI and fNIRS) are excluded. ECoG, although has high temporal and spatial resolutions, is excluded as an invasive method is not ethically justified for able-bodied people. MEG, although has a similar temporal resolution and a higher spatial resolution than EEG, is still paving its path toward being used as wearable devices [BHL+18]. Consequently, this led to using EEG signals to build a Brain-Computer Interface (BCI) that bridges the communication gap between humans and computers/machines [MC10].

1.3 EEG-based Paradigms

There already exist relevant studies decoding cognitive states from EEG signals. There are two major paradigms for decoding. One employs task-irrelevant stimuli with known timing while the other one focuses on the EEG correlates of the task-related behaviors and cognitive states. This thesis will focus on the EEG correlates because using stimuli has some disadvantages.

The rationale of employing a task-irrelevant stimulus is that the subject may not perceive or be attentive to the stimuli when her attention is mostly drawn to the main task. For example, Macdonald and Lavie designed a visual discrimination task and inserted a brief pure tone together with the visual display on a final trial. They showed that the subjects were not able to notice the tone in 79% of cases when it was a high-visual-load condition [ML11]. The lack of auditory attention can be detected through different amplitudes of some specific Event-Related Potential (ERP) in EEG signals [Fow94, MRMH11]. Therefore, this phenomenon can be used to measure the level of concentration which can be highly related to the task difficulty level. Employing this paradigm can be reliable in terms of decoding, although online decoding accuracy is still to be verified as the analysis was done in offline. Nonetheless, there are three potential drawbacks. One is that the sampling rate of cognitive state measurement may not be

high enough. Both the studies of Barry and Miller *et al.* requires the stimuli to be oddball or rare novel sounds [Fow94, MRMH11]. If the stimulus is being repeated at a short interval, the user may quickly start to ignore it or no longer be able to elicit the targeted ERP. Another is that user experience can be compromised with the irrelevant sound. A scenario is that a game (task) with good music will be interfered with by the irrelevant sound from time to time. Lastly, the irrelevant auditory stimuli may interfere in communications in human-human interaction. This can be problematic in critical situations, for example, when a pilot communicates with an air-traffic control manager in an urgent landing. Given these reasons, this kind of paradigm is not focused.

The other paradigm studies the patterns of EEG oscillations on the frequency domain that correlate with the targeted cognitive states [BAV+14]. Early studies usually focused on discovering statistical differences in frequency bands without verification of online decoding. Commonly reported patterns are decreasing and increasing of α band (8-13 Hz) and θ band (4-8 Hz), respectively, while increasing the task difficulty [APGvG10]. However, the θ band is not always found to increase [GTH+08]. The increase of workload can also correlate with the power in β band (13-30 Hz) band [Kli99, BWS96]. Patterns in the γ band (> 30 Hz) had also been reported, but less frequently [BLL+07]. These findings are not universal, for example, different spatial regions were found to be activated for skilled and normal subjects in a sentence verification task [APGvG10]. This suggests subject-specific decoders are likely to perform better.

These early discoveries served as a basis of later studies with online experiments. For example, Pope *et al.* evaluated different combinations of the frequency features on whether they really reflect the engagement and eventually proposed an index based on three frequency bands [PBB95]. Szafir *et al.* further utilized the engagement index in the mentioned study of adapting an educational robot [SM12]. Gilbert and Andujar were also developing an educational application based on the engagement index [AG13]. This engagement index in the original article used the sum of power from Cz, Pz, P3, and P4 [PBB95]. However, in the later studies, the acquisition systems did not even capture EEG signals from those electrodes [SM12, AG13], while the index seemed to work well in the educational robot. The exact definition of the engagement index becomes vague. Alternatively, in the previously mentioned Tetris study, the θ band at Fz and α band at P4 were specifically used for an online experiment [EFG16]. On the other hand, a recent trend is employing machine learning or advanced signal processing techniques for better modeling and decoding accuracy. This is reviewed in the next section.

1.4 Design of EEG Decoders

EEG signals have been considered in decoding different kinds of cognitive states, where the motor imagery is an example different from those proposed of this thesis. As the rationale of designing EEG decoders are nearly the same in most cases, this section therefore reviews some useful techniques to design EEG decoders.

Designing an EEG decoder usually involves the following building blocks: pre-processing (in-

cluding artifact reduction), feature engineering,[1] state inference (classification or regression), and post-processing. Their techniques are therefore accordingly revisited, while the real-time (low latency and causal) constraint is a strict requirement as the usability in online decoding is necessary.

1.4.1 Pre-processing

The electrodes on the scalp do not necessarily pick up signals directly generated by neurons. Other signals such as Electromyography (EMG) and Electrooculography (EOG) can also be present. Apart from that, skulls and cerebrospinal fluid have high electrical resistance and therefore blur the targeted neural signals [DLWK11]. Consequently, it is typical to utilize known characteristics of signals to purify the targeted signals from noisy observations.

Spectral filtering is a common technique to remove signals from unwanted frequency bands. The powers in EEG and ECoG signals are known to have an inverse relationship over frequency [HDP+18]. The higher the frequency is, the lower the Signal-to-Noise Ratio (SNR). Therefore, power in high frequency is easier to be contaminated by environmental noises or other physiological signals. If one would like to have good quality in γ band (> 30 Hz), it is important to record in an electromagnetic isolation room. This setting is not practical in general as suitable places are highly constrained. For the very low-frequency power, it is also known to have a drift in the signal over time. As a result, band-passing between 1 and 28 Hz is typical for example. Spectral filtering will have minimal influence on the power of spectral features due to the nearly flat frequency response. However, the filtering can benefit other techniques such as artifact detection and reduction based on the time-domain analysis.

Spatial filtering technique is also a common approach. This thesis uses a BioSemi amplifier which picks up the signals at the scalp and referenced to the CMS and DRL electrodes around the midline of the parietal region. Therefore, it may attenuate more local activities around the references than in the other regions. Moreover, manufacturers recommend always re-reference the recordings to increase the common-mode rejection ratio. Given that there is no concrete evidence that a few electrodes possess critical signals of the targeted states, a Common-Average Re-referencing (CAR) [BPP85] is therefore preferred as it does not introduce a strong spatial bias. However, a potential issue of referencing on any electrodes is spreading artifacts on other electrodes; if a referenced electrode has a strong artifact, the re-referencing will be largely affected. As a result, electrodes with a high chance of contamination should be avoided as a reference. Re-referencing in most cases is a spatial high-pass filter and not many studies considered or combined spatial low-pass filtering for reducing potential artifacts. The assumption is based on the fact that EEG signals have strong spatial correlations with neighboring electrodes. If a huge difference is found between neighbors (*i.e.* high spatial frequency), it is likely to be caused by artifacts or noises which should be excluded. The

[1] Feature engineering is crafting meaningful features from those collected signals based on domain knowledge, while feature selection is reducing number of features from a given set and is included in part of the pre-processing and the state inference in this thesis.

challenge is designing the filter on a non-uniform spatial sampling grid. Recently, there was a proposed method, Spatial Harmonic Analysis (SPHARA), tackled the challenge and can be applied for noise reduction [GEF+15].

Spatial filtering often includes methods that only employ linear combinations of EEG signals on the spatial domain. Another two common techniques are Principal Component Analysis (PCA) [WEG87] and Independent Component Analysis (ICA) [MBJS96, Vig97]. PCA is usually applied for dimension reduction but is not suitable when huge artifacts are presented. Artifacts added more variations to the data. As a result, PCA may consider artifact-induced variation as important. This, in fact, might improve decoding accuracy if the subject has a consistent artifact in a certain cognitive state. However, the decoder would be far from really decoding cognitive states. ICA on the other hand is more suitable for dealing with artifacts. ICA retrieves potential components by maximizing their independence. By inspecting each mixing vector or component, one can identify abnormal patterns. For example, a vector with a very high coefficient compared to others is aligned with the concept of a high spatial-frequency component. Another case is when the activity in a component does not resemble EEG signals. However, the challenge in either case is finding a reliable way to define artifacts. As the studies in the thesis inevitably involve visual activities such as blinking and target searching (saccade and ocular movements), EOG signals are recorded and are therefore reliable artifact references for those ocular components.

Data-driven filters were also developed. Supervised spatial filters are xDAWN [RSAG09], Common Spatial Filtering (CSP) [WGG06], and Fisher spatial filters [HVE06] for example. Naumann *et al.* also proposed a time-consuming source power co-modulation analysis using a data-driven spatial filter to select components that co-modulates with the target responses. It had been used to predict the difficulty level when playing video games. The predicted level was on average one level different from the true level; nevertheless, the method was not tested in online experiments [DMH+14, NSKDB16]. Co-design of spectral and spatial filters is also possible. A known one is filter-bank CSP [ACZG08]. These filters are in fact more like feature engineering or selection, as the discrete labels for later classification or regression are required to train these filters.

1.4.2 Feature Engineering

The quality of features can be more crucial than applying advanced statistical learning algorithms for improving decoding accuracy. Based on a given input, the selection of features can be completely data-driven or manually selected based on domain knowledge (feature engineering). A data-driven approach can be implemented by deep learning or some other advanced machine learning algorithms to automatize the process. However, deep learning, which has shown its powerfulness in many other fields [LBH15], is not yet the leading technique in EEG signals [LBC+18]. The reason is probably the insufficiency of data. The backbone of data-driven methods with many parameters is a huge amount of data, which is not easy

to collect in this field. Considering that the amount of collected data in this thesis is unlikely to be enough, which may prone to overfitting, especially when subject-specific models are preferred. The focus is therefore on the manual selection.

The power spectrum is typically used features in EEG studies where a short-time sliding window is usually implemented for real-time decoding. There are several methods to estimate the power-spectrum of short-time windows. Hamming window is typically used nowadays and the cancellation of side lobe reaches nearly the lowest level [Har78]. The estimation of power nonetheless may be inaccurate as there is only one realization of the window. A multi-tapper method employing different types of windows was therefore introduced and favors more accurate estimation with a higher computational cost [Tho82, Tre95].

Based on the relevant studies, the engagement and attentional indices are likely to represent the relevant cognitive states in this thesis. The engagement index is defined as the ratio $\beta/(\alpha+\theta)$ without consistent report of electrodes [PBB95, CRBP11, SM12, AG13], and the attention index is defined as θ/β where Fz, F3, and Fz were used [PVAG$^+$14]. The definitions of frequency bands are 4-7 Hz for the θ band, 8-12 Hz for the α band, and 16-28 Hz for the β band.

Recently, using the covariance matrix of temporal samples as features has drawn attention in the BCI research community. This is promoted by introducing a manifold named Riemannian geometry to compute the distance between features (covariance matrices). The method using Riemannian geometry has won several competitions in offline analysis [CBB17].

1.4.3 State Inference

The outcome of decoding cognitive states can yield either a continuous value or a descriptive label. For example, the perception of difficulty level can be reported as a value between 0 and 100 or can also be described as easy, hard, or extremely hard. As a result, the inference of cognitive states can be modeled as a regression or classification problem.

For classification, Lotte *et al.* made a review article for most EEG protocols, while the majority being motor imagery [LCL$^+$07]. The typically used classifiers are k-Nearest Neighbours (kNN), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Artificial Neural Network (ANN) *etc.* According to the article, SVM stood out of the crowd in many aspects. On the other hand, LDA is frequently employed in motor imagery studies. No matter which classifier is used, the curse-of-dimensionality requires the use of feature reduction in order to avoid overfitting [Fri97]. The filter techniques such as Fisher score is a typical method to reduce the dimensionality, while wrapper methods co-optimize the features with the assigned classifier to have a better accuracy–usually at a cost of large computational time [LBC$^+$18]. One example of wrapper method is an automatic stop stepwise (wrapper) LDA developed by Aricò *et al.* [ABD$^+$16b].

Regression usually can be found as an intermediate step of a linear and binary classifier, *e.g.* the

principle of popular SVM computes the distance between a feature and a hyper-plane, where the distance is to be binarized by a threshold. Although this can be done easily, regression is less investigated or used in BCI. This is probably because that the majority of BCI studies concern trial-based protocols with discrete commands which are naturally formulated as a classification problem. Regression, however, is a suitable method for decoding perception of difficulty or cognitive workload in this thesis, because the nature of such cognitive states is better described in a continuous scale rather than discrete states. For example, when providing self-evaluation in the tasks used in this thesis, some subjects found it hard to decide whether the difficulty level was easy or hard, meanwhile, giving a relative value is easier.

Ordinary least square solver is the standard regression approach. The solver assumes the error distribution is Gaussian with the mean being the hyperplane of regression. One can also utilize Generalized Linear Model (GLM) to change the targeted distribution [DB08]. As the decoder is assumed to be a binary classifier most of the time, using a binomial distribution is a reasonable choice. Similar to classification, regression also has several techniques with an embedded feature selection process. This is typically done by adding regularization terms, such as sparsity, Tikhonov (ridge regression), and elastic net regularizations [ZH05].

There exist a huge number of classification methods. For interested readers, comprehensive reviews and update of classifiers based on EEG signals are given by Lotte et al. [LCL+07, LBC+18]. However, most methods in the reviews were evaluated in an offline setting where the subjects did not have BCI feedback. In the online (asynchronous in the articles) cases, there is no single classifier considered as superior to the rest [LCL+07]. This probably suggests that the offline validations were overfitting, or the non-stationarity of the EEG signals from the user side is larger than expected when tested in online. The non-stationarity can come from different setups, background states in different days, or different efforts on interacting with the decoder. A future direction to improve is probably mutual learning [PTS+18]. It has been mathematically shown that having non-zero learning rates of both the human and the machine can further improve the whole performance, under the assumption that the human and machine use linear models in learning each other. The mathematical theory was verified by a behavioral experiment that was designed based on the assumption [MVS+17].

1.4.4 Post-processing

EEG signals are known to have a low SNR which can result in unreliable decoding output. As noise is usually assumed with a zero mean, the noise can be suppressed by looking into more samples or trials. This has been shown effective in many studies [LSC04, SS09, PBLM11]. This kind of approach also eliminates a common assumption of classification algorithms that each sample is independent. However, this assumption is mostly incorrect in physiological signals, because close-in-time samples are highly correlated. There are various techniques, such as averaging the decisions across trials [SS09], and applying a low-pass filter on the probabilistic output of decoder [LSC04, PBLM11]. The concept of using more samples can also be applied

to feature extraction by using low-rank decomposition for example [JCM18b]. In general, the more samples or trials are used, the more reliable the decoder is. This, however, increases the latency of decoders and needs to be compromised.

1.5 Organization of Thesis

The main goal of this thesis is to investigate the effect and the practice of integrating the EEG-based cognitive state decoder inside the interaction loop between humans and machines. As a result, the thesis is divided into several objectives.

One objective is examining how well a cognitive state decoder can replace or even outperform a behavior-based decoder under the principle shown Figure 1.1. Different from the previous study using a similar protocol [FCSS19], this thesis adapts the interaction from the task side and has more degree of freedom in controlling the drone, which in turn, has the potential of higher generalizability. Additionally, this thesis will focus on decoding cognitive states with the latency as short latency as possible to enable more flexibility on future applications that was not considered in many studies presented in Chapter 1.1.

The second objective is studying the generalization of the cognitive state decoder in different tasks. To the best of my knowledge, the surveyed articles of relevant BCI did not test the applicability in different tasks. It is therefore worthy of investigation if similar neural correlates can be identified and whether the same decoding framework is applicable in different tasks.

The third objective is studying practical issues that should be concerned in designing such HMI system. For example, when will the output of a cognitive state decoder become reliable, and how to design a decoder with good decoding accuracy.

Following these objectives, this thesis is organized as the following chapters.

Chapter 2 Bi-directional Online Regulation This chapter aims at an early proof-of-concept that the proposed protocol can benefit from decoding EEG signals. This is done by comparing the EEG condition with a sham condition based on the user's behavior. In this study, the regulation of cognitive states is bi-directional where the task difficulty can increase or decrease, and change on a scale of less than one second.

Chapter 3 Decoder Optimization A practical issue is building a reliable decoder. According to the result of Chapter 2, a commonly used decoder in BCI applications should be further improved. Based on the data of Chapter 2, this chapter compares several decoders, including a new one, and selects the decoder has the highest accuracy for the later chapters.

Chapter 4 One-directional Online Regulation This chapter continues the study in Chapter 2. With a revised protocol based on the feedback of subjects, this study mainly regulates the cognitive states from the left-hand side of the inverted U curve. The revised protocol

also allows an evaluation of online decoding accuracy to complement Chapter 3. This however requires a slower interaction but the decoder was still running at a scale of less than one second.

Chapter 5 Dual Task and Temporal Dynamics Previous chapters are restricted to a single task with different difficulty levels. To enable more scalability, I further introduce a second task to the protocol. Furthermore, a practical issue in HMI system is that a delayed output from the decoder may degrade the user experience. The temporal dynamics of decoded cognitive states are evaluated to reveal latencies contributed by the decoder and user.

Chapter 6 Conclusion The contributions are summarized and future works are discussed.

2 Bi-directional Online Regulation

Disclaimer: This chapter is adapted from the following article–with permissions of all co-authors and the organizer:

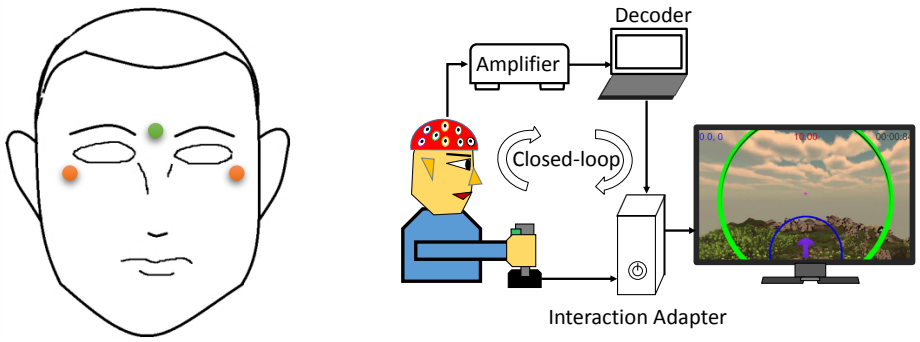
P-K Jao, R. Chavarriaga, and J.d.R. Millán. "Analysis of EEG correlates of perceived difficulty in dynamically changing flying tasks." in IEEE International Conference on Systems, Man, and Cybernetics, 2018.

2.1 Introduction

User's cognitive states normally can locate at any point on the x-axis of the curve in Figure 1.1. Adjusting the task difficulty from either side of the curve helps reach the sweet spot; lower the task difficulty when it is too difficult, and raise the task difficulty when it is too easy. The freedom of bi-directional regulation on task difficulty preserves the possibility of convergence from any cognitive states but has a cost of potentially oscillating around the sweet spot.

In order to conduct such a study, a protocol is needed to collect data for building decoders and to ensure a principle of regulating the task difficulty. As defined in Chapter 1, drone piloting is the main task to study cognitive states. The designed task (to be detailed in Chapter 2.2.3) requires the subjects to steer the direction of a simulated drone toward a certain waypoint. At least two extreme difficulty levels will be presented to elicit the cognitive states for bi-directional regulation.

There are two approaches to regulate the difficulty level with the designed task. One approach is switching on or off an assistive piloting system. In order to avoid the loss of engagement, the subject should remain control of the drone while the assistive system tries to adjust the drone with a computer-decided orientation. This system belongs to shared control [MAC15]. The other approach is directly controlling the task difficulty, given that the protocol is implemented in a simulated environment. Although directly controlling the task difficulty level is unlikely to happen in reality, this approach ensures well-defined task difficulty levels. Hence, it is also



(a) EOG setup (b) Protocol setup

Figure 2.1 – Experiment setup.

considered for its suitability of scientific studies.

Following the above ideas, the EEG correlates under different levels of difficulty will be studied. Meanwhile, a simple shared control mechanism will be tested for its usability in online interaction. This thesis does not dedicate much to the shared control as it is task-specific.

2.2 Material

2.2.1 Recording Setup

All the physiological signals were acquired using a Biosemi ActiveTwo system running at a 2,048 Hz sampling rate. EEG signals were recorded with 64 channels following the extended 10-20 international system. Apart from EEG signals, EOG and EMG were also recorded for artifact removal. The setup of EOG is illustrated as Figure 2.1(a), one was placed on the midline of eyebrows, and the other two were on the bulge bones below the outer sides of canthi. For EMG, two bipolar channels were placed on the biceps and triceps of the right arm.

2.2.2 Participants

Twelve graduate students (three females; mean age: 23; standard deviation: 1) participated in three different recordings (see Chapter 2.2.4). As shown in Figure 2.1(b), each subject sat in front of a computer screen and used their right hand holding a joystick[1] to pilot a drone in a simulated environment, which was implemented with Unity (https://unity3d.com/). One male subject was left-handed but reported to be comfortable using the right-handed joystick. The protocol was approved by the local ethical committee and all subjects provided written consent.[2]

[1] Logitech Extreme 3D Pro, https://www.logitechg.com/en-us/products/gamepads/extreme-3d-pro-joystick.html

[2] This paragraph is adapted from ©2018 IEEE.

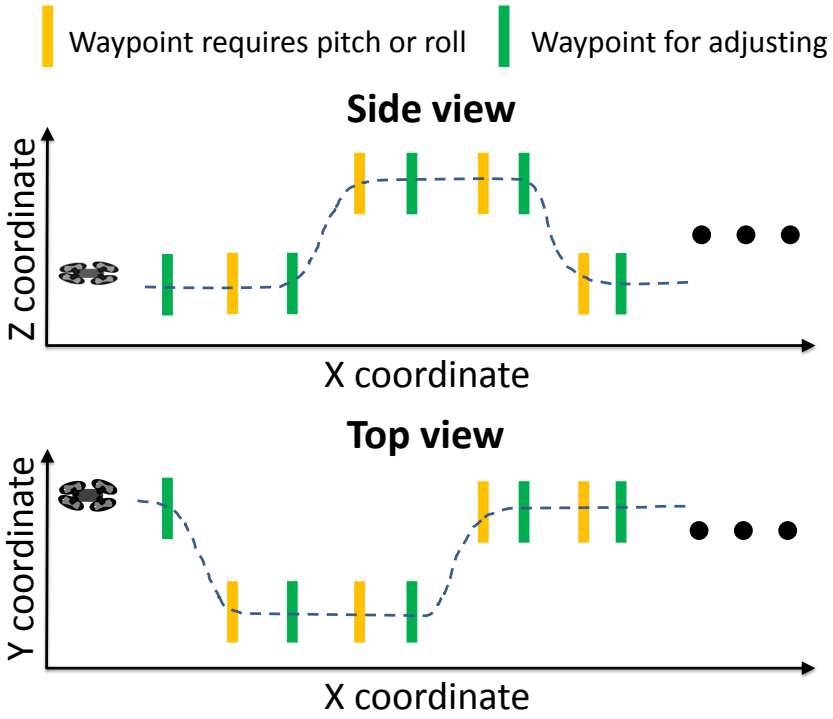


Figure 2.2 – Layout example of waypoints

2.2.3 Task

The task is the same for the three different recording sessions. Each session includes multiple trajectories to be passed through by the subjects. An example of trajectory is presented in Figure 2.2, where a side view and a top view are provided to handle three-dimensional coordinates. Each trajectory was composed of many waypoints for which the subjects were instructed to steer through with the drone which flew in a constant velocity. The only allowed movements of the drone are pitch and roll.

Figure 2.1(b) presents a screen-shot of the task. The green circle is the current waypoint to go through while the next one appears in blue. Further waypoints are not visible. The purple cross indicated the looking direction of the drone and was used to determine whether the drone passes (hit) a waypoint. Therefore, the subjects were informed that a good strategy is to aim at the center of targeted waypoint with the cross. If the subjects failed (miss) to go through the waypoint, they were instructed to proceed to the next one. The purple arrow below was a 3D indicator pointing to the center of the current waypoint in case of losing the waypoint on the screen. The numbers on top display the number of passed waypoints, score, the radius of waypoint, and elapsed time. When hitting a waypoint, $1/r$ of points were gained, where r is the radius of waypoint.

2.2.4 Design of Sessions

Recording sessions were spaced for at least one week. Two of the recording sessions were open-loop and the final one was closed-loop. In the open-loop case, the protocol followed the pre-designed parameters, *e.g.* radius of the waypoint, of task difficulty without dynamically changing them. The major goal of open-loop recordings was collecting preliminary data in order to build EEG decoders for the designed task. For the closed-loop recording, the decoder passes estimated cognitive states to the interaction adapter for regulating the parameters of task difficulty. The objective was to test the decoders and to compare the result with a pseudo ground truth condition based on the subject's behavior.

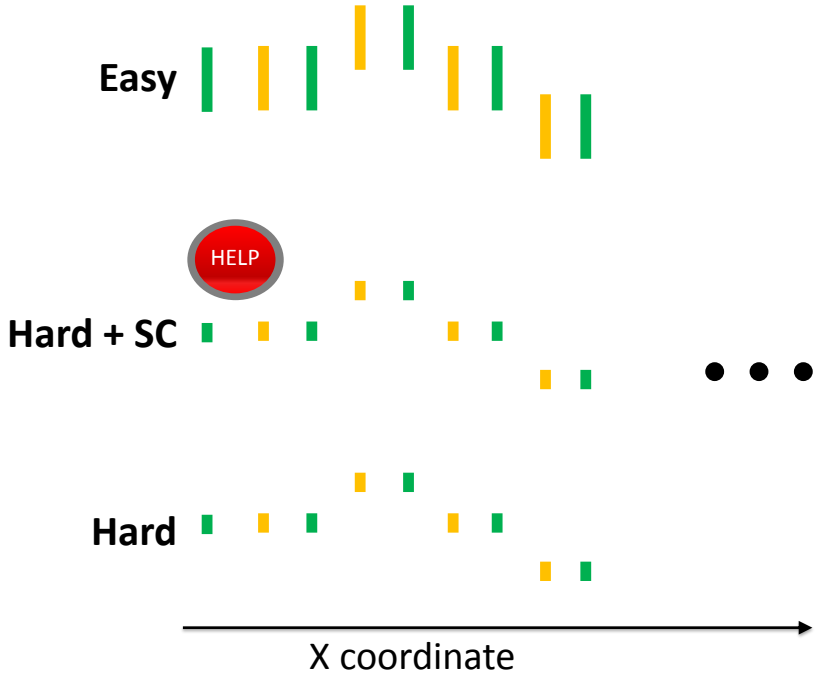
Session 1

As demonstrated in Figure 2.3(a), each trajectory in the first session was designed with a uniform radius without any dynamical changes. There are 15 trajectories where each comprises 34 waypoints. The radius of a waypoint was either 6.0 or 0.3 Arbitrary Unit (A.U.), referred to as Easy or Hard conditions, respectively. For the 0.3 A.U., there was an extra condition, the subjects flew with a pre-designed shared control mechanism to see if it can help reduce the perceived difficulty level. If it can, the shared control will be included in the future online adaptation to adjust the difficulty level.

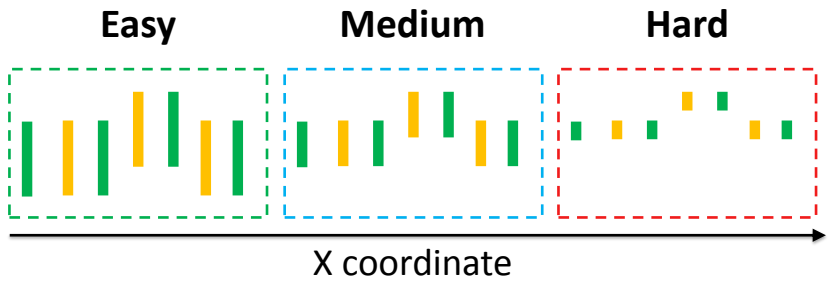
The designed shared control aims at not competing with the user for authority over the drone, and thus tries to provide a hardly perceivable assistance in changing the steering direction. The design can be expressed with the following equations,

$$\begin{aligned} d_f &= d_h + c_0 \times d_c\,, \\ d_c &= (d_a - d_h) \times \exp(-\hat{e}(t)/c_1)\,, \\ \hat{e}(t) &= |d_h - d_a| + c_2 \times \hat{e}(t-1)\,. \end{aligned} \tag{2.1}$$

The first equation defines the final steering direction, d_f, which is a simple linear combination with the user's control signal, d_h, and the final assistance, d_c. Assuming an automatic steering direction, d_a, is generated, the $(d_a - d_h)$ term in the second equation makes the final steering as d_a. However, as hardly perceivable assistance is of interest, scaling is needed for computing the final assistance. Therefore, a dynamical scale according to $\hat{e}(t)$ is implemented with an exponential factor tunable by a variable c_1. The third equation defines $\hat{e}(t)$ at time t and introduces a forgetting factor c_2 to accumulate previous error, $\hat{e}(t-1)$, in order to smooth the transition of d_c. Under this design, the final steering direction is compromised between the input of subject and an automatic navigator. However, if both directions differ too much, the steering direction will be dominated by the input of subject. Only when the difference between both is small enough, the navigator can contribute to fine-tuning the steering direction.



(a) Session 1



(b) Session 2

Figure 2.3 – Designed conditions in session 1 and 2

There were in total of five pre-designed trajectories that were used in each of the three conditions. In other words, each trajectory was used once in each condition. The presentation order of the fifteen trajectories was random. After each trajectory, the subjects filled the NASA-TLX and reported the perceived level of difficulty on a scale from 0 to 100.

Session 2

Each trajectory in the second session has 42 waypoints. As depicted in Figure 2.3(b), each trajectory comprised three different radii, where one radius forms a group of waypoints. The three radii were 6.0, 3.0, and 0.3 A.U. corresponding to Easy, Medium, and Hard levels,

respectively. The radii decreased twice within a trajectory. To avoid habituation, the first time of decreasing happened at 8^{th}, 10^{th} or 12^{th} waypoint while the second time at 16^{th}, 20^{th} or 24^{th} waypoint. Subjects were required to pass through at least 34 of them to make sure each condition has sufficient data. Missing a waypoint after 34^{th} would immediately terminate the trajectory in order to motivate the subjects to stay engaged. In total, the subjects conducted fifteen trajectories.

Session 3

The third session was designed with a dynamically changing radius. The radius was bounded between 6.0 and 0.15 A.U. with a change rate of 1 A.U. per second in a 16 Hz update rate. There are two conditions, sham and EEG. Although the sham condition used a behavioral decoder, the subjects were informed that it is an EEG decoder in order to have similar behavior as in the EEG condition. The decoder decides whether the current moment is Easy, Hard, or Comfort/Uncertain. In order to optimize the engagement, the radius decreases (increases) if it is Easy (Hard) and stays the same if it is Comfort/Uncertain.

There were eight trajectories for each condition. The first eight were either all EEG condition or the sham condition, and the other condition for the last eight. Each trajectory had 66 waypoints, but a trajectory could finish earlier if ten waypoints are missed, where the remaining chance of missing was displayed on the screen. The subjects were instructed to try their best to hit as many waypoints as possible–achieving the highest points.

The EEG decoder was built based on the data of the second session and the implementation details are presented in Chapter 2.3. On the other hand, the behavioral decoder used the aiming error to adjust the radius. The aiming error was computed by the distance between the purple cross and the center of current waypoint on the plane where the waypoint is. The decoder would consider as Hard if the distance is larger or equal to two times the current radius and as Easy if less than 1.1 times the radius. For the case in between, the decoder would output uncertain. As the subjects were supposed to consider it as another EEG decoder, the output of the behavioral decoder was added with a little noise to favor incorrect decisions and was low-passed to introduce a delay.

Placement of waypoints

For all the sessions, waypoints composing a trajectory were arranged in a way that the subject needed to either perform a pitch or roll every other two waypoints as illustrated in Figure 2.2; the waypoints in between were simply placed 30 A.U. in front of the previous one for letting the subject to adjust the orientation of drone. The numbers of required pitch and roll maneuvers were balanced for all the directions. The Euclidean distance between waypoints for pitch was 31.6 A.U.. and 42.43 A.U. away for roll maneuvers.[3]

[3]This paragraph is adapted from [JCM18a] ©2018 IEEE

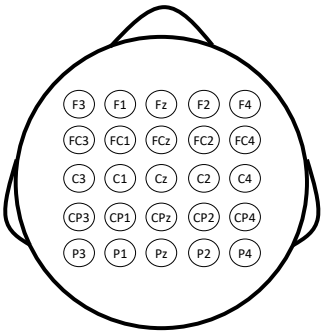


Figure 2.4 – The main twenty-five electrodes being analyzed.

2.3 Decoding Perceived Difficulty

2.3.1 Signal Pre-processing

Peripheral EEG electrodes were left out of the analysis to reduce the likelihood of muscular contamination, as some subjects made movements either with the neck or body. The remaining twenty-five channels, as plotted in Figure 2.4, are: 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.

EEG, EMG, and EOG signals were down-sampled to 256 Hz and bandpass filtered using a 13^{th} order Butterworth from 1–40 Hz. For the EEG signals, CAR was first applied to reduce common noise [BPP85], and a 20^{th} order SPHARA was applied to reduce high spatial frequency components [GEF+15]. EMG signals were extracted by computing bipolar signals. EOG signals were separated into vertical and horizontal components. The vertical one was computed by having the top electrode subtracts the average of the other two, while the horizontal one was the subtraction between the two electrodes on the sides.

To evaluate potential contaminations in EEG signals, ICA was performed [MBL+00]. The number of components was assumed to be fifteen and their correlations between EMG, vertical and horizontal EOG components were computed. None of the EEG components had a correlation larger than 0.6, and was not further removed from the analysis.

To extract the power spectrum, Power Spectral Density in logarithm (log-PSD) was estimated for all electrodes using a one-second sliding window with a 2Hz shift rate in the offline analysis and 16 Hz in online decoding. The frequency range of interest is from 1 to 28 Hz and yielding a total of 700 features. During the extraction, each window was also examined if a strong artifact presents. If the absolute potential in any electrode exceeds 50 μV, the window is considered as contaminated and discarded for later analysis. In the online decoding, if it happens, the window is considered as Uncertain.

2.3.2 Classification

Based on the previous reports of engagement index and correlation in the frequency bands, the 700 features were further reduced by using the following frequency bands: the δ (3 Hz), θ (4-7 Hz), α (8-13 Hz) and β (16-21 Hz) bands. They were computed by the average of frequency bins within the ranges and eventually reduced to 100 features. To further refine the number of features based on the collected data, Fisher scores were computed and the top twenty features were kept. These features were then feed-forward to a regularized LDA with a uniform prior to classify between Easy and Hard conditions.

2.3.3 Offline Validation

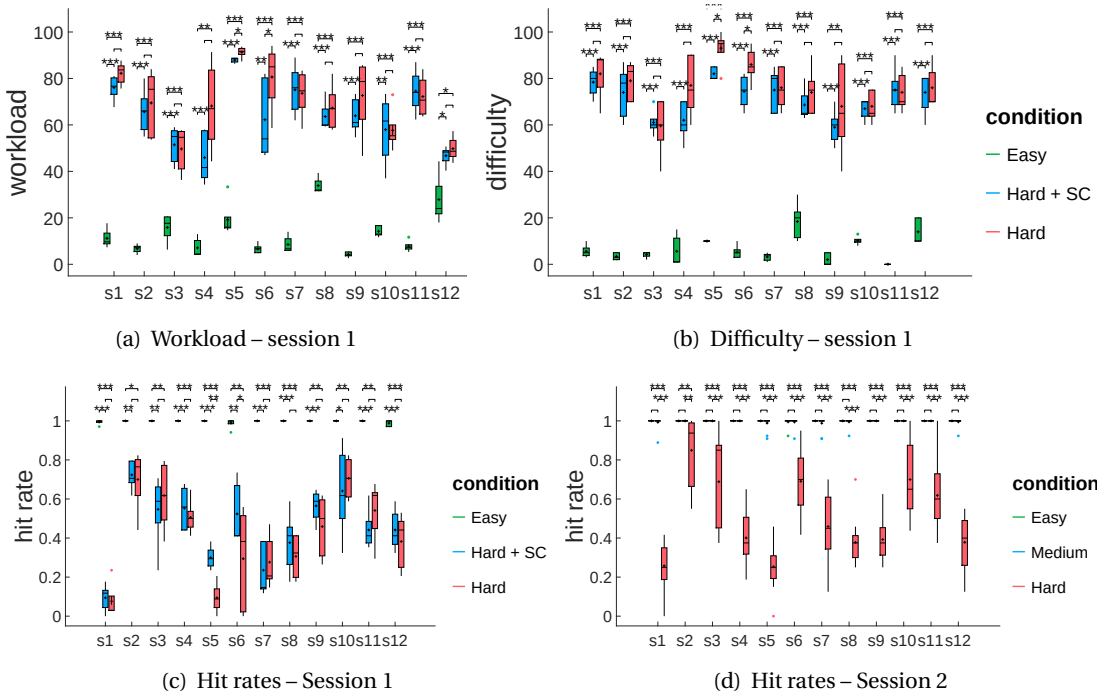
Although each session involved three conditions, only Easy versus Hard were analyzed. This is to fit the bi-directional regulation in online decoding. Further analysis is presented in Chapter 3. A leave-one-trajectory-out cross-validation was applied to estimate the decoding accuracy of the modeling process for each subject. The accuracy was estimated by the class-balanced accuracy in order to neutralize the effect of potentially imbalanced samples of each class. Normal accuracy can be biased if the number of samples in each class is not the same [FWM+08]. The class-balanced accuracy first computes the accuracy of each class, and then averaged across classes such that the accuracy of each class has the same weight regardless of the number of samples. However, the decoding accuracy in the first session degenerated to the standard accuracy–the ratio of correct prediction–because each trajectory had only one class as ground truth.

2.3.4 Online Tuning

The decoders in the online sessions were trained by using the data from the second session with the window labeled as Easy or Hard. This selection of data is the best fit as a larger signal variation was expected between the two classes.

The decoders might not work well in the online session as they were built with data recorded on different days. In order to alleviate this potential situation, the model was manually fine-tuned by trisecting the output of the decoder. The output of decoder was a probability value and two thresholds were to be determined for dividing into three regions: Easy, Comfort, and Hard. The tuning procedure took place before the first and fifth trajectory of both the EEG and sham conditions.[4] The subjects orally communicated to the operator their perceived difficulty level in order to determine the two thresholds on-the-fly when steering the drone in a training trajectory. The subjects were informed that tuning the thresholds should help them achieve the highest scores.

[4] The tuning procedure in the sham condition was to make the subjects to consider it as another EEG decoder.



(a) Workload – session 1

(b) Difficulty – session 1

(c) Hit rates – Session 1

(d) Hit rates – Session 2

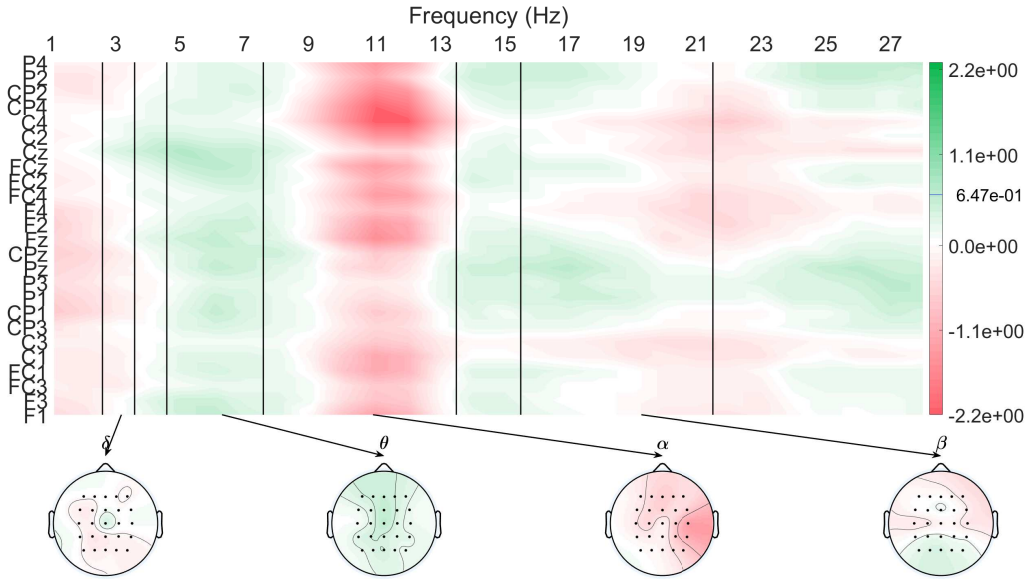
Figure 2.5 – Subjective assessments and behavioral outcome in the offline sessions. (a) and (b) are subjective assessments in the first session. (c) and (d) are the hit rates in the two sessions. Paired-wised t-tests were applied (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). The crosses indicate mean values. Colorful dots are outliers. Horizontal lines in the boxes are median values.

2.4 Results

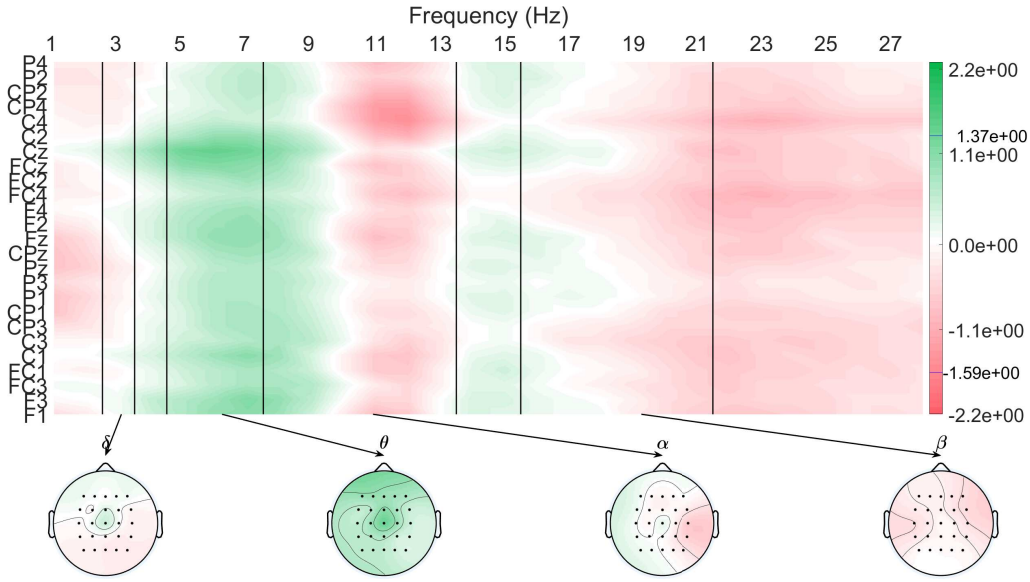
2.4.1 Offline Behavior

Figure 2.5(a) and 2.5(b) plot the reported workload and perceived difficulty levels, respectively, in the first session. Each box represents a condition of a subject, where one color stands for one condition as indicated in the legend. Pair-wised t-tests were applied between each pair of conditions for each subject ($n = 5$). Significant results are mostly found between (Easy, Hard) and (Easy, Hard + SC (Shared Control)) but nearly none between (Hard + SC, Hard). Clearly, the designed shared control mechanism was not powerful enough to reduce the workload or perceived difficulty level in a difficult scenario. On the other hand, the Pearson's correlation between workload and difficulty levels is $r = 0.95$ ($p < 10^{-95}$, $n = 195$). This suggests that only reporting difficulty levels should be sufficient in most cases.

Figure 2.5(c) and 2.5(d) plot the first and second sessions, respectively, their hit rates of each condition and subject. Similarly, pair-wise t-tests were applied ($n = 15$). The hit rates in the first session are consistent with the subjective assessment such that only a few significant cases are observed between (Hard + SC, Hard) and all are significant in the other two pairs. As the



(a) Session 1



(b) Session 2

Figure 2.6 – Grand average log-PSD of Hard minus Easy. The grand average was performed over windows, and then subjects. Red (Green) means that a lower value for the Hard (Easy) condition. White means no difference. The blue line on the color bar indicate the extreme values on the data, otherwise, the extreme values of the data are the same as the limit of the color bar.

shared control was not able to significantly influence either hit rates or subjective assessment of cognitive states, the shared control was therefore not further employed in the other sessions. On the other hand, in the second session, the medium difficulty was not enough to cause a significant difference in hit rates from the Easy condition, but there were more mistakes in the Medium case as many outliers can be spotted in Figure 2.5(d). As a result, it was not clear if the Medium condition should be categorized as Easy or separated as Comfort.

2.4.2 Power Spectrum Correlates

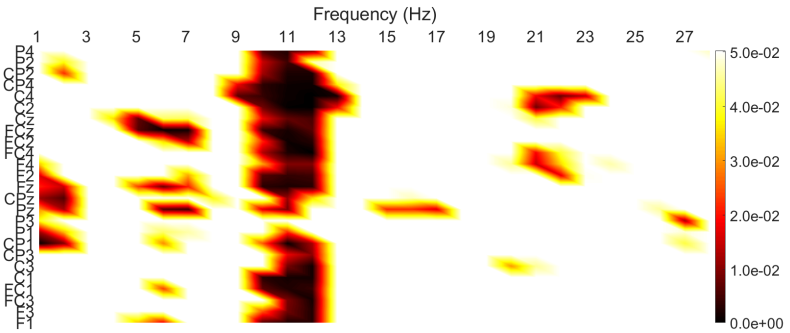
Figure 2.6(a) and 2.6(b) are the difference of grand averaged power spectra between the Hard and Easy condition for the first and second sessions. The topoplots below show the difference in the four frequency bands selected in Chapter 2.3.2 for all the electrodes of interest (Figure 2.4). White means there is no difference between the two conditions, while the darker the red or green means the two conditions are easier to be distinguished.

The power spectra have some consistency across both sessions. For the region around Cz, the powers in the δ and θ bands are green, which means higher power for the Hard condition. The powers in α bands around C4 and CP4 are red, which means higher power for the Easy condition. The powers in the β bands are also similar but are found in both central-lateral regions. As a higher power is usually associated with the Easy condition, it suggests that potential muscular activities were not dominating the measured signals. The reason is that, in the Hard condition, the subjects may prone to contract facial muscles and contaminated signals with a higher power than EEG signals.

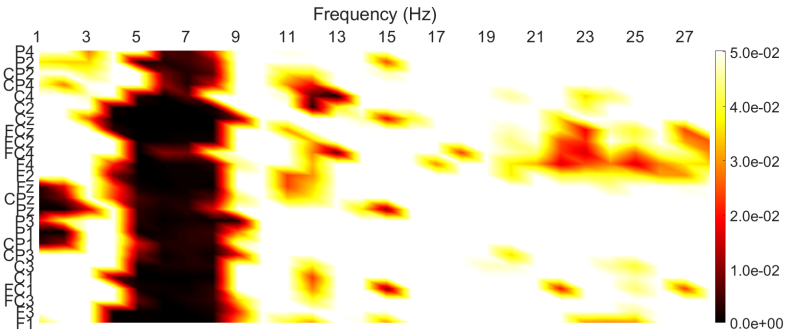
Figure 2.7(a) and 2.7(b) depict the p-values of paired t-tests (n=12, one sample is one subject) based on the data of Figure 2.6. Darker color indicates that the p-value is 0.05 or below. Figure 2.7(c) further emphasizes the frequencies and channels that yielded significant results across the two sessions. It can be seen that the previously mentioned δ and θ bands at Cz, α band around C4 and CP4, β band around central-lateral region are all consistently significant.

2.4.3 Offline Decoding Accuracy

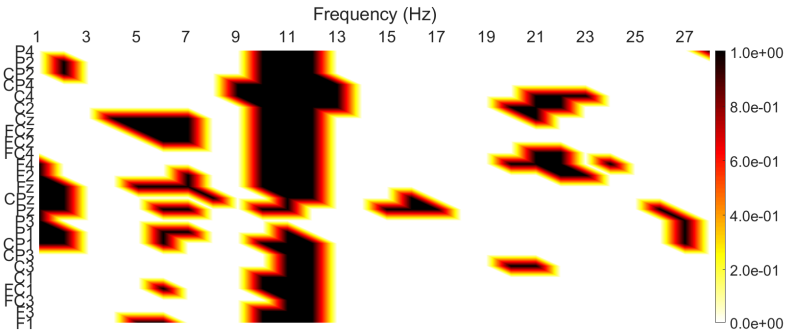
Figure 2.8(a) and 2.8(b) depict the decoding accuracy in the first and second sessions, respectively. For the first session, eight subjects out of twelve yielded results that were significantly better than the chance level (t-test, n = 10). The mean accuracy across subjects was 0.61 with a 0.05 standard deviation. In the case of second session, all the subjects had class-balanced accuracies significantly higher the chance level (t-test, n = 15). The mean accuracy across subjects was 0.64 with a 0.05 standard deviation.



(a) Session 1

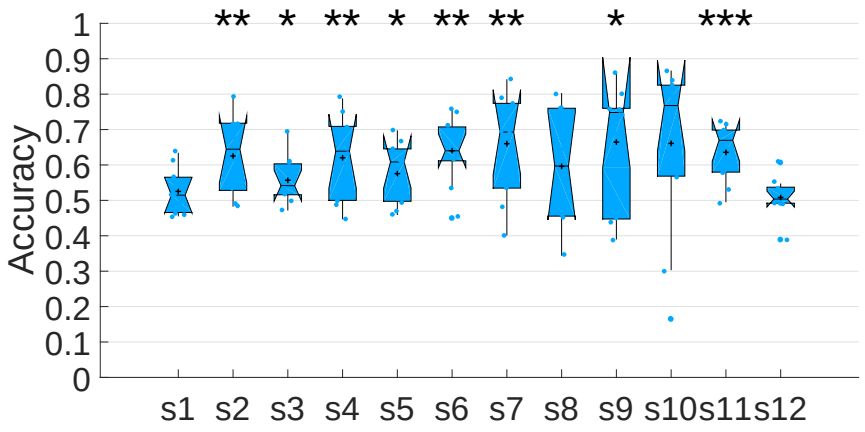


(b) Session 2

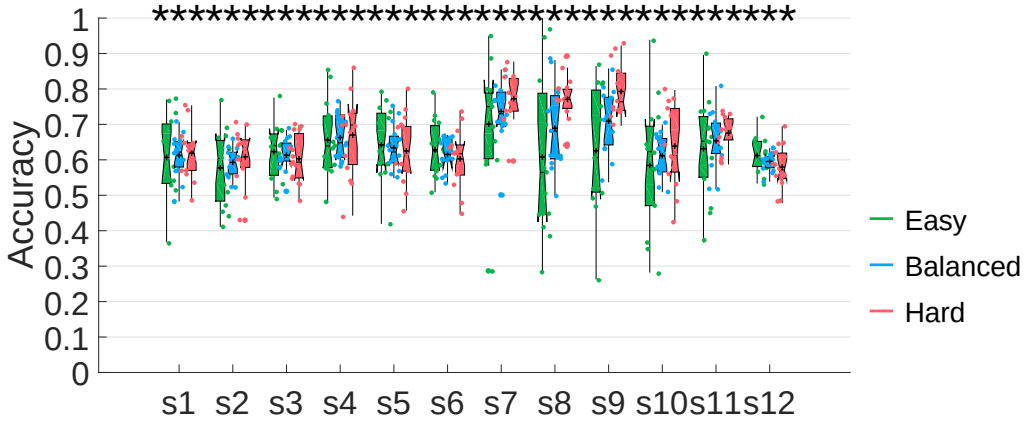


(c) Consistent significant results

Figure 2.7 – (a) and (b) show the p-values of t-tests (n = 12) based on the data of Figure 2.6. Values equals to or below 0.05 are highlighted as darker color. (c) illustrates consistent significantly different features in dark while white indicates inconsistency (at least one session is insignificant).



(a) Session 1



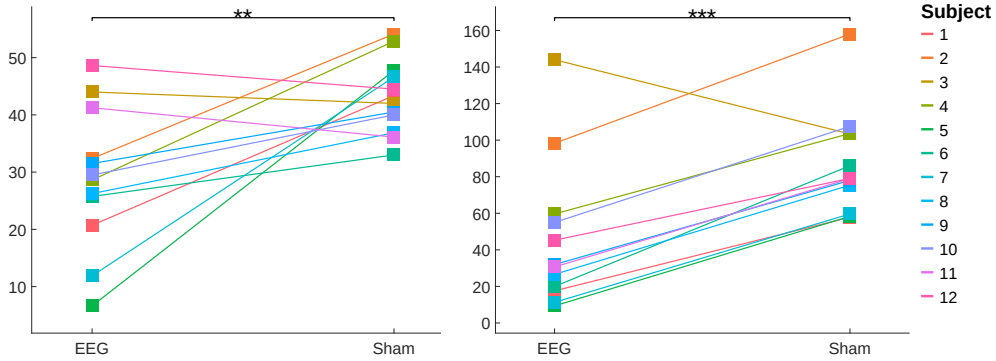
(b) Session 2

Figure 2.8 – Decoding accuracy in offline validation for Easy versus Hard. T-tests over the balanced accuracies were applied against the chance level, 0.5 (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

2.4.4 Online Behavior

Figure 2.9 illustrates the number of hit and the scores for each subject in the EEG and sham conditions. Only three subjects passed more waypoints than the sham condition. On average, the sham condition passed 27.1 (± 14.2) more waypoints than the EEG condition. In terms of points, one subject earned remarkably higher points in the EEG condition than in the sham condition. On average, in the sham condition subjects collected 41.3 (± 27.1) more points than in the EEG condition. For the subjects with more hits and fewer points, the radius was likely biased to a larger value. On the other hand, a subject with a lower hit count and fewer points indicates that the radius was likely too small.

Figure 2.10 draws the distributions of radius in each condition and subject, where a curve is a probability density function of radius computed from all the time points and all the trajectories. A blue curve stands for the EEG condition while a red one is a sham condition. As



(a) Number of hit (b) Points

Figure 2.9 – Behavioral result in the third session

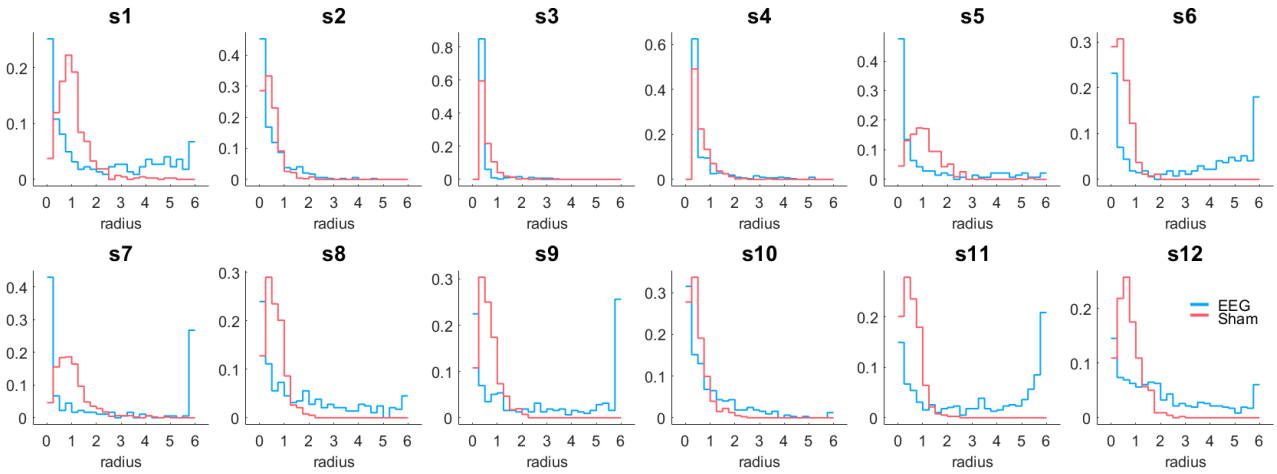


Figure 2.10 – Distribution of waypoint size in online sessions. Each curve is normalized to represent a probability distribution function.

can be seen, s5, who had a very low hit count and low points in EEG condition, had the radius mostly drifting between 0.15 and 0.5. The range was obviously too small compared to the sham condition which mostly centered around 1, and resulted in many misses during the recording. Conversely, s11 had the radius mostly concentrated above 5 while the sham condition was around 0.7, this resulted in a high hit count and low points. For s3 who had higher points in EEG than the sham condition, the curve of EEG condition is well centered around 0.5 and even sharper than the sham condition. This probably suggests that the EEG condition better reflected the demand of s3 than the sham condition. This can happen as there is no guarantee that a behavioral decoder is the best, especially, the sham condition was intentionally designed to slightly deviate from a perfect behavioral decoder. Another explanation can be that the cognitive state was really able to oscillate around the sweet spot, for which Faller *et al.* had also shown improved task performance in the case of down-regulating from a high arousal state [FCSS19]. It is worth noting that s4 also had very similar curves in both conditions as s3.

The EEG condition was however not better than the sham condition. The reason is not clear, but it probably relates to the radius when getting close to a waypoint. The curves here do not give such information.

2.5 Discussion

The open-loop experiment showed that the designed protocol was able to induce different levels of workload and perceived difficulty for the two extreme radii in the first session. The designed shared control was however not able to influence neither the levels nor the task performance. Therefore, it was not further studied as it is likely to consume much time to optimize. In the second session, the Medium condition did not differ too much in hit rates from the Easy condition, but the observation during recording was that the subjects were more engaged than in the Easy condition. However, without a proper evaluation of the cognitive state, it was not clear how to define the class of Medium condition. The current radius was subjectively decided by the operator. A better solution would be finding a method to standardize the radii for each subject such that similar subjective assessments are expected.

The power spectrum correlated to the Easy and Hard conditions in both sessions. The correlates in θ and α bands are coherent to some previous reports. The increasing of power in θ band was reported when subjects were tracking displacement errors [Coh16, SSJS16]. This is consistent with the protocol used here. In the case of Hard condition, the subjects needed to put more effort into pointing the purple cross toward the center of waypoint which is similar to tracking. The increasing of power in α band in the Easy condition is consistent with the case that the subjects were performing easy tasks or was relaxing, such as the well-known α peaks when eyes closed [Kli99, APGvG10]. However, for the used protocol, the α band has the highest difference in the sensory-motor region which raises the concern of decoding motor activities rather than cognitive states. It is well-known that the sensory-motor cortex maps to the contra-lateral side of the body [PB37], and therefore, neural correlates on the contra-lateral side imply that the correlates are hard to disentangle from the motor-sensory activity. Given that the subjects in this study were using their right hands which are ipsilateral to C4 and CP4 that have the highest difference in the power spectrum, the neural correlates are more likely to be cognitive states. Ipsilateral and neural, not limited to EEG, correlates on the motor cortex were also reported for difficult conditions in different motor tasks, from complex finger-tapping [VDA$^+$05], quickly moving a cursor towards a fixed zone [BRS$^+$14], to tracking a moving target [Per18].

The observed modulations of powers are, however, not always coherent with the literature. For example, the engagement index is known to be $\beta/(\alpha+\theta)$. The higher the β, the more engaged which should correspond to the hard condition. However, Figure 2.6 shows that high power in β band correspond to the easy condition. Additionally, the denominator suggests that the powers in α and θ bands are better to co-vary for consistency, while Figure 2.6 shows that both bands behave in opposite. One explanation for the incoherences is that the subjects

were not really more engaged in this study. Another explanation is that the literature has some limitations. For example, the index can work with previous protocols but not in visuomotor-based protocols. Another situation can be that the nonlinear combination destroys intuitive observations and really needs to be computed. Given there are so many possibilities, the best approach would be mining data from the same task and the same subject to build a subject-specific decoder with robust data-driven approaches.

On the other hand, the closed-loop experiment aimed at how well an EEG decoder can approach the sham decoder. The behavioral result was mostly expected. Exceptionally, s3 largely outperformed the sham condition with an EEG decoder. It is however needed to admit that the distributions of radius in EEG condition were mostly different from the sham condition, and some are mostly biased to either the Hard or Easy condition, even though two subjective calibrations were performed during the EEG conditions. Indeed, the offline decoding accuracies were not really high and likely caused the situation. Therefore, the decoders need further improvement in later experiments.

The online interaction also needs to be refined. A common opinion among most subjects was that the current interaction made the task more difficult. For example, a radius that is considered as easy in a non-changing scenario becomes difficult if it is changing all the time. The issue was that the continuously changing radius made the subjects had to pay extra attention most of the time. Without knowing the changing direction before seeing, the subjects were likely not able to properly decide their steering direction. Most subjects preferred or did not always aim at the center. The aiming was not easy for most subjects. As a result, a better strategy for them somehow was optimizing the flying trajectory together with the next waypoint. Otherwise, they might fall into a situation that the facing direction is difficult for hitting the following few waypoints. Also, a continuously changing radius made the evaluation of decoding accuracy impossible. Otherwise, the subject needed to report their perceived difficulty level or preferred changing direction all the time. This can be particularly an issue especially around the sweet spot, because the behavioral latency is unknown and may vary from time to time. Another crucial reason is that asking a subject to continuously provide feedback of ground truth is an extra and heavy mental task. Likely, the subject would not be able to provide reliable feedback, neither to properly perform the main task.

2.6 Conclusion

Although the closed-loop experiment did not yield a promising result on average, the concept of online interaction based on EEG decoder is proven, given the case of s3 was unlikely to be by chance. Some issues of the online interaction were identified and should be considered in the next experiment. It is also confirmed that reporting perceived difficulty level is strongly consistent with the workload level in the designed task, and time can be saved in future experiments by neglecting NASA-TLX as it has many more variables to be reported. However, before proceeding to the next closed-loop experiment, optimizing the EEG decoders is necessary.

3 Decoder Optimization

As pointed out in the previous chapter, decoding accuracy is one factor that needs to be further improved. This Chapter is dedicated to settling a general framework for building decoders of difficulty-related cognitive states. Based on the review made in Chapter 1.4, I will select components to build several decoders that can meet real-time constraints (low latency and causal) in online experiments, and validate which one is the best based on the offline sessions of Chapter 2. The final framework is presented in Chapter 3.4.

3.1 Selection of Components

3.1.1 Pre-processing

My main purpose in pre-processing is automatically reducing as many artifacts as possible. Therefore, spectral and spatial filtering techniques are considered for their typical adoption for reducing artifacts in EEG studies. The idea is to band-pass the signals in (1) spectral domain to avoid high-frequency power with low SNR and too low frequency to avoid signal drift caused by sweating or movement-related slow cortical potential [GCM11], and (2) spatial domain to avoid high frequency which is likely to be artifacts [GEF+15] while low frequency is for increasing the common-mode rejection ratio due to the amplifier. On the other hand, those data-driven filters directly aim at improving classification accuracy can utilize artifacts that may be persistent with a specific condition, and therefore, those filters are not considered.

As ocular activities are likely to contaminate EEG signals, removal of ocular artifacts is also considered, where ICA is one of the automatic approaches. At last, some rules of thumbs can automatically identify problematic EEG signals at some specific time and eventually discarding them [DSBMP15]. For example, a 100 μV amplitude is unlikely to be a normal neural activity.

3.1.2 Feature Engineering

The purpose here is not developing new types of features, but instead, comparing existing and well-known features for their performance. The selected features are power spectrum from all the selected EEG electrodes, the engagement index which had been utilized in other studies [PBB95, CRBP11, SM12, AG13], and the attention index given that attention can be a useful cognitive state [PVAG+14]. Both indices are to be evaluated in all electrodes for a comprehensive evaluation. Especially, in the case of the engagement index, the used electrodes were not consistent in those reports.

3.1.3 State Inference

The purpose of this part is to explore useful statistical machines for inferring cognitive states. Based on Lotte *et al.* [LCL+07], which was the available review article by the time of investigation, SVM outperforms several other different classifiers. As a result, SVM is being included in the comparison with different kernels. LDA is also included for its frequent employment in motor imagery studies. Regression-based approaches also yielded nice decoding results in relevant studies [ABD+16a, SSJS16], and therefore, are considered. For the ease of implementation, a regressor with elastic net regularization is chosen in the study while it is capable to compromise between the sparsity and Tikhonov regularizations.

3.1.4 Post-processing

The main purpose of post-processing is to study the effect of different numbers of samples. As a result, moving average with the past samples is adopted for the sake of simplicity.[1]

3.2 Comparison of Methods

The comparison is separated into controlled variables and independent variables to avoid exhaustive enumeration. The controlled variables are mainly in the pre-processing step such that they do not change during the validation, while the independent variables are the methods to be compared. The validation method is introduced after the two variables.

3.2.1 Controlled Variables

EEG and EOG signals were downsampled from 2,048 Hz to 256 Hz, to save computational time. The signals were then band-passed between 1 and 40 Hz by a 14^{th} order Butterworth filter with backward and forward processing.[2]

[1] If desired, one can design an Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filter to match the desired frequency response that should depend on the actual need of the application.

[2] This Chapter only involves offline validation. Therefore, backward and forward processing is allowed and minimizes the potential delay. However, only forward processing is used during online decoding in other chapters.

For EEG signals, out of the 24 recordings, electrode C1 was removed twice from the offline analysis due to abnormal patterns by visual inspections. A 20^{th} order spatial low-pass filter, SPHARA [GEF+15], was applied to interpolate signals for the removed electrodes and more importantly to reduce high spatial frequency components, likely corresponding to artifacts. Peripheral electrodes were left out of the analysis to reduce the likelihood of muscular contamination, yielding twenty-five channels centered at Cz, as shown in Figure 2.4, 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. CAR was then applied [BPP85].

The vertical EOG component was computed by subtracting the signal from the sensor between eyebrows by the average of signals from the other two sensors. The horizontal EOG component was derived from the bipolar signal between the two sensors close to canthi. Both components were used as the reference for removing ocular artifacts.[3]

Potential EEG artifacts due to ocular movements were alleviated by ICA. Specifically, RUNICA is performed [MBL+00], starting with 15 components, and iteratively searching for the maximum number of components with a proper solution. A proper solution is found when the returned weight matrix has no imaginary number and that the maximum and minimum values in the weight matrix were similar, for which 5 was picked as a threshold. If any of the final independent components had a correlation higher than 0.7 with the vertical or horizontal EOG component, the independent component was then dropped out from the future analysis. The reconstructed EEG signals therefore did not contain those independent components. On average, first and second sessions yielded 15.9 and 15.6 components, respectively. The numbers of removed components were 0.25 and 0.58 for the first and second sessions.

Multi-tapper algorithms were chosen for computing the power spectrum of each electrode, given that more reliable signal power can be estimated [Tho82, Tre95]. Thomson's multitaper algorithm was used to compute Power Spectral Density (PSD) with the time-half bandwidth product being two, and $10\log_{10}$ was used to compute the log-PSD features. The features were extracted over a two-second sliding window, with a 500ms shift for evaluating the decoder. This resulted in a 0.5 Hz resolution, and only the power bands between 2 and 28 Hz were extracted, in order to avoid the movement-related slow cortical potentials [GCM11] and a too low SNR on high frequencies. If a window had one or more EEG time sample with a peak value larger than 50μV after the previous pre-processing, the window was rejected from the validation.

3.2.2 Independent Variables

The features to be compared are (1) log-PSD with 0.5 Hz resolution, and (2) the attention and engagement indices which are defined in Chapter 1.4. Apart from the features, several decoding models were evaluated. All the evaluated combinations are listed below.

[3] During online decoding in other chapters, the EOG signals are not needed because the employed ICA only estimates the model in offline analysis.

Indices + LDA Engagement and attention indices are extracted which leads to 50 features. These features are the input of an LDA. The predicted labels of LDA are averaged with the previous 11 labels, and then another LDA is applied for the final mapping. The LDA was implemented by assuming equal distributions among the targeted classes.

Indices + Fisher + LDA Similar to Indices + LDA. The only difference is using Fisher score to preserve top 20% of discriminative features.

PSD + Fisher + LDA Fisher score is used to preserve the top 20% of discriminative log-PSD features, resulting in 140 features. The selected features are forwarded to LDA. The predicted discrete labels from LDA are averaged together with the previous 11 samples, and then another LDA is applied to find the threshold(s) for mapping onto the targeted discrete labels. The LDA was implemented by assuming equal distributions among the targeted classes.

PSD + Fisher + L-SVM Fisher score is used to hold the top 20% of discriminative log-PSD features. These features are given to SVM with a linear kernel [CL11]. The searched values of hyperparameter C (cost) are [2^{-3}, 2^{-2}, 2^{-1}, 1, 2^{1}, 2^{2}]. The predicted labels are averaged with previous 11 samples, and an LDA is employed to find the threshold(s) for mapping onto the targeted discrete labels. The LDA was implemented by assuming equal distributions among the targeted classes.

PSD + Fisher + R-SVM Almost the same as PSD + Fisher + L-SVM, but SVM comes with a radial basis function kernel [CL11].

PSD + Fisher + S-SVM Similar as the above, but using a sigmoid kernel [CL11].

PSD + GLM The log-PSD features are given as input of GLM-based regression. The assumed distribution is binomial with logistic regression as the link function. The regression is regularized with elastic net, where the regularization parameters are scanned with α = [0.15, 0.5, 1] (1 = LASSO, 0 = ridge regression)), and nonnegative λ = [0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20] for the penalty term. The target response in the first session is the self-assessed difficulty levels which are divided by 100. This normalizes the values to 0 and 1 in order to meet the requirement of a binomial distribution. In the case of the second session where difficulty levels are not accessible, the easy, medium, and hard conditions are assigned as 0, 0.5. and 1, respectively. The output of regression is further averaged with the previous 11 samples of the regressor. In the end, an LDA is applied to find the threshold(s) for mapping onto the targeted discrete labels. The LDA was implemented by assuming equal distributions among the targeted classes.

The effect of different lengths in post-processing is examined after deciding the best model from the above-listed methods. The investigation uses samples spanning from 0 s (without post-processing), 2 s, 4 s, 6 s, 8 s, to 10 s before.

Table 3.1 – Validation settings of session 1 and session 2

	Session 1	Session 2	Chance level
Set 1	Easy vs. Hard+SC	Easy vs. Hard	0.5
Set 2	Easy vs. Hard	Easy vs. Medium	0.5
Set 3	Hard+SC vs. Hard	Medium vs. Hard	0.5
Set 4	Easy vs. Hard and Hard+SC	Easy vs. Medium and Hard	0.5
Set 5	Easy and Hard+SC vs. Hard	Easy and Medium vs. Hard	0.5
Set 6	Easy vs. Hard+SC vs. Hard	Easy vs. Medium vs. Hard	0.33
Cross-validation	Leave-one-trajectory out	Leave-one-trajectory out	
Metric	Accuracy	Class-balanced accuracy	

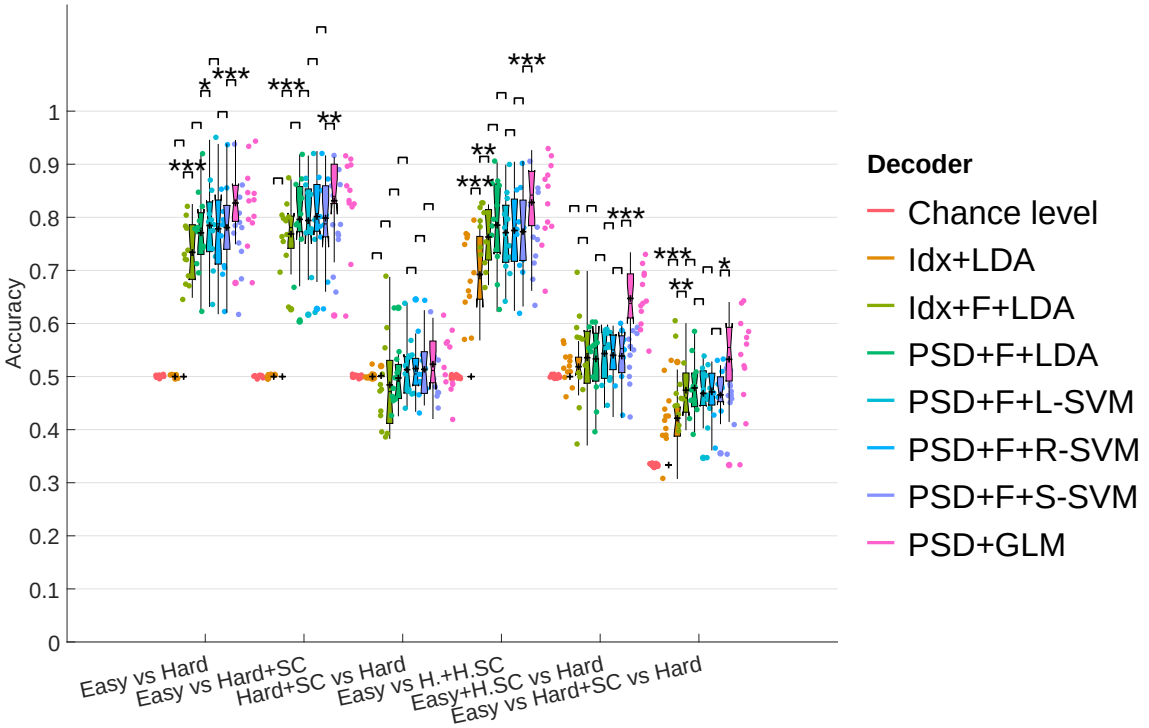
3.2.3 Validation

The offline sessions from Chapter 2 are used for assessing the performance with the same validation strategy as described in Chapter 2.3.3. Table 3.1 summarizes the validation strategy. Leave-one-trajectory cross-validation is used for both sessions, where one trajectory is held as a test set while the other as a training set. It is important not to partition samples of the same trajectory into both the training and test sets. Otherwise, the high temporal correlation of physiological signals will yield optimistic result [Mil04, VRE+17]. Although hyper-parameters are involved, nested cross-validation is not considered here for the sake of computational time. The performance metric is the class-balanced accuracy for not biasing any class in case of imbalanced samples, where the metric in the first session becomes accuracy because there is only one condition in the test set. Apart from only validating Easy versus Hard, the Medium and Hard + Shared Control (SC) conditions are also included for a comprehensive investigation. As a result, each session has six sets of classification problems for all the two-class and three-class cases that are grouped according to the difficulty levels. For example, Easy and Hard versus Medium is not considered.

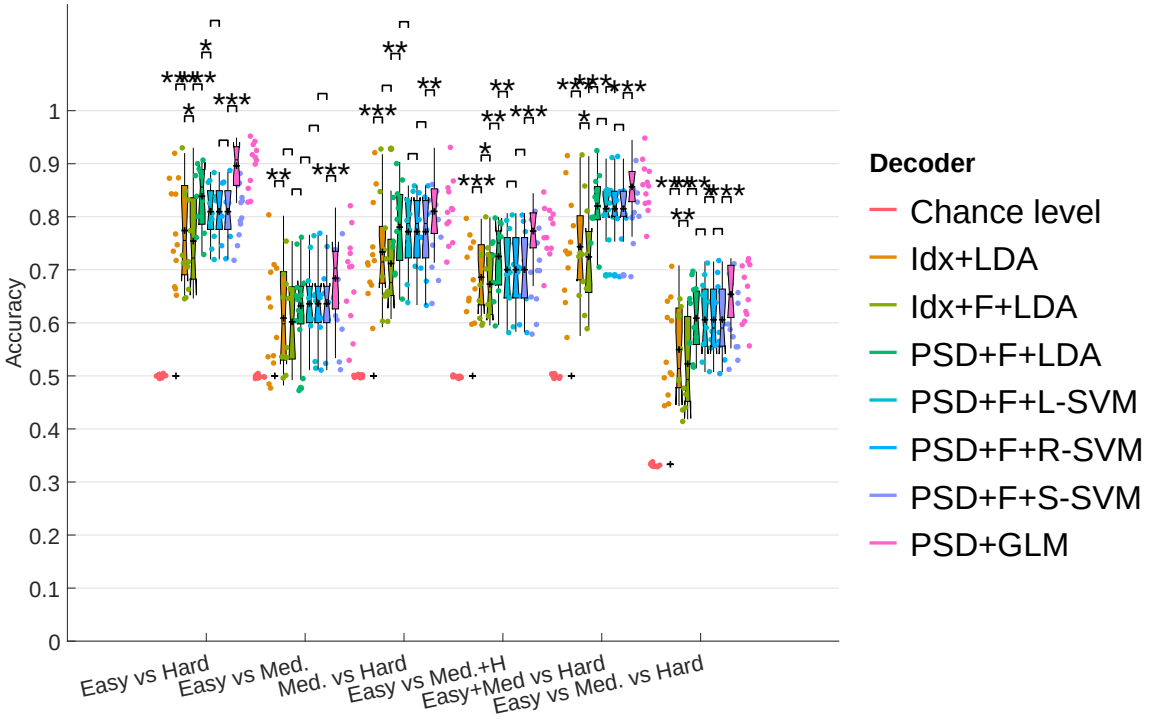
3.3 Result

3.3.1 Features and Decoders

Figure 3.1(a) and Figure 3.1(b) present the validation result of different classifiers for session 1 and session 2, respectively. Each sub-figure has six groups of boxes, where one group represents one specific set of targeted condition as indicated on the x-axis. For example, the first group is Easy vs. Hard while the last group is Easy vs. Hard+SC (Med) vs. Hard for the first (second) session. Each box represents quartiles of the (class-balanced) accuracy during validation from all the subjects, where one data point corresponds to one subject. The average of all subjects is shown by a cross. T-tests ($n = 12$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$) were conducted between any two adjacent methods in Figure 3.1(a) and Figure 3.1(b). The Hard+SC vs. Hard basically yielded chance-level accuracies for all the methods. This is



(a) Session 1



(b) Session 2

Figure 3.1 – Decoding accuracies using different decoders. T-tests were applied on adjacent methods ($n = 12$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). The crosses indicate mean values. Horizontal lines in the boxes are median values.

expected as the subjective assessments between the two were nearly identical.

Using the two indices without feature selection evidently performed much worse than other methods, as it sometimes yields chance-level accuracies. Even using the Fisher score, the result can easily be worse than purely using log-PSD. Among the rest of the evaluated methods, it appears that using either LDA or SVM, the decoding accuracies do not differ much in the first session. On the other hand, using SVM seems better while the chosen kernel does not make much difference. The best method all the time is the PSD+GLM.

3.3.2 Length of Post-processing Window

Taking PSD + GLM, Figure 3.2(a) and Figure 3.2(b) illustrate the effects of using different window lengths in post-processing for the first and second sessions, respectively. The legends are the same as Figure 3.1. T-tests ($n = 12$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$) were conducted between any two investigated lengths that are adjacent.

In most cases of either session, decoding accuracy significantly increased when the length is extended. One exception is the Hard + SC vs. Hard in the first session. This is not surprising, as both conditions did not induce sufficiently different cognitive states. As a result, close-to-chance-level accuracies are expected regardless of the length.

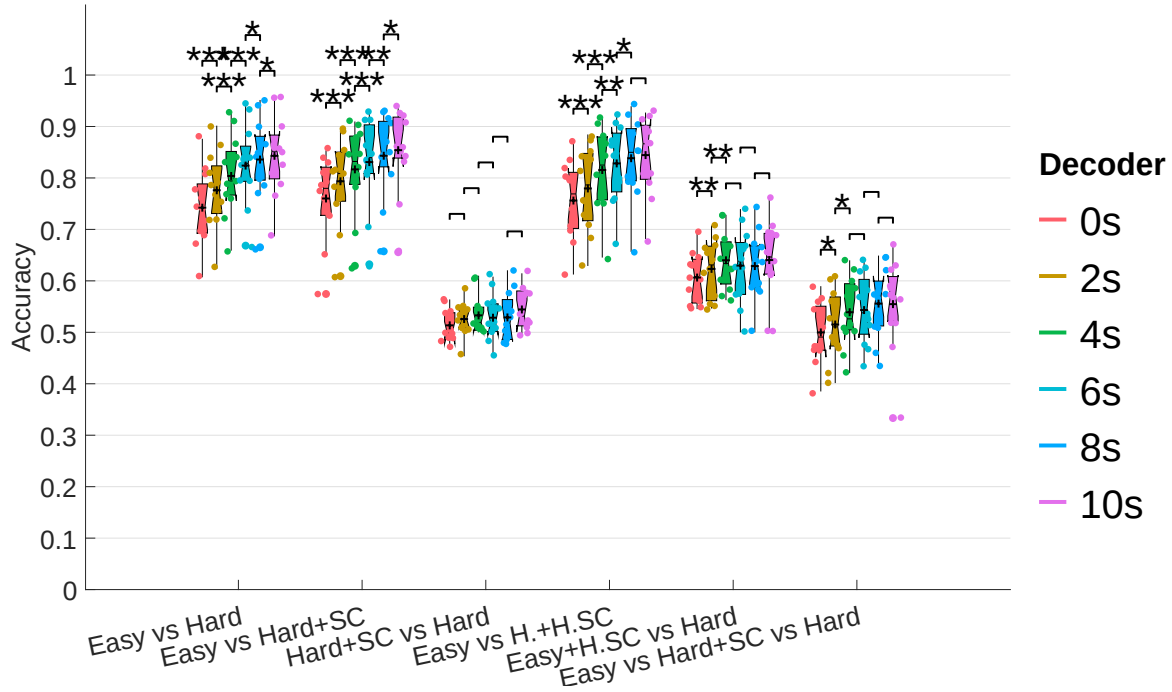
Significance can still be found between using 8- and 10-second windows in many cases. This suggests that there is still space to improve, although the increasing trends seem to reach their margins. It can be seen that post-processing is a very effective method to boost accuracy. However, when low-latency feedback is important, one should carefully make a trade-off. In the used dataset, a window of 6 seconds seems to be a good trade-off.

3.4 Final Method

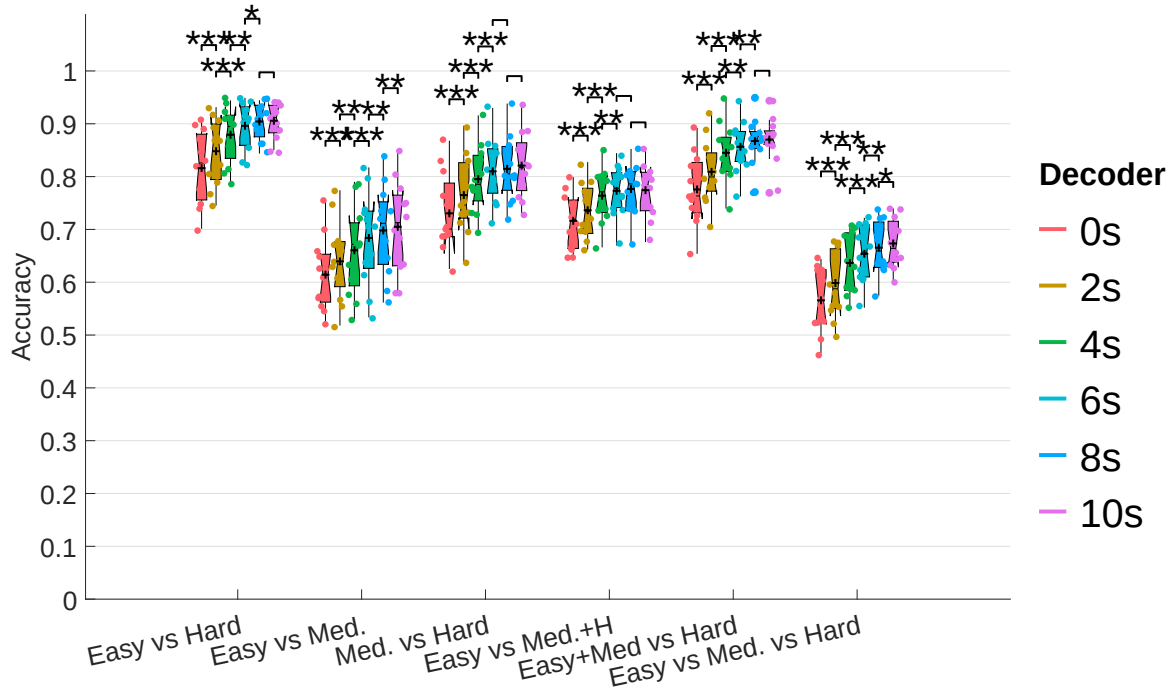
Based on the above results, the final signal processing architecture is adopted as below for the following chapters.

Pre-processing The details of pre-processing are the same as described in Chapter 3.2.1. A quick summary is that EEG and EOG signals are downsampled and then band-passed between 1 and 40 Hz. EEG is further visually inspected to remove contaminated electrodes, and then spatially filtered (and interpolated if necessary) by SPHARA. Then filtered again on the remaining twenty-five electrodes by CAR. EOG were spatially decomposed into vertical and horizontal components, and were used as reference signals for removing ocular artifacts. This is done by an iterative ICA method. A sliding window is used to extract log-PSD features by a multi-tapper algorithm, and each window is examined for potential artifacts.

Cognitive State Inference The method being adopted is the PSD + GLM, where the hyperpa-



(a) Session 1



(b) Session 2

Figure 3.2 – Decoding accuracies using different lengths of post-processing. T-tests were applied on adjacent window lengths ($n = 12$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). The crosses indicate mean values. Horizontal lines in the boxes are median values.

rameters are to be selected by a cross-validation. The window length of post-processing is 6 seconds unless specified.

3.5 Conclusion

Different components of decoders were compared. The pre-processing becomes more solid than Chapter 2 with less EOG artifacts. Different decoding approaches were also investigated, where the GLM-based method using log-PSD features is the best among the compared methods. On the other hand, integrating the current decoder output with previous outputs is to be very effective in improving decoding accuracy by trading off the latency. According to the application, this trade-off needs to be further evaluated. In short, as the decoder has shown a much higher accuracy, it is worth testing on another closed-loop experiment.

4 One-directional Online Regulation

4.1 Introduction

This chapter focuses on an online interaction with a one-directional regulation. Although bi-directional regulation has the possibility to converge to the sweet spot regardless of the current state, the unpredictable regulating direction actually increased the difficulty level in Chapter 2. To alleviate from this issue, this chapter considers one-directional regulation with a learning-like setting.

A learning process is usually progressive; the learners begin at a certain easy level which is increased after becoming more proficient. Progressive learning is a reasonable design according to the theory of challenge point [GL04]. Both the skill of learner and the task difficulty decide the optimal challenge point which has the highest potential learning benefit. As shown in the left curve of Figure 4.1, if the curve of potential learning benefit is drawn against different functional task difficulty level, the curve will look like an inverted U which is of the same shape as Yerkes-Dodson curve [YD08].

The functional task difficulty is a unified way to describe the relationships between the task difficulty level and potential learning benefit as well as task performance. The functional task difficulty level is different from an objective difficulty level. The functional difficulty level is

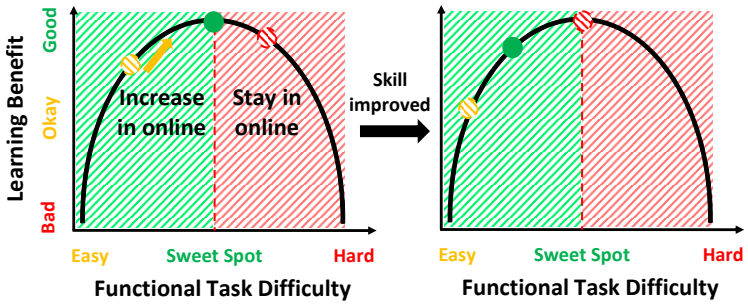


Figure 4.1 – Principle of one-directional regulation.

subjective which means that the difficulty is normalized to the skill of learner. For example, an objectively difficult level can be subjectively easy for experts. This is consistent with the point made in Chapter 2.5; a better solution is to standardize the difficulty level for each subject. As can be seen from the left curve in Figure 4.1, there are three states (circles in yellow, green, and red). Once the skill of learner is improved, as shown in the right curve, the three states will be shifted toward the left side of the curve if their objective difficulty levels are re-mapped with the new skill level. As a result, the protocol this time will introduce a standard to define functional task difficulty, where objective difficulty levels are normalized based on the task performance. In this case, each subject should have a similar functional (subjective) task difficulty level compared to other subjects.

Another advantage of considering progressive learning is solving the issue of continuously changing the difficulty level. Normally, a level in a learning process should remain the same for a certain amount of time for training. On the one hand, this will make the learner easier to adapt to the current level instead of adjusting all the time. On the other hand, it also allows the learner to provide simple feedback about the current level which permits evaluation of online decoding accuracy.

As a result, the principle of interaction is summarized in Figure 4.1. The designed protocol will only increase the objective difficulty level when the current level is considered as easy, either by an EEG decoder or the learner itself. If the level is not easy, the objective difficulty level will remain the same. The level increases from the easy (left) side of the learning curve, and is therefore different from another similar work regulating arousal states from the right side of Yerkes-Dodson curve [FCSS19]. Plus, the learning curve is not necessarily equivalent to the Yerkes-Dodson curve. In this chapter, even if the difficulty is higher than the sweet spot, the skill should be improved after a certain time and eventually be shifted toward the sweet spot without regulating the objective difficulty level.

The goal of this chapter is to compare the behavioral outcome between using an EEG decoder and self-paced decisions on when to increase the difficulty level. Andrieux *et al. al* had shown on a motor learning task that self-paced decisions of task difficulty level have a better task performance than a control group. The control group was yoked which means that the subjects were informed of having random difficulty level but instead having the same sequence of difficulty level from the group of self-paced decisions [ADT12]. One question is whether the use of an EEG decoder would yield the same outcomes as the self-paced decision process. A related question is that how does skill improve in different conditions. The rest of this chapter investigates these questions.

4.2 Materials

4.2.1 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 a normal or corrected-to-normal vision and reported no history of motor or neurological disease, except one subject ever experienced vasovagal syncope but not during this study.

4.2.2 Recording Setup

A Biosemi ActiveTwo amplifier was used to record EEG and EOG signals at 2,048 Hz. Sixty-four EEG electrodes were placed according to the international 10-10 standard. Three additional channels for EOG were placed in the same positions as in Figure 2.1(a).

Subjects sat comfortably in front of a twenty-inch screen showing the protocol with a 1680x1050 resolution. They hold a joystick, the same as in Chapter 2, to provide inputs to the protocol. One female subject was left-handed but reported being comfortable using a right-handed joystick.

4.2.3 Task

The task is similar to the one described as in Chapter 2. The subjects were instructed to steer the simulated drone to fly through a series of waypoints. The drone flew at a constant velocity with roll and pitch as the only allowed maneuvers. As shown in Figure 2.1(b), 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 subjects were forced to proceed to the next waypoint. If it is a hit of difficulty level k (from 1 to 16), k points were rewarded. 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 waypoint, and the elapsed time.

In the case of online sessions (see Section 4.2.5), the radius (difficulty level) could decrease (increase) or stay the same after every other four waypoints, forming a decision group. 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 steering, the numbers of points and hits were hidden to avoid over-thinking on optimizing points. On the other hand, feedback on difficulty level was given to the subject in the middle red number. The current level and radius were displayed, and a ☺ symbol appears if the current level is indicated as easy by the subject.

4.2.4 Personalizing Difficulty Levels

As discussed, personalizing the radius of each level is necessary, in order to transform into the same scale of functional task difficulty levels and to properly compare the behavioral data in the online sessions. The personalization is based on a skill evaluation process which is illustrated in Figure 4.2(a). The top of Figure 4.2(a) shows the collection of data, and then, the lower part indicates that personalization is done by a sigmoid regression.

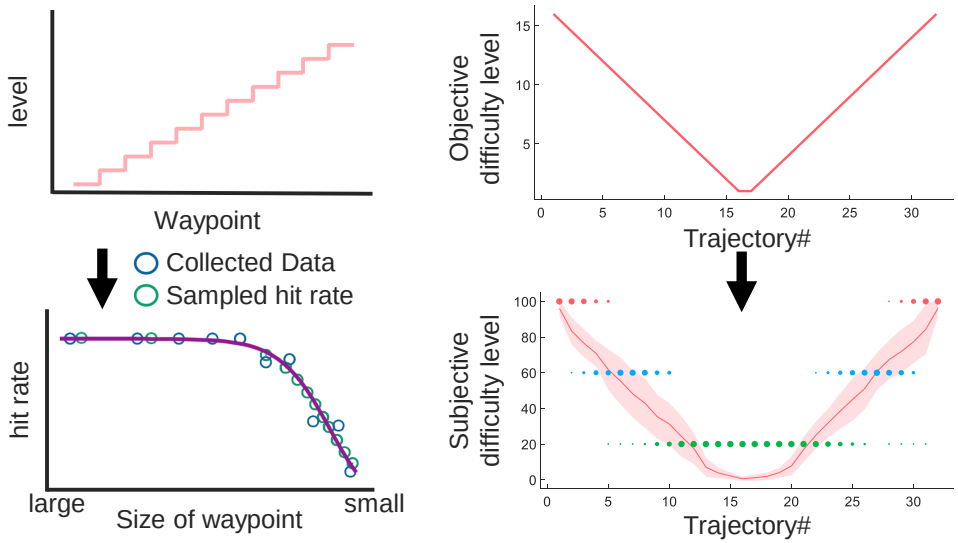
For the data collection, the subject needed to navigate the drone through a pre-defined trajectory of eleven levels (radii), each level has eleven waypoints. The hit rate of each level was computed. Each hit rate corresponds to a blue data point in the bottom of Figure 4.2(a). Then a sigmoid regression was performed with the x-axis being the radius of waypoint and the y-axis being the hit rate. Thirteen radii were firstly sampled from the regression curve with 0%, 8.3%, ..., and 100% of hit rates. Three additional levels were included with the radii being 1.5, 2, and 2.5 times larger than the radius of 100% hit rate, in order to include some extremely easy levels. This resulted in defining sixteen levels in the recordings, and the sampled points are plotted as green dots in Figure 4.2(a).

The eleven levels used to personalize the sixteen levels in the offline recording were manually defined by the operator. While in the online sessions, the eleven levels were calculated from conducting another sigmoid regression based on the hit rates of the offline session. The offline session, as shown in Figure 4.2(b), has thirty-two trajectories, where each is associated with a difficulty level (waypoint size). As a result, the sigmoid regression for defining the eleven levels in the online sessions was computed from thirty-two data points, and the targeted eleven radii had the hit rates of 0%, 10%, ..., and 100%.

4.2.5 Experimental Protocol

Each subject participated in one offline and two online recording sessions that took place on different days within three months. The aim of the offline session was to collect 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 between EEG-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 simply placed 32 A.U. 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, respectively.



(a) Personalize difficulty

(b) Design of offline session

Figure 4.2 – Offline session is composed of (a) personalizing objective difficulty based on skill evaluation before the recording and (b) 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).

Offline

During the setup, each subject familiarized with the protocol by at least 10 minutes of steering with a pre-defined training trajectory of 121 waypoints (*c.f.* Chapter 4.2.4), each steering lasted about five minutes. The subjects were asked to try their best for the last one, as the result would be used to personalize difficulty levels in the later experiment (as presented in Chapter 4.2.4).

Before the actual recording, each subject was recorded with eye closed and opened, each for one minute. They were used as an EEG baseline for calibration (to be addressed in Chapter 4.3.1). Then, each subject navigated through thirty-two trajectories, each having only one difficulty level. As depicted in the upper figure of Figure 4.2(b), the level descended from level 16 to level 1, and then ascended from 1 to 16. Each trajectory was composed of thirty-two waypoints and lasted around 90 seconds.

As shown in the lower figure of Figure 4.2(b), the subjective difficulty level was reported at the end of each trajectory. The subjects were asked to give a numeric level between 0 and 100 for the assessment of perceived difficulty level, and to declare whether the trajectory was easy (green dots), hard (blue dots), or extremely hard (red dots). The definition of easy was that the subject felt in good control of the drone, while the opposite for the other two. The extremely hard was defined as a level that the subject feels herself cannot manage the level in reasonable training time.

Online

During the setup, the subject firstly practiced for at least 10 minutes, and then steered the drone another time for personalizing the difficulty without recording signals. Eye-open and eye-close were then recorded for a preliminary calibration of the EEG decoder (see Chapter 4.3.1).

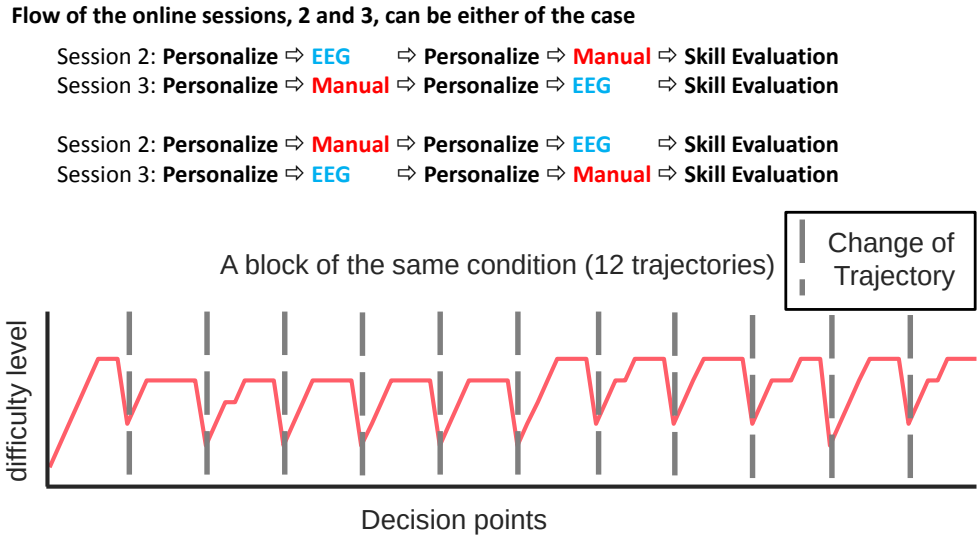
Figure 4.3(a) shows the flow of online sessions. Each session was composed of two personalization processes, one skill evaluation, one EEG block, and one Manual block. The personalization processes (same as Figure 4.2(a)) before EEG and Manual blocks ideally map the new 16 levels to the same functional task difficulty levels. These re-mappings contribute to a more reasonable comparison of the behavioral result of the two conditions. The first personalization was done without recording signals as described in the previous paragraph. The skill evaluation is the same task without further tuning the levels. In the later texts, when skill evaluation is mentioned, it will include the skill evaluation parts in the personalization processes unless specified.

As addressed in Figure 4.3(a), if a subject begins with the EEG condition in the first online session, Manual condition will become the first block in the second online session. The subjects were divided 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.

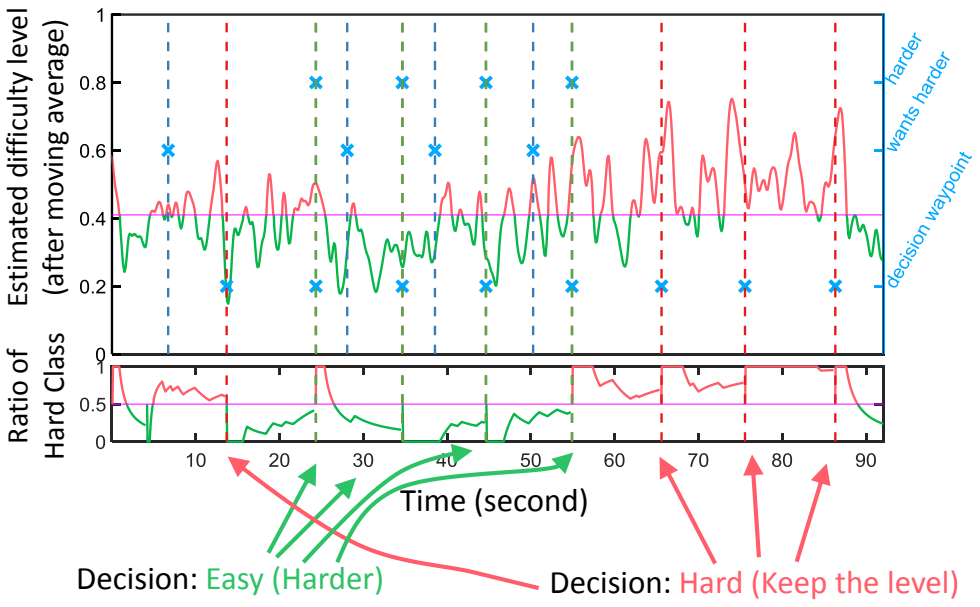
The main difference between EEG and Manual condition is the decision rule of increasing the level. In the EEG condition, the level was increased if the EEG decoder decides as Easy (see Section 4.3.3 for details of interaction and Figure 4.3(b) for an illustration). On the other hand, the Manual condition was based on whether the subject pressed the button before the decision point.

As addressed in the bottom of Figure 4.3(a), each condition was conducted in a block of twelve trajectories. The beginning level of 1^{st} trajectory was always level 1 while the next trajectory began with four levels lower than the final level of the previous one. For example, 2^{nd} trajectory began at level 3 if the final level of 1^{st} 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, with the exception of changing trajectories as indicated by the vertical lines.

Each trajectory consisted of 33 waypoints with eight decision points (4^{th}, 8^{th}, ..., and 32^{nd} waypoints). After the decision point, the level either increased or stayed the same. In the case of EEG condition, the operator checked the reliability of the decoder before the 1^{st} trajectory. This also happened before the 5^{th} and 9^{th} trajectories if the level was not moving between Easy and Hard. During the checking, adjustment of the bias term of regression (see Section 4.3) was based on the feedback of the subject, where the subject additionally steered one time or a few times of an extra and long trajectory. The adjustment did not require the subject to



(a) Design of online sessions



(b) Online interaction mechanism

Figure 4.3 – Online session. (a) Flow of online sessions, where the personalize and skill evaluation refers to the same task as in Figure 4.2(a). Each EEG or Manual 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.

finish the trajectory and the time was made as short as possible.

4.3 Decoding Perceived Difficulty

4.3.1 Signal Pre-processing

The signal processing is mostly the same as described in Chapter 3.4. Out of 39 recordings, electrode P2 was removed twice from the offline analysis or online decoding due to short-circuit with a reference electrode. On average, 15.8 components were returned by ICA and 1.07 components were then removed during the construction of online decoders.

The main difference is that the instantaneous log-PSD features were corrected by subtracting a baseline log-PSD vector computed during the baseline recording (eye-open and eye-close). The vector was computed by averaging the log-PSD features across time. In online sessions, another baseline recording was conducted in order to capture the non-stationarity of EEG signals across days. The features were then calibrated by subtracting the new baseline log-PSD vector.

4.3.2 Inference of Perceived Difficulty

The inference was done in the same way as addressed in Chapter 3.4. The reported difficulty level was divided by 100 to normalize between 0 and 1 for the regression. One difference is in the online sessions, where the bias term of the regression model from the offline session was shifted from the original value when the decoder cannot properly decode the cognitive states. The criterion of whether to shift the bias term is described at the end of Chapter 4.2.5. Although subjects were asked to describe the levels in three classes, the targeted classes were 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 sliding window.

4.3.3 Online Interaction

Figure 4.3(b) depicts the online interaction in a trajectory, where the blue asterisks are linked to the y-axis on the right and indicate the event of making decisions and ground truth given by the subjects. 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, only one decision at the protocol side was made at a decision waypoint (those vertical lines extended to the bottom panel). The decision 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 decision was to increase the level (harder). Otherwise, the level was kept. The counter of each class was reset to zero when a decision was made or at the beginning of a trajectory.

4.3.4 Performance Validation

Both the offline and online decoding performances were assessed by the class-balanced accuracy. The best α and λ 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×mean (across test sets) of class-balanced accuracy - 0.5×mean (across classes) of standard deviations (across test sets).

The offline validation strategy is four times of leave-one-pair-out cross-validation, where each test set holds a pair of trajectory. One trajectory was 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 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 the non-cross-validation case, the Easy trajectories were less or more frequently chosen than the Hard plus Extremely Hard trajectories.

In the online sessions, 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 ×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.

4.4 Analysis of Online Behavioral Data

4.4.1 Task Scores

One question of interest is whether subjects could achieve similar task performance in both conditions. As a result, two-tailed t-tests over task scores (n = 12) were conducted between both conditions for each subject. 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.

4.4.2 Skill Curves

Apart from the task scores, skill improvement is also worthy of examination. The sigmoid regression performed in Chapter 4.2.4 can represent a subject's skill curve. The skill curves are characterized by several parameters, the two important ones are x50 (the x-value has 50% of the range in y response) and slope. Generally speaking, a larger (smaller) x50 (slope) represents a worse skill. The comparison was conducted by a two-tailed 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.

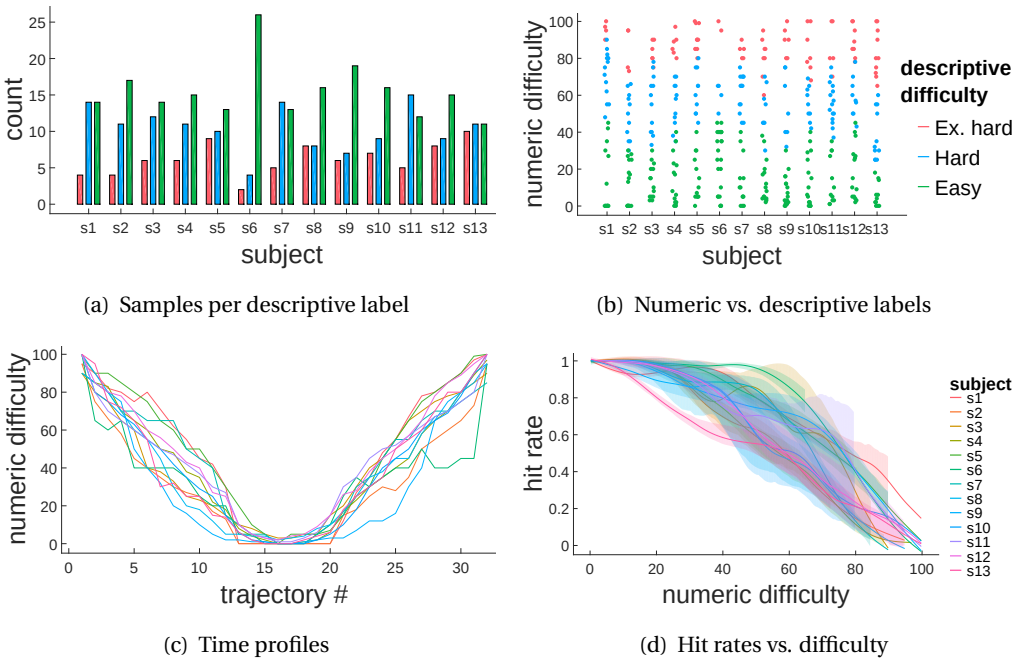
Apart from the x50 and slope, another index is the area beneath the skill curve. The area was computed by an integral over level 1 to 16 which leads to overall hit rates across all the levels. The larger the area is the better the skill. With the overall hit rates, their differences between both conditions 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.

4.4.3 Final Levels

Behaviorally speaking, if both conditions are similar in terms of the decision process, the curve of final levels vs. trajectories should also be similar among 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:

- Pearson's correlation: this gives the similarity of a co-varying trend between two patterns (vectors, $n = 12$). If desired, 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 and also tolerates any magnitude of shift.
- 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.
- 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.

Each index has its pros and cons. Therefore, having at least two indices with good values is convincing to conclude that both profiles are similar.



(a) Samples per descriptive label

(b) Numeric vs. descriptive labels

(c) Time profiles

(d) Hit rates vs. difficulty

Figure 4.4 – Subject's reported difficulty labels and hit rates in the offline recording.

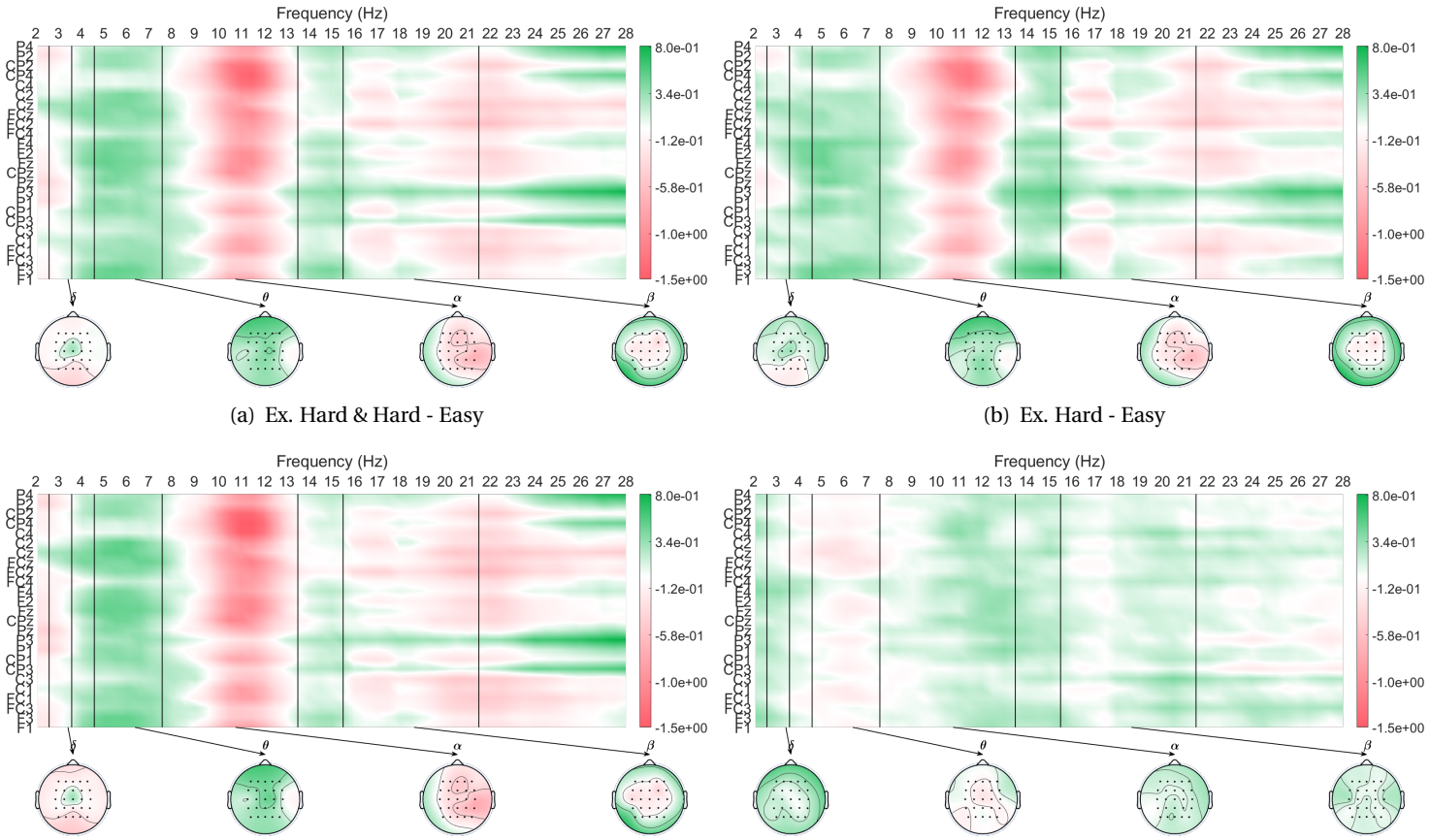
4.5 Results

4.5.1 Offline Behavioral Result

The plot of subjective difficulty levels in Figure 4.2(b) 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 could nicely induce subjective difficulty level in the designed V-shape.

Subject-wise results are reported in Figure 4.4. Figure 4.4(a) provides the amount of trajectories being categorized as Easy, Hard, or Extremely Hard. Six subjects had an amount of Easy trajectories similar to the sum of the other two labels. Subject s6 had a much higher number of Easy trajectories because the personalization was conducted without sufficient training time during setup. Although the distributions across subjects were not necessarily the same, the protocol generally induced sufficient samples for validating under the targeted binary-classification framework. Figure 4.4(b) further scatters the numerical and descriptive labels, some intra-subject inconsistency can be observed in s8, s10, and s11, where a few numerical levels were described as Hard while another time as Extremely Hard. Roughly speaking, a level below 45 is considered as Easy while above 70 is considered as Extremely Hard.

Figure 4.4(c) displays how the numeric levels evolved across trajectories, where one line stands



(a) Ex. Hard & Hard - Easy

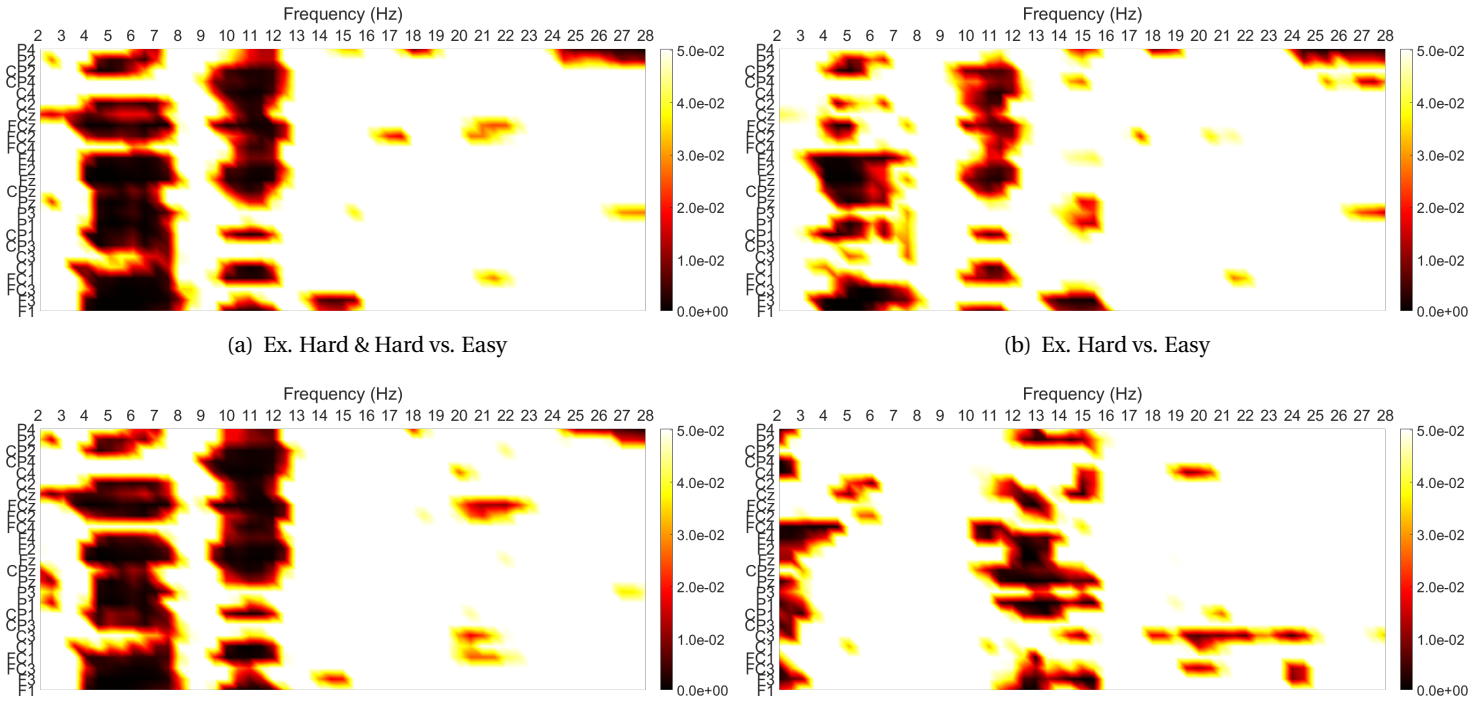
(b) Ex. Hard - Easy

(c) Hard - Easy

(d) Ex. Hard - Hard

Figure 4.5 – Differences between two grand averages of log-PSD. The grand average was first performed over windows, and then subjects. Red (Green) means that a lower value favors the Hard (Easy) condition. White means no difference. The blue line on the color bar indicate the extreme values on the data, otherwise, the extreme values of the data are the same as the limit of the color bar.

for a subject, similar to the grand average of Figure 4.2(b). Some smaller waypoint sizes were reported as easier than the previous larger one, but the trend in a larger scale shows that the difficulty levels changed as expected. Figure 4.4(d), on the other hand, illustrates the relationship between hit rates and reported numerical difficulty for each subject for the thirty-two trajectories. The data was spline-smoothed with the shaded areas as standard deviations [MO18]. The hit rate generally decreases when the numeric difficulty level increases. The trend is similar to the bottom of Figure 4.2(a), although far from perfect and the x-axes are not exactly the same.



(a) Ex. Hard & Hard vs. Easy

(b) Ex. Hard vs. Easy

(c) Hard vs. Easy

(d) Ex. Hard vs. Hard

Figure 4.6 – p-values of t-tests (n=13) using the data from Figure 4.5.

4.5.2 Power Spectrum Correlates

Figure 4.5 shows the difference of the grand averaged power spectrum between different sets of conditions. Each sub-figure has topoplots showing the differences in the four frequency bands as in Chapter 2.3.2, and one electrode-frequency plot illustrating the power for each frequency bin and electrode. White means there is no difference between the two conditions while the darker the red or green means the two conditions are easier to be distinguished.

Figure 4.5(a) plots the difference between the targeted binary set, [(Ex. Hard, Hard), Easy], in the online sessions, while Figure 4.5(b) and Figure 4.5(c) are [Ex. Hard, Easy] and [Hard, Easy], respectively. The difference between Ex. Hard and Hard are presented in Figure 4.5(d). It can be seen that the first three sub-figures are very similar among themselves and are also similar to Figure 2.6(a) and Figure 2.6(b). The major and consistent differences lie in the α band in the right hemisphere (C4 and CP4 in particular) and the θ band around Cz. On the contrary, in Figure 4.5(d), the two consistent differences seem to be reversed. This somehow implies that the trends in log-PSD is not exactly ordinal to the difficulty levels. However, the differences in Figure 4.5(d) are rather small compared to the other sub-figures. Therefore, the differences in Figure 4.5(d) might be relatively subtle and it is hard to conclude the finding at this moment.

Each sub-figure of Figure 4.6 plots the result of paired t-tests (n=13) between the two groups as indicated and are the same as the corresponding sub-figure in Figure 4.5. White means

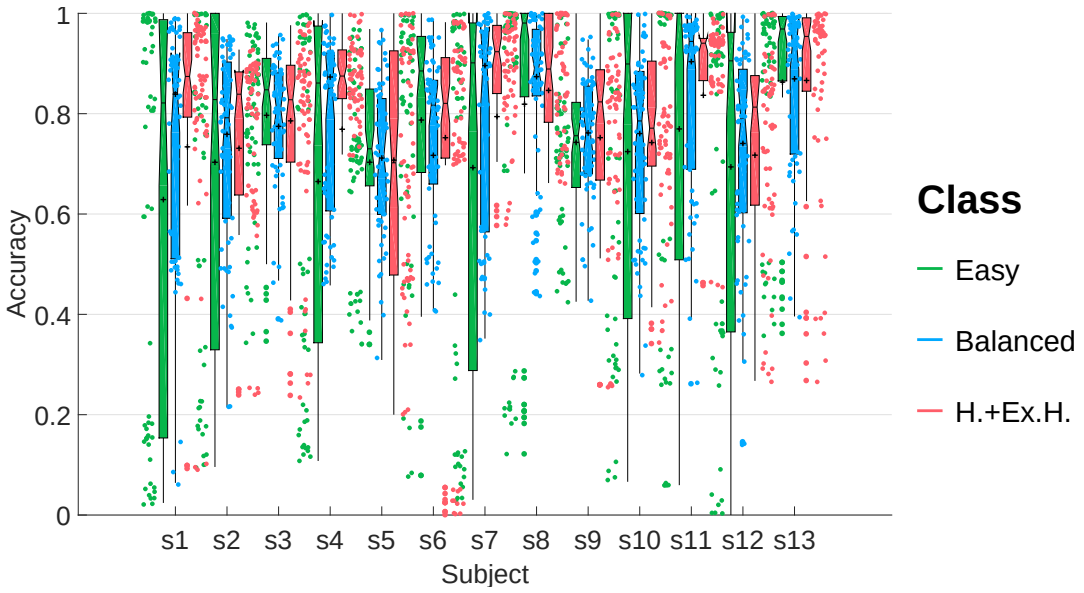


Figure 4.7 – Accuracy at window level in offline validation.

p-value is 0.05 or above. As can be seen from Figure 4.6(a), 4.6(b), and 4.6(c), the δ and θ bands at Cz, and α band around C4 and CP4 still yielded significant result as in Figure 2.7(c), while the β band around central-lateral region is less consistent.

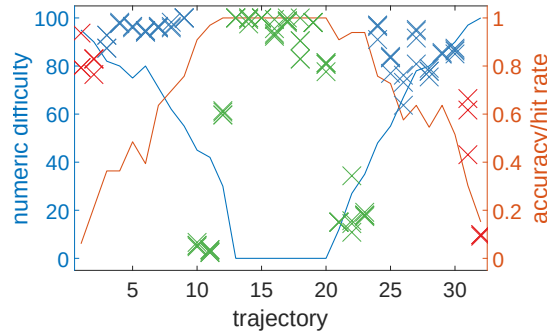
4.5.3 Offline Accuracy

For each subject, the best hyper-parameter set was selected. The mean class-balanced accuracy in offline validation across subjects was 76.7% with standard deviation being 5.1%. These values indicate plausible decoders are available for the online sessions.

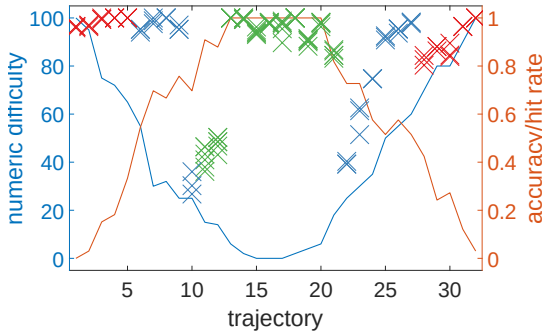
Figure 4.7 further depicts the per-class accuracy (green and red) and class-balanced accuracy (blue) at window level for each subject with her/his own best hyper-parameter. In the figure, the H(ard)+Ex.H(ard) class was grouped as one class according to the online interaction principle. The boxes provide the quartiles while the data points are test accuracies during the validation. It is easily observed that some test sets were often largely misclassified and appeared more frequently in the Easy class.

Figure 4.8 provides a clue to a possible explanation of this phenomenon. For the sake of illustration, Figure 4.8(a) and Figure 4.8(b) shows large plots for the first and last subjects, respectively, while Figure 4.8(c) shows plots for all the subjects. The red lines are hit rates and blue lines are the reported numerical difficulty levels. A cross stands for the accuracy of a tested trajectory and a color stands for a specific descriptive label. A trajectory may have multiple data points because of the validation strategy.

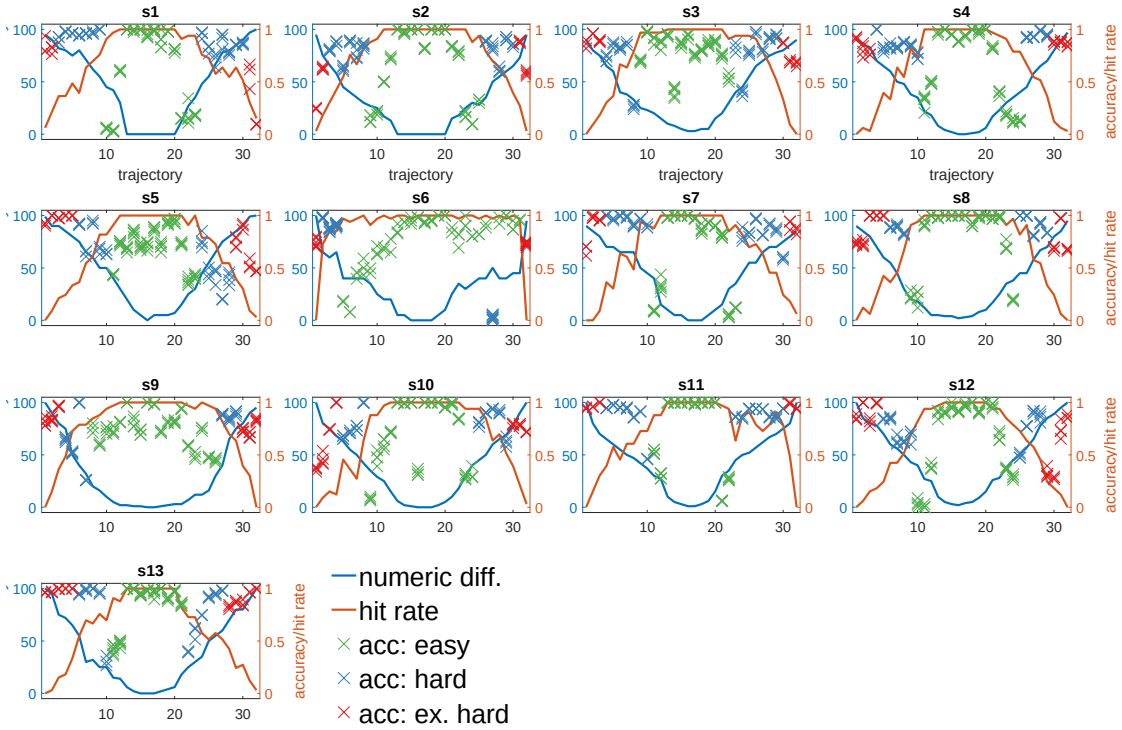
The finding is that the data points with low test accuracies are mainly located between the



(a) s1



(b) s13



(c) All subjects

Figure 4.8 – Accuracy at window level in offline validation per trajectory for each subject. Although there are three labels, please notice that the classification was done between Easy vs. (Hard + Ex. Hard).

transition trajectories, when the descriptive labels are switching from the Easy (green) to Hard (blue) with a few exceptions. Taking s1 as an example, the accuracies at around 10^{th} and 23^{rd} trajectories suddenly drop. It is easily observed that these trajectories have a numeric difficulty level around the boundary between Easy (green) and Hard (blue). Similarly, s13 has the same situation at around 11^{th} and 22^{nd} trajectories. It has to be remembered that the classification was done between Easy vs. (Hard + Ex. Hard). Therefore, the transitions between Hard (blue) and Ex. Hard (red) was mostly not affected as in s13. One exception can be found is around the 32^{rd} trajectory of s1, where a potential explanation is that s1 was tired and disengaged, given that it was extremely hard and the long recording was about to finish. As illustrated in Figure 4.8(c), similar trends can be identified from all the subjects. Therefore, I believe that the regression properly captured the near-ordinal trends in the power spectrum, but the responses of regression at transitional cases were too similar. As a result, the one-dimensional LDA was not able to perfectly separate them, but instead, yielded a threshold with the best class-balanced accuracy.

4.5.4 Online – Decoding Accuracy

EEG condition

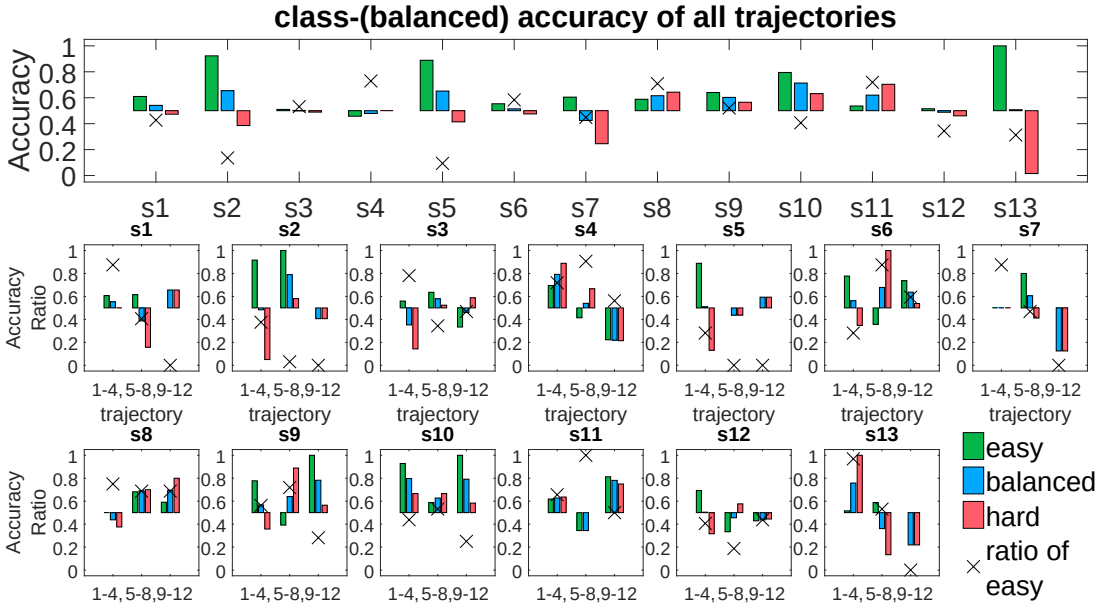
Figure 4.9 illustrates the online decoding accuracy of the EEG condition for each session and subject. Blue bars show the class-balanced accuracy while green and red bars are class-accuracy for Easy and Hard, respectively. The 'x' markers are the ratio of Easy samples in the ground truth. If the ratio is 0 or 1, the class-balanced accuracy corresponds to accuracy.

The top panel in each sub-figure (session) is the result per subject computed with 96 samples (8 decision points × 12 trajectories) of the entire EEG block. For both sessions, in total 16 out of 26 recordings have a class-balanced accuracy higher than the chance level.

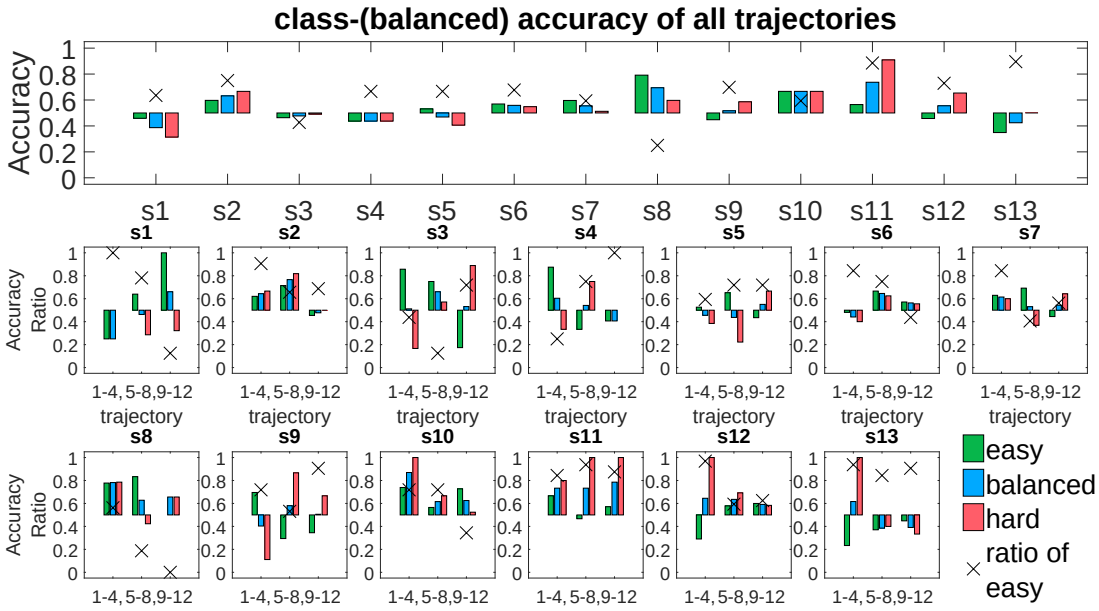
Other panels are the result grouped by four trajectories for each subject. Each group ensures the same bias term and the accuracy is based on 32 decision points. For both sessions, 50 out of 78 blue bars are above the chance level.

Figure 4.10 shows the used shift of bias term during online decoding, where zero means that the bias term of regression was the same as the best model learned from the offline session. The red bars highlight the recordings using the same shift for all the 12 trajectories, in total, 9 out 26 kept the same shifts, where s10 had zero shifts across both recordings.

The online decoding accuracies were not as high as the offline validation. This can be a typical case for EEG decoders. Thus, the accuracy in open-loop is further evaluated, to know which accuracies can be achieved with the best bias term.



(a) Session 2



(b) Session 3

Figure 4.9 – Online decoding accuracy in EEG condition at decision-point level, in terms of each class and class-balanced accuracy.

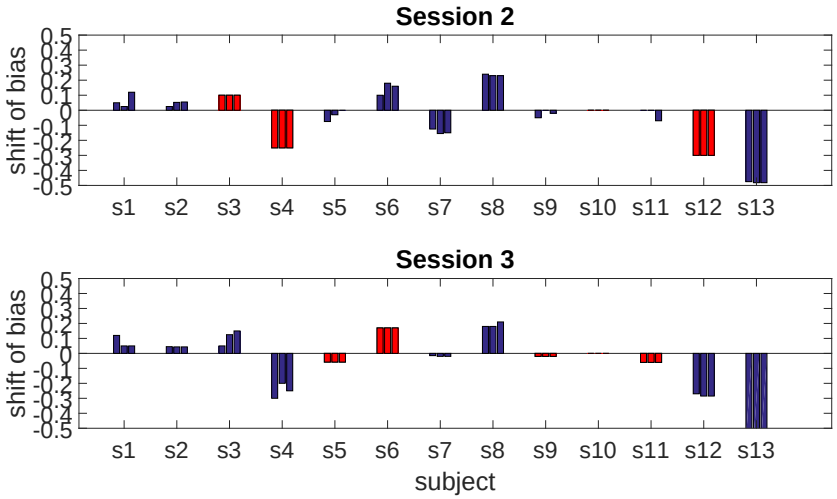


Figure 4.10 – Shift of the bias term in online sessions, where red bars indicate the same shift in a session.

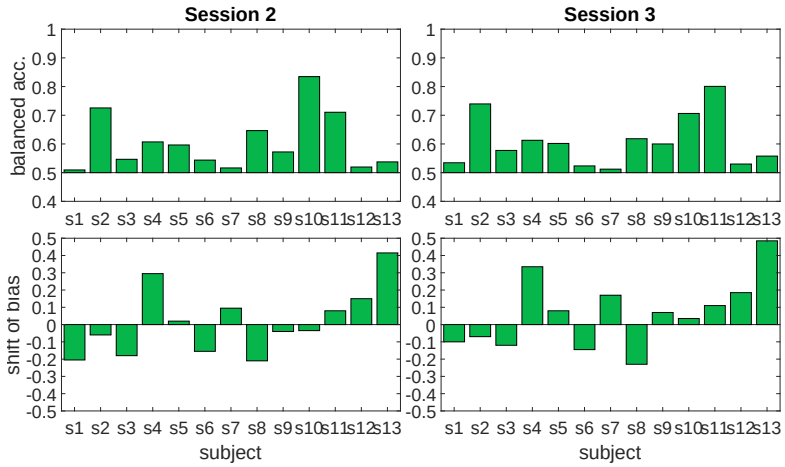
Maunal Condition & Skill Evaluations

Figure 4.11 plots the best class-balanced accuracy that could be achieved with the best bias term in the open-loop cases, where no EEG feedback was provided to the subject. Specifically, the open-loop cases refer to both the Manual condition (see Figure 4.11(a)) and skill evaluations (see Figure 4.11(b)). The first skill evaluation (personalization) is not shown because it was done during the setup without recording physiological signals.

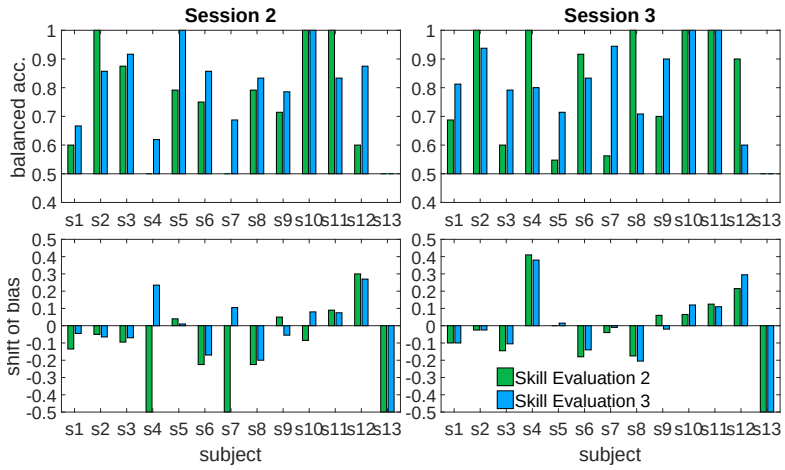
The evaluation was done by a post-simulation which has the same decoding procedure as in the closed-loop case. Different shifts of the bias term were scanned. The shift ranged from -0.5 to 0.5 with 0.005 as a step. The upper panels of Figure 4.11 report the best accuracy per subject and the lower panels show the corresponding best shift of bias term. In Figure 4.11(b), each skill evaluation is drawn in a specific color.

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 is therefore highly dependent on the subject. The average difference of the best shift between sessions was 0.05 with a standard deviation of 0.04, showing that the best bias term can be relatively stable across sessions.

On the other hand, the skill evaluation part did not yield consistent rankings in accuracies between sessions ($r = 0.49$, p-value = 0.12, $n = 13$ for the 2^{nd} evaluation and $r = 0.09$, p-value = 0.71 for the 3^{rd} one). The average differences of the best shift between session were 0.12 and 0.01 for the 2^{nd} and 3^{rd} 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.

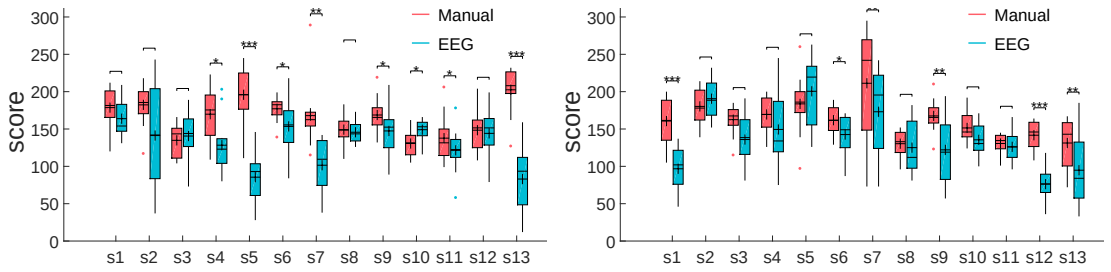


(a) Manual condition



(b) Skill evaluation

Figure 4.11 – Open-loop class-balanced accuracy at decision-point level in the Manual condition and the last two skill evaluations.



(a) Session 2 (b) Session 3

Figure 4.12 – Task scores and one-sample t-tests (*: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

Table 4.1 – Parameters of Skill Curves

		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

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 as can be seen in Figure 4.8.

4.5.5 Online – Task Scores

A paired t-test comparing the scores of the Manual and EEG conditions at the group level was performed for each session.[1] Session 2 yielded a p-value of 0.01 with 13 samples and an effect size of 0.8. Session 3 yielded a p-value of 0.005 with 13 samples and an effect size of 0.95. Both tests showed that there is a huge effect between the two conditions.

In order to investigate at the subject level Figure 4.12 shows the comparison of task scores between the two conditions 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.

T-tests were conducted between both conditions for each subject ($n = 12$, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$). Although the scores for Manual condition are generally higher than the EEG condition and has a smaller variation, they are not necessarily significantly better. This suggests that both conditions can be similar at the subject level. Interestingly, a few cases show that the average score (+ sign) for EEG condition could be higher than the Manual one. Specifically, s3 and s10 in session 2 are the cases, and s10 is statistically higher; s2 and s5 in session 3 are also higher. This implies that using an EEG decoder still has the potential of having a better task score.

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

Table 4.2 – Parameters correlates of Accuracy in 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 4.3 – Difference of Overall Hit Rates

	Manual - EEG				1st block - 2nd block			
	Hit rate difference		Correlation		Hit rate difference		Correlation	
Session	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

4.5.6 Online – Skill Curves

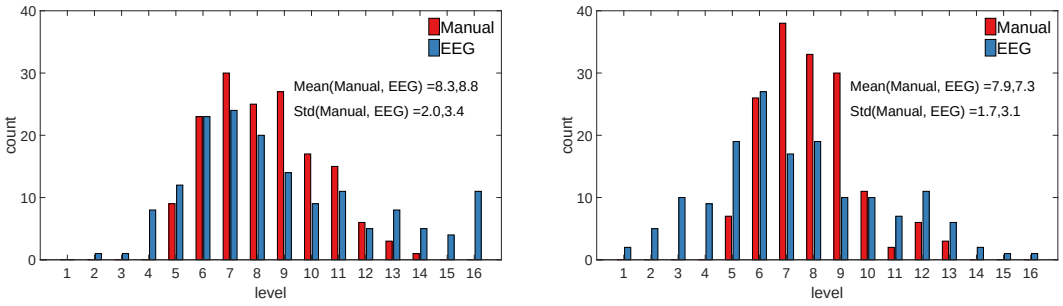
Following Section 4.4.2, Table 4.1 lists the result of t-test for x50 and slope. The condition columns mean that the test was conducted between the Manual and EEG conditions, while the Time columns refer to the result between the first and second presented blocks. Row p is the p-value and Power is the t-test statistical power, where a positive value means that Manual (1^{st} block) is higher than the EEG condition (2^{nd} 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 4.2 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, as presented in Chapter 4.4.2, Table 4.3 lists the result of overall hit rates in different sessions for comparing the two conditions and different blocks. Positive values in the Avg columns indicate that Manual (1^{st} block) is better than the EEG condition (2^{nd} block). No significant correlation was found suggesting that the decoding accuracy does not correlate to either of the checked factors.

4.5.7 Online – Final Levels

Figure 4.13 summarizes 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 final level located between level 6 to 9 regardless of the condition. EEG condition as expected to be less stable than the Manual, shows a higher standard deviation and higher counts on the tails of the distributions.

Figure 4.14 illustrates the curves representing the final levels of each trajectory for each subject



(a) Session 2

(b) Session 3

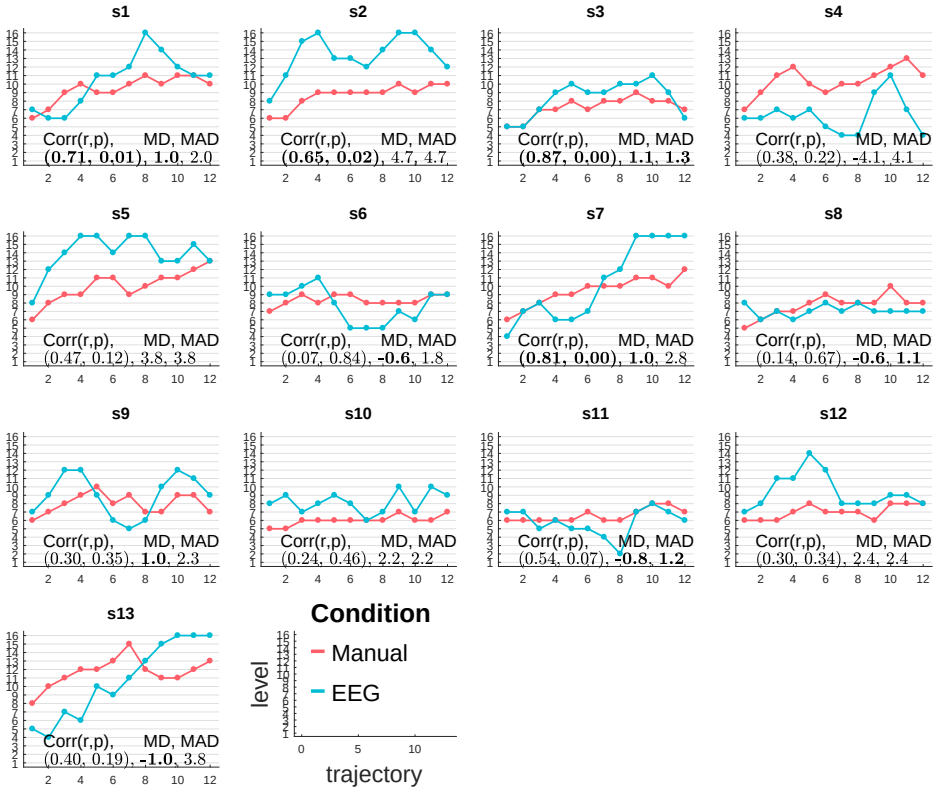
Figure 4.13 – Statistics of final level.

and session. The red curves stand for the Manual condition while the blue ones are for the EEG condition. 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 [EFG16]. 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, the 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 had similar patterns as the Manual cases. The adjustments of the bias term probably favored similar patterns, but the adjustments were two times at most and some subjects did not need.

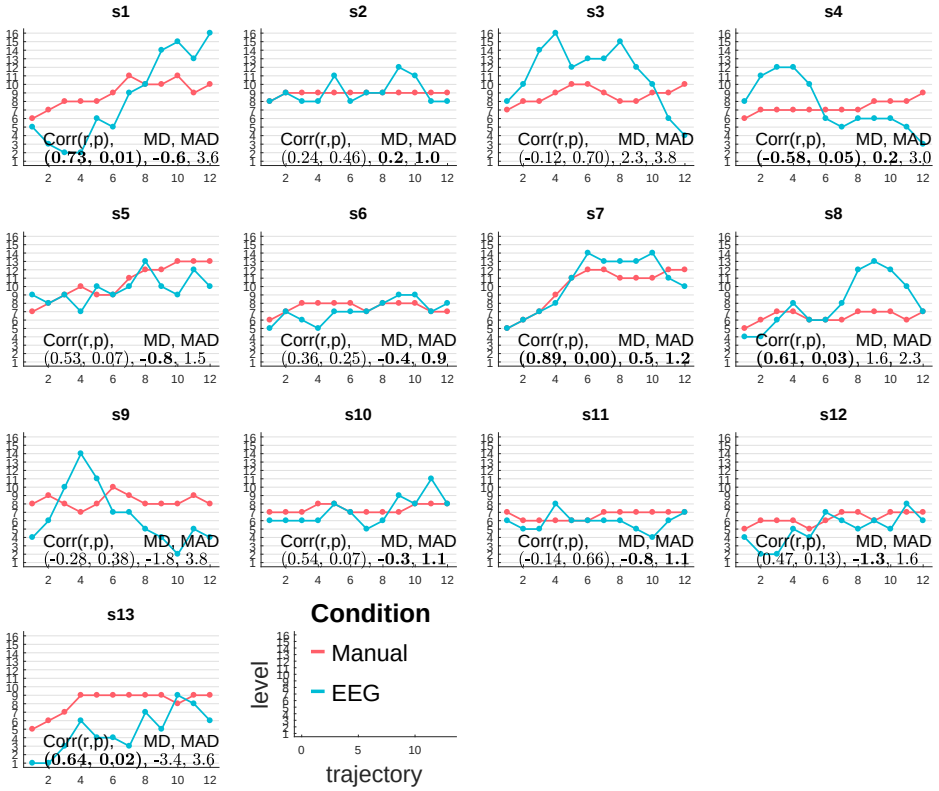
Apart from the curves of final levels, the correlation, MD, and MAD are also indicated for each subject as texts. Boldface is used to emphasize the values with high similarity. Specifically, the highlighted values are correlation coefficients with significance, and also for |MD| (MAD) which are below 1.5 level (level^2). The proposed indices adequately played their roles as predicted. For example, subject s11 has very good values in MD and MAD for both sessions but the r is rather low. Referring to the curves in both sessions, s11 has the patterns of EEG conditions being similar to Manual ones. However, there are a few overshoots being over-emphasized by the correlation. In total, five and seven subjects in session 2 and 3, respectively, have two indices indicating high similarities.

4.5.8 Correlation to Decoding Accuracy

A question of interest is the role of decoding accuracy in the interaction, given that the behavioral results are similar even if the online decoding accuracy was not very high. In addition to the correlation tests conducted in Table 4.2 and 4.3, other additional correlation tests were performed. Between the task scores and the online accuracy, the correlations in session 2 was 0.15 (p = 0.62, n=13), and 0.13 (p = 0.67, n=13) in session 3. The correlations with r in session 2 was -0.28 (p = 0.34, n=13), and 0.01 (p = 0.98, n=13) in session 3. The correlations to MAD in session 2 was -0.01 (p = 0.98, n=13), and -0.67 (p = 0.01, n=13) in



(a) Session 2



(b) Session 3

Figure 4.14 – Final levels across trajectories.

session 3. The significant negative coefficient indicates that the higher the accuracy is, the lower the dissimilarity. The correlation with MD in session 2 was 0.46 (p = 0.11, n=13), and 0.27 (p = 0.37, n=13) in session 3. Although there was one significant result, it is probably by chance.

4.6 Discussion

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. This indicates that either (1) hit rates did not consistently reflect the subjective feeling, (2) some subjects largely improved their skills than others during the 32 trajectories, or (3) the estimated hit rates were not robust enough as only 11 data points were used. 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 of some subjects biased towards one class, the tuning of the bias term in regression in the online sessions should be able to avoid or alleviate such a situation.

Many behavioral results in online sessions suggest that the EEG and Manual conditions were similar, except that the task scores of Manual condition were significantly better than the EEG condition at the group level. At subject level, the task scores of both conditions could be similar; about half 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. In addition, the curves of final levels can be very similar between both conditions, and nearly every subject had one session with a close pattern. These are positive results to support 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 [WRB+17], 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.

Unlike the study of Faller *el. al* which was also using a simulated drone [FCSS19], this chapter did not directly aim at regulating the user's state but rather to make the task level fit the user's need. Faller *el. al* instructed the subjects to regulate the arousal state while performing the navigation. On the contrary, the subjects in this chapter were instructed to complete the task without explicitly paying attention to the decoder, and the subjects should try to be consistent with the definition of the descriptive label when pressing the button. That means, pressing the button identifies the same perceived level of difficulty (Easy in our case) in both the online and offline sessions. As a result, these differences made our Manual condition more or less a ceiling instead of a baseline as in [FCSS19], and therefore, not really comparable. An interesting finding is that in rare cases the EEG condition overpassed the ceiling. The only

significant one 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. In other words, the decoder tended to make the level harder, forcing the subject to challenge himself by leaving his comfort zone.

Although the offline validation accuracy was high, it did not necessarily guarantee high online decoding accuracy, especially when they were recorded weeks after. This cross-day variation might come from fatigue, different background states, different setups, *etc.* It is hard to confirm the main reasons, but still, there are some computationally expensive approaches to alleviate this cross-day variation [LJY17, JCM18b]. Instead, I chose the baseline subtraction approach, because it can be combined with the adjustment of the bias term to save the computational power and still worked well for some cases. Also, as long as the best shift of the bias term can be found, the decoding accuracy could still be high.

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 leaning to level 1 or 16. The correlation to the accuracy in Table 4.2 and 4.3, however, only yielded one significant result. Other correlation tests in Chapter 4.5.8 also yielded limited evidence of the role. As a result, it is hard to draw any conclusion between the decoding accuracy and the behavioral results. This indecisive conclusion, however, is not necessarily due to accuracy, but might also be largely influenced by the personalization of difficulty without many samples.

4.7 Conclusion

It is generally believed that leaving a comfort zone benefits learning. Under this belief, the designed one-directional regulation demanded the subjects to leave their comfort (Easy) zone. With an EEG decoder with adequate offline accuracy, the closed-loop experiment demonstrated that the behavior results could be similar between using an EEG decoder and self-paced decision. 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 subjects with higher online decoding accuracy are still needed for drawing 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 only around twelve seconds. The dynamics of cognitive states around an event of changing difficulty levels should be investigated. Apart from improving the accuracy, another future research direction is examining the applicability of the decoding framework in other tasks. This helps to see whether the decoded cognitive state is bound to a certain kind of visuomotor activity, *i.e.*, whether it is task-specific.

5 Dual Task and Temporal Dynamics

5.1 Introduction

The previous chapters have demonstrated that (1) decoding difficulty-related cognitive states from EEG signals is possible and (2) users can potentially benefit from integrating the decoder in the loop. *Another question of interest is the applicability of the modeling of decoders in different scenarios.* Before, the subjects focused on a single task while using only their right hands. Would the modeling achieve a satisfactory outcome in a different task? In order to answer this question, a second task is introduced to modulate the task difficulty, and the performance of decoders will be analyzed. Additionally, this time the drone-navigation task will be performed with left hands. This may help eliminate the confounding factor whether the EEG decoder is decoding cognitive states or muscular activities.

As identified in Chapter 3, in order to reach a satisfactory decoding accuracy, a trade-off with latency in post-processing is essential. This latency can be considered as extrinsic from the subjects since the latency comes from the signal processing (*e.g.* sample averaging). *Another question of interest is whether an intrinsic latency, embedded at EEG signals and non-reducible, can be observed.* Assuming that there is a perceivable latency in EEG signals, I denote a *transition period* as the time ranging from the moment when difficulty changes to the moment we can reliably decode the new cognitive state (*i.e.* a *stable period).* With an adequate analysis, we can understand the effect of dynamical changes of difficulty on users' cognitive states and further reveal whether an intrinsic latency can be observed and estimated. This is different from previous HMI studies on decoding cognitive states (*e.g.*, workload, levels of attention) which have mostly focused on setups where the conditions of the task remain constant and assume that the user states will remain it as well [EFG16, FCSS19, JCM18a, ABD$^+$15]. The protocol in this chapter additionally changes the level of difficulty within a trial instead of only one condition per trial.

5.2 Materials

5.2.1 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. of 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 resolutions. They held a game-pad[1] to provide inputs to the protocol including a joystick on the left side and four colored buttons on the right side. All subjects had a 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 EOG signals at 2,048 Hz. Sixty-four EEG electrodes were placed according to the international 10-10 standard. Three additional channels were placed as Figure 2.1(a) to measure the EOG .

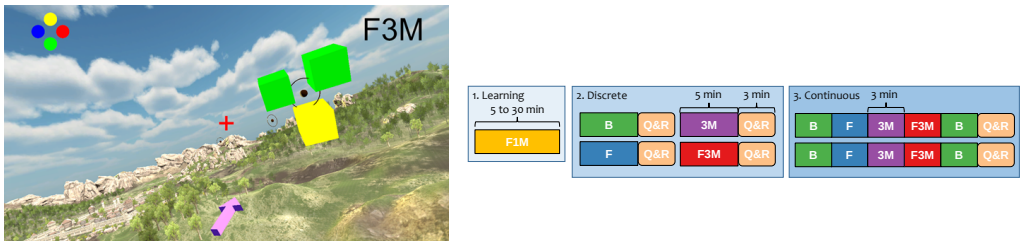
5.2.2 Protocol

As stated at the beginning of this chapter, one objective is to induce different levels of task difficulty with another task. Each recording session was therefore composed of multiple conditions including flying a simulated drone and visual recognition, as well as their combination [DCA18]. The first one is a low-difficulty baseline condition (B), where the subject watched the drone automatically flying through the waypoints. The subjects were instructed not to use the game-pad. Any attempts to control would take no effect.

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

The third condition was a visual recognition task, where the subject was mapping targets (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 in total of 80 groups of objects over 122 waypoints. Subjects were asked to press the corresponding color-coded buttons on the game-pad upon the appearances of the objects. If the subject pressed the correct color-button (hit) the corresponding object disappeared; otherwise (miss) there was no effect in the visual feedback. The subjects were told that a miss reduces their task performance, in order to avoid random pressing. The colors of the three

[1] Logitech, “Gamepad F310.” http://support.logitech.com/en/product/gamepad-f310



(a) Protocol screen-shot (b) Design of experiment

Figure 5.1 – Protocol and the design of experiment. (a) Top-right indicates current task. Top-left is the layout of button colors. The purple arrow below is an indicator to the current waypoint. The red-cross is the center of drone. Circles are waypoints to be passed. Colored cubes are the objects of the mapping task. (b) Each session consists of one learning phase, (presenting different conditions in) one discrete phase and one continuous phase. The presenting orders of 3M, F, and F3M in discrete and continuous phases were pseudo randomized in the two sessions for each subject.

cubes could be the same, where one pressing would only eliminate one cube. The order of pressing did not matter.

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

Figure 5.1(a) is a screen-shot of the protocol. On top-right, the current condition was displayed (in this case F3M). On top-left, the layout of the buttons on the game-pad was presented as a reminder to the subjects to prevent them from looking at the game-pad and to help to fixate on the screen, reducing possible EEG artifacts from muscle activities during the recording. On the bottom-center, there was an arrow indicating the direction to the next target waypoint. This was particularly helpful in cases when the target waypoint falls outside the current view. The waypoint has a small sphere in the center. The size of waypoint was small such that most subjects needed to focus on the navigation task if they want 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 miss of waypoints. The red-cross in the center was the drone's center used to determine if it is inside a waypoint. The cubes to be mapped appeared around the target waypoint and completely disappeared when the target waypoint changed, even if the subject failed to eliminate them.

In order to study the transition period, each recording is designed with three phases as depicted in Fig. 5.1(b). The *first phase* was devoted to allowing the subject to get familiar with the task and there was no recording of neural signals. During the setup of sensors, 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 try as many times as she wishes before starting the experiment. Setting up of all sensors normally took between 30 to 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 filling and a resting period (Q&R). Q&R period was introduced to reduce the effect of performing an experimental condition on the following one and to allow studying the transition effects from a resting state. During this period, the subject reported their perceived difficulty level from 0 to 100 and their workload level through the NASA-TLX [HS88]. Both questionnaires were used to assess if 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 session. With the five-minute constraint, the subjects finished 93.4 waypoints out of 122 on average.

During the *third phase (continuous)*, experimental conditions were presented without interleaving resting periods, allowing to study the effect of dynamical changes in the level of difficulty on the neural signals. There were two trials per recording, each composed of five segments corresponding to the different conditions (See Fig 5.1(b)). 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. At the end of a trial, there was also a Q&R period as in the second phase. It should be noticed that due to the continuous nature of this task, it was impossible to immediately gather subjects' self-reports for each condition. Each condition lasted 3 minutes. In total, the two trials provided 30 minutes of effective signals.

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

5.3 Decoding Cognitive States from EEG

5.3.1 Signal Pre-processing

The pre-processing method is nearly the same as stated in Chapter 3.4. One specific thing to mention is that the spectral filter was done with forward and backward processing to eliminate the latency introduced by the filter. This is applicable as this chapter only involves offline analysis.

Contaminated EEG electrodes were manually removed based on visual inspection. Out of 48 recordings, CP1, POz, and PO3 were removed once, PO4 was removed twice, and P2 was removed four times.

As the validation will be divided only by the two phases, the selection of relevant components was done in each training fold (see Chapter 5.3.3 a description of the cross-validation) instead of using all the data as in Chapter 4. On average, 15.3 components were kept for analysis, and

Table 5.1 – Different settings of numeric and descriptive labels in classification

Condition			Workload or Difficulty level		
Target	Numeric	Descriptive	Target	Numeric	Descriptive
B	0	0	1^{st} lowest	mean level/100	0
F	0.33	1	2^{nd} lowest	mean level/100	1
3M	0.66	2	3^{rd} lowest	mean level/100	2
F3M	1	3	4^{th} lowest	mean level/100	3

1.7 components were removed when using both phases to train.

5.3.2 Classification Method

The classification method is mostly the same as stated in Chapter 3.4. A GLM-based regression is applied, followed by a moving average for the post-processing. Then, an LDA is employed to classify. The post-processing here used the previous five seconds instead of six. The search range of λ in the regression was reduced to [0.05, 0.1, 0.5, 1] for the sake of computational time. The numeric labels were all normalized between 0 and 1 to fit the chosen link model, binomial.

The classification method here considers using three different labels, namely, condition labels, workload levels, and difficulty level to approach the problem. Table 5.1 summarizes the settings for using the three labels.

When the decoded cognitive states are to be directly mapped with the task condition, the numeric labels for regression of B, F, 3M, and F3M are mapped as 0, 0.33, 0.66, and 1, respectively. The descriptive labels for the LDA are evidently B, F, 3M, and F3M and are mapped to 0, 1, 2, and 3, respectively, in the implementation.

When the decoded cognitive states are to be mapped with the order of self-reported workload level or difficulty level of one subject, the average level of each condition was first computed from all the three trials and then sorted. The numeric labels for regression are the average levels divided by 100 for the normalization, and the descriptive label of the lowest level is being assigned as 0, second lowest as 1, and so on for the LDA. As each average level is still associated with a condition, this setting still can be considered as decoding the task condition, in the sense of their order in the workload or difficulty level.

For the post-processing, the previous 9 samples were averaged to reach a compromise between accuracy and latency; since a log-PSD-based feature was computed every 500 ms, this resulted in a five-second delay. In the special case of the samples at the beginning of a trial, padding was needed. The first sample of a trial was replicated instead of averaging with zeros.

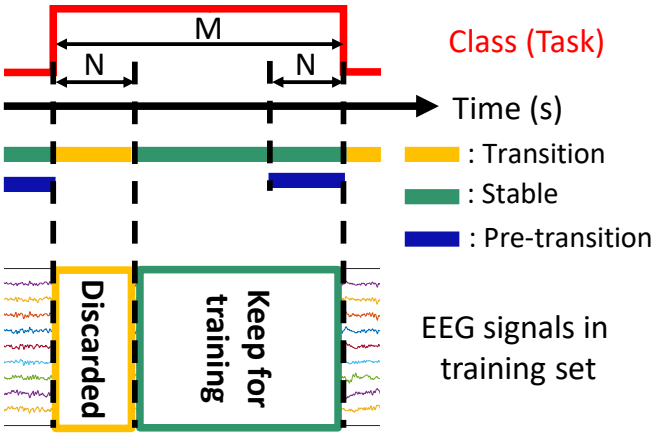


Figure 5.2 – Validate the transition period. The three defined periods using one task as an example. M is the duration of the task while N is the assumed duration of transition. For training a decoder, only the signals from the stable period were used. The test accuracies were computed for each period for different values of N on test sets.

5.3.3 Assessment of Decoder Performance

Decoding performance was assessed using the previously defined class-balanced accuracy, where a two-phase cross-validation was performed for each recording. This helps to distinguish different conditions and characterize the transition period. In other words, either the discrete or continuous phase was served as training data while the other was treated as test data. It is important for this type of studies not to use a cross-validation method that does random partitions; since a random partition may easily have the training and test samples being close in time which violates the principle of independent training and test sets [VRE+17], and thus yielding optimistic performance estimations.

Different subsets of the four conditions are to be evaluated for a comprehensive analysis. Following the descriptive labels defined in Table 5.1, the analysis includes all the six combinations of two classes, [0, 1, 3] and [0, 2, 3] for three classes, and [0, 1, 2, 3] for four classes. In the case that only a subset of classes is targeted, training data used in Chapter 5.3.2 only contains the data from the subset. However, the decoder was still simulated on all the data but only evaluated on the subset. The full simulation guarantees the transition effect will not be distorted by truncating the signals.

5.4 Transition Analysis

5.4.1 Transition Period

Assuming N seconds are needed for EEG to reach a *stable* state (*i.e.* a state where the cognitive state can be reliably decoded) from the onset of changing condition. Then, the neural activity

in the first N seconds will not reflect the new state. Hence, a decoder should 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 Figure 5.2. A 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 one is a pre-transition period that is a subset of the stable period and corresponds to the last N seconds before the task change. The pre-transition period was defined as a control term where accuracy should be higher than in the transition period in the case there exists the hypothesized transition effect.

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 in between. Therefore, only the data in the stable period was kept for training; *c.f.*, Figure 5.2. This ensures the training samples are rather reliable and can build a better model. Testing performance is reported as the class-balanced accuracy for each period and different N values.

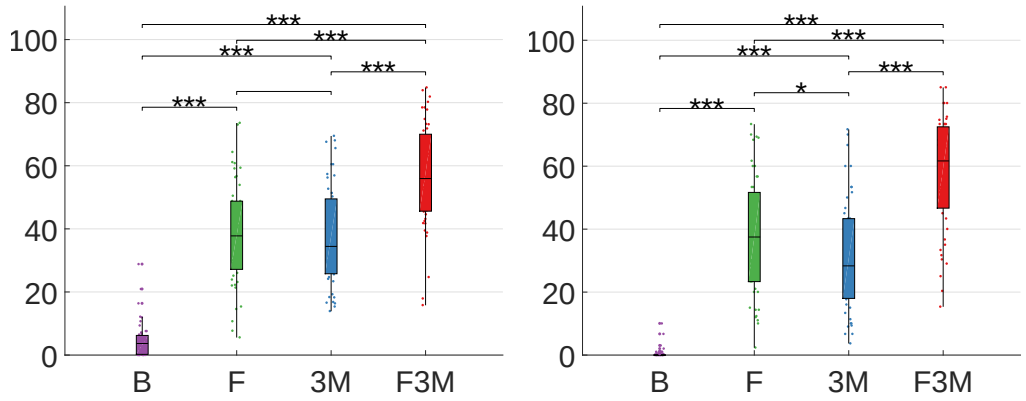
According to the hypothesis, EEG correlates will only reliably reflect the current state after the transition period. Thus, it is expected that the accuracy curve of the transition period should improve as N increases. One reason is that the decoder may have been trained with a larger portion of reliable 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, both reasons are non-separable. If the underlying hypothesis does not hold true, the accuracy curves should be flat for all the 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.

5.4.2 Estimating the Intrinsic and Total Latency

Once the transition period is confirmed, the next objective is to estimate the delay. The idea is to choose a large N (60 s) such that the training data only belongs to the stable period. Then, using a sliding window of two seconds to compute the class-balanced accuracy on the transition periods of the test set.[2] In Chapter 5.4.1, the assessed accuracy is coarse since all the samples before N seconds are considered. It is hard to precise at which point the state becomes stable. As a result, computing the accuracy inside a smaller window can better reflect the exact time point. If a window to compute accuracy contains only stable data, the accuracy should be higher than those containing any data from the transition period.

However, the sliding window approach still cannot exclude the latency from signal processing,

[2]The two-second window includes any log-PSD window that has its beginning time inside the two-second window.



(a) Workload level (b) Difficulty level

Figure 5.3 – Subjective assessment averaged across trials. Two-tail t-tests with 48 samples were applied (*: $p < 0.05$, and ***: $p < 0.001$). The quartiles are represented by boxes, where the dots are the data points.

especially the one from post-processing.[3] In order to exclude this kind of extrinsic latency as much as possible, a forward and backward, instead of forward, post-processing was conducted which shifts the extrinsic delay from the transition period to the pre-transition period. Consequently, this minimizes the potential effect from the post-processing, although not feasible in online decoding.

5.5 Results

5.5.1 Self-Reporting Questionnaires

Figure 5.3 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,[4] shown as colorful dots; within-session values were averaged across the three trials. The workload and perceived difficulty levels increases in the order of B, 3M, F, and F3M.

One-way ANOVAs were 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-tail 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 of difficulty level ($p < 0.05$).

The Pearson's correlations between workload level and difficulty level for each condition are all statistically significant (p-value < 0.001; $n = 48$). The correlation coefficients in B were 0.53, 0.57, and 0.56 for trial 1, 2 and 3, respectively; for F, they were 0.81, 0.83, and 0.82; 3M were 0.88,

[3] Delay from the spectral filter can be neglected, since a forward and backward processing was used.
[4] Each data point was averaged across the three trials.

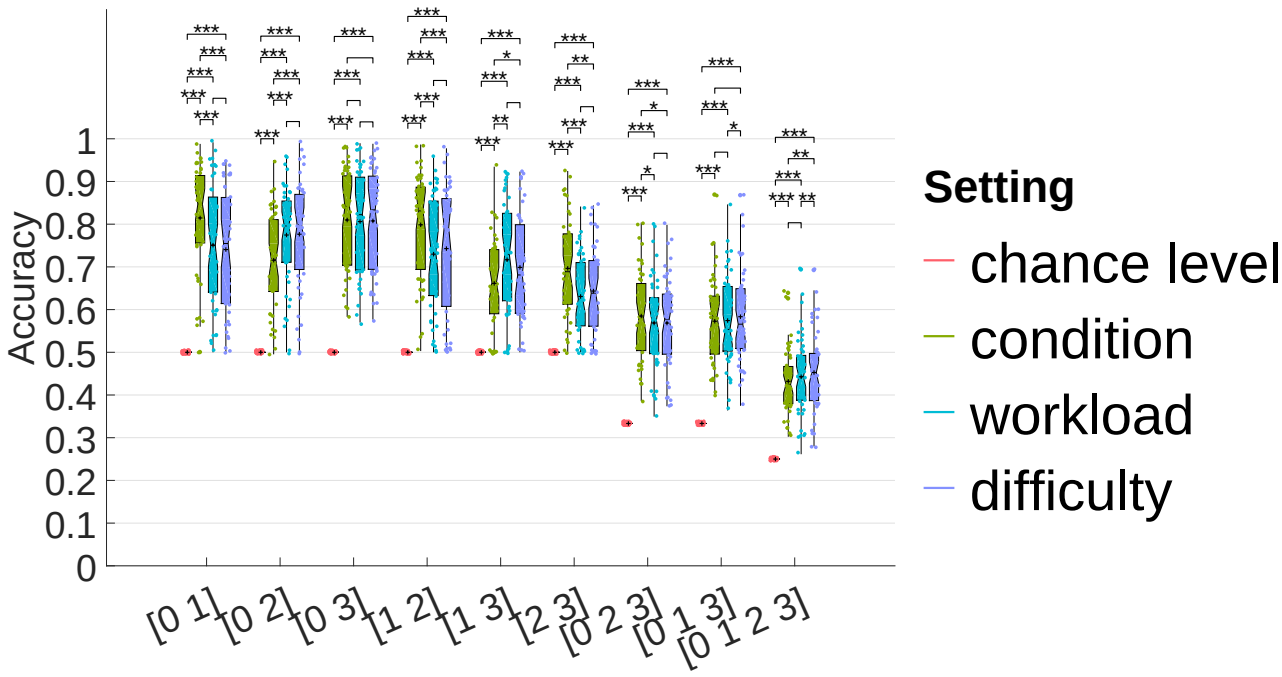


Figure 5.4 – Decoding accuracies of each targeted classes under different settings of labels.

0.83, and 0.88; F3M are 0.77, 0.86, and 0.83. The high correlations suggest that workload level and difficulty level are highly intertwined and using either of them as ground truth should not result in big differences regarding decoding approaches.

5.5.2 Decoding Using Different Labels

Figure 5.4 shows the decoding accuracy of different combinations of the targeted subset (x-axis) under different settings (color). Each color stands for either the chance level accuracy, using the conditions, workload levels, or difficulty levels for building the decoders. The black crosses are the mean values while the boxes are quartiles.

One-way ANOVA was performed to examine if there is an effect of using the three different labels (excluding the chance level). The test yielded non-significant result with $F(2,24) = 0.01$, $p = 0.99$. Another one-way ANOVA including the chance level was conducted and yielded significant result with $F(3,32) = 8.3$, $p = 0.0003$.

Paired-wise two-tail t-tests were also performed ($n = 48$, *: $p < 0.05$, and ***: $p < 0.001$) between each setting as shown in Figure 5.4. Using either workload or difficulty level for the decoder only makes little differences, as both levels are highly correlated. The only exceptions are in the cases of [0 1 3] and [0 1 2 3], where using the difficulty level yielded significantly higher accuracies than using the workload level. In the case of [0 1 2 3], the accuracy was even significantly higher than using conditions.

However, for unknown reasons, using either the difficulty or workload level yielded significantly worse results than using the conditions for the cases of [0 1], [1 2] and [2 3]. As noticed, a major difference is that, when using the conditions as labels, the continuous label of one condition has the same value in all the three trials, but the difficulty or workload levels have three values (trials) of each condition. Ideally, more values are better as more variations of signals and cognitive states can be captured, especially when a trial is long. However, the results probably indicate that zero variation in the labels is better than the case of limited variation when training the regressors or decoders.

Given that using the condition-based regression has more significantly better accuracies than using the workload or difficulty level, the later analysis will focus on this setting. Even though it is not directly regressed onto the difficulty levels, those conditions still have different difficulty levels.

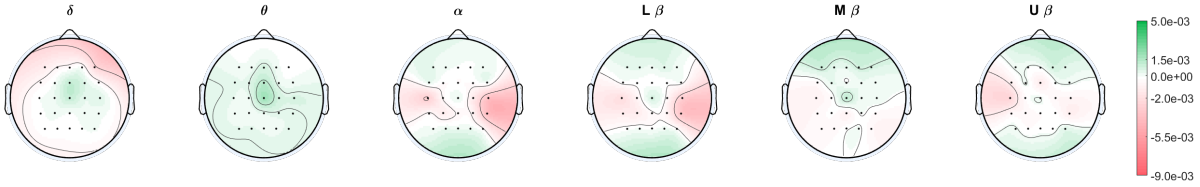
5.5.3 Neural Correlates

Figure 5.5 shows the regression coefficients averaged across all recordings, where the coefficients of a recording were learned from all the three trials. In each sub-figure, topographic plots show the coefficients averaged for six different frequency bands, namely δ (2-5 Hz), θ (5-8 Hz), α (8-12 Hz), lower β (12-16 Hz), mid β (16-18 Hz), and upper β (18-28 Hz), such that all frequency bins are included in the analysis. Negative values (shown in red) mean that low log-PSD values favor the most difficult condition indicated below the sub-figure, while positive values (in green) embrace the easiest condition, Baseline. The white color stands for a zero coefficient (*i.e.* the feature carries no or little information about the target output).

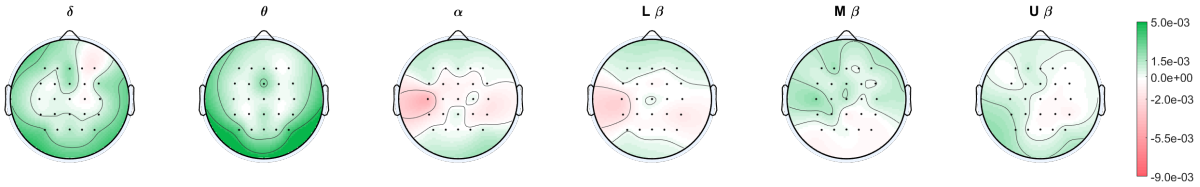
As can be seen from Figure 5.5(a) and 5.5(b), while using B as the reference, F and 3M both have a common negative pattern within the α band over the left sensorimotor region. F additionally has stronger α features in the right sensorimotor region, specifically at CP4 and C4, and a positive pattern at the center (Cz) within the θ band. 3M also has positive patterns within the δ and θ bands at the parietal region. The θ band is also rather evident in the frontal central site (Fz). For F3M (see Figure 5.5(c)), the patterns are similar to condition F. In addition, within θ band, there are stronger components around the peripheral electrodes. FCz and Fz are also stronger within the δ band while the α band in red is also stronger. In the case of three- and four-class classification (Figure 5.5(d), 5.5(e) and 5.5(f)), the patterns are pretty similar to [B, F3M], but with smaller coefficients. In short, the most useful patterns are the α bands in C3, CP4, and C4, δ band at FCz, and θ band at Fz.

5.5.4 Transition Period

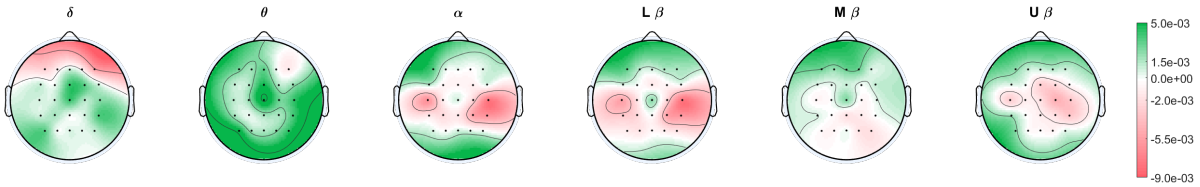
Figure 5.6 confirms that the presumed transition period exists. Figure 5.6(a) and Figure 5.6(b) illustrate the accuracies tested on the discrete and continuous phases, respectively. Whenever testing the discrete phase, transitions considered are from rest to a condition, B, F, 3M, or F3M.



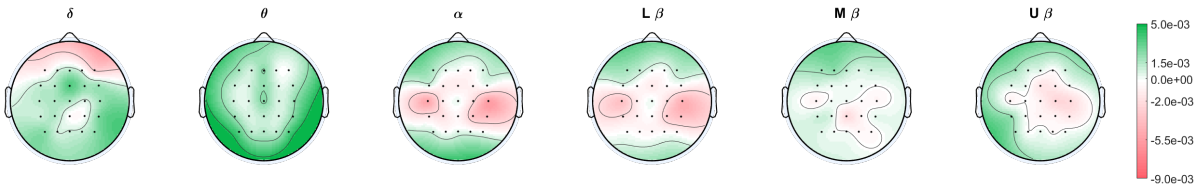
(a) [B, F]



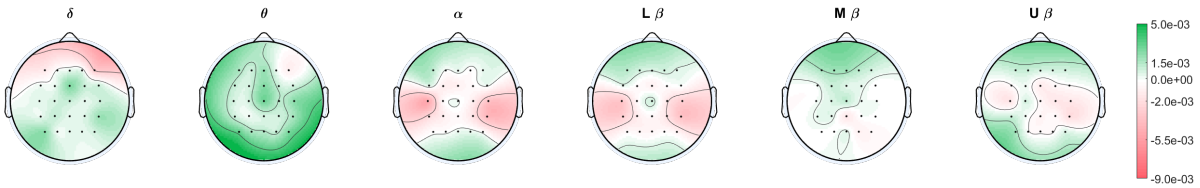
(b) [B, 3M]



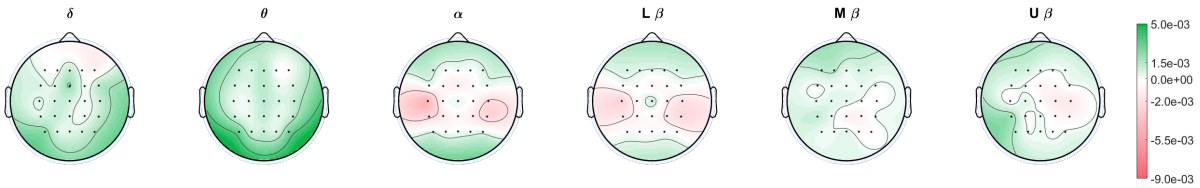
(c) [B, F3M]



(d) [B, F, F3M]



(e) [B, 3M, F3M]



(f) [B, F, 3M, F3M]

Figure 5.5 – Regression coefficients averaged across 48 recordings using conditions as labels. Red (Blue) means that a low (high) log-PSD power favors the Baseline condition below each sub-figure The values range between 5×10^{-3} (Red) to -9×10^{-3} (Blue), and white means 0.

While tested on the continuous phase, a transition was either between rest and a condition, or any two conditions adjacent in time, even one of the condition is not in the targeted classes.

Each panel represents a different set of target classes. The red, green, and blue curves stand for the defined pre-transition, stable, and transition periods, respectively, over different N (x-axis). The accuracies were computed over the 48 recordings, and the shaded areas plot the 95% confidence interval of the mean values with the assumption of Gaussian distribution [MO18]. The tested N are marked as circle dots on the curves. It is clear that nearly each transition curve exhibits an increasing trend and reaches a stable state between 10 and 20 seconds, while the stable and pre-transition curves are flat. The confidence intervals clearly indicate that the accuracies in the transition periods are significantly different from the other two curves. Consequently, the transition period is a phenomenon worthy of further examination and 60 seconds is a large enough N to train a decoder for estimating the latency in the next section.

5.5.5 Estimation of Latency

Figure 5.7 provides more precise information regarding the latency by computing accuracies through a two-second sliding window. The curves and shaded areas follow the same idea as Figure 5.6, except that the data used here includes all the recordings and the two phases in the same figure.

Figure 5.7(a) provides the information of total latency, intrinsic plus post-processing.[5] It can be seen that the transition curves reached a rather stable state between six to twelve seconds, depending on the targeted classes.

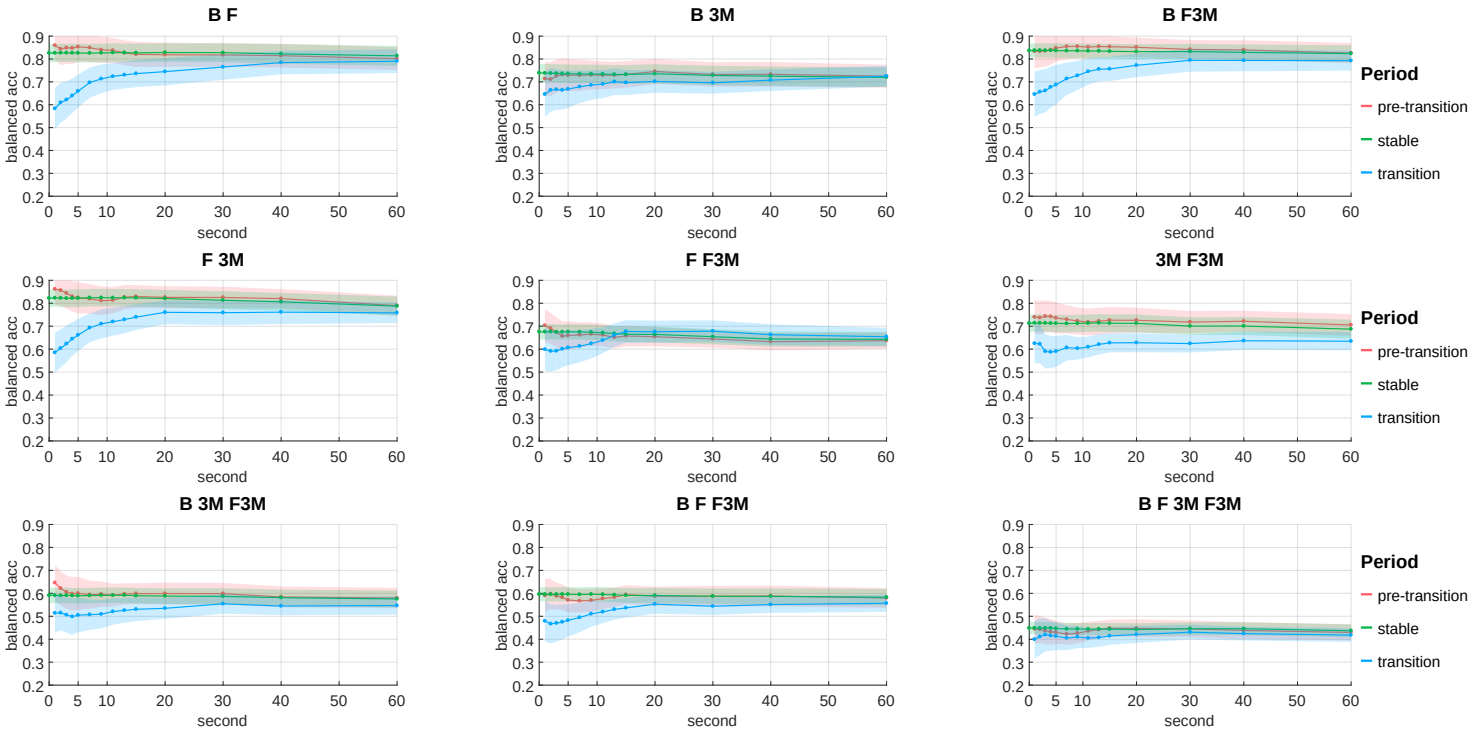
Figure 5.7(b), on the other hand, provides the intrinsic latency by using the forward and backward moving average. In the case of [B F3M] and [B 3M F3M] for example, there are evident transition periods, and the estimated latency is around four to eight seconds. In the case of [F F3M], the transition is not evident. This is probably because the decoding accuracy is not high enough. The accuracy corresponds to the green box of [1 3] in Figure 5.4 and is the lowest among all the binary classification problems.

5.6 Discussion

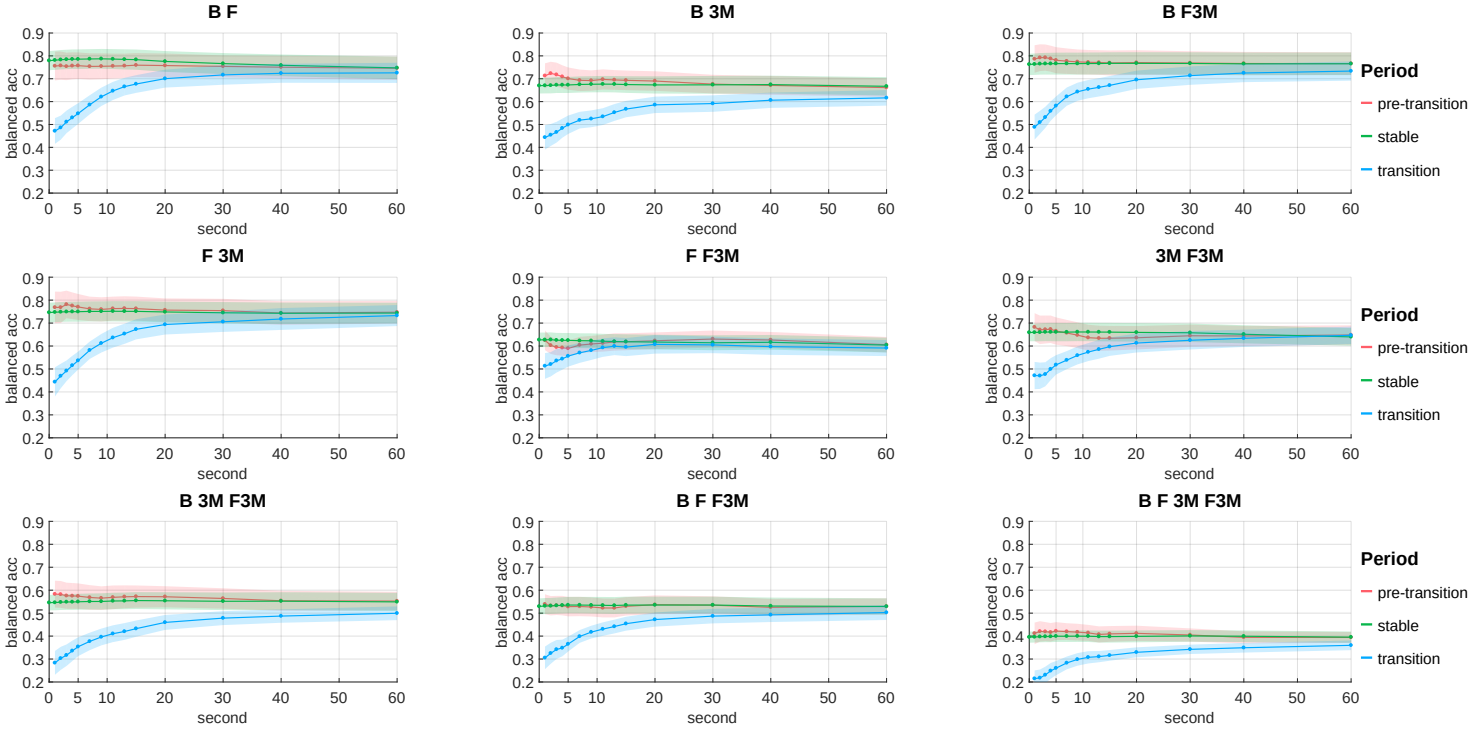
Except for the workload level between 3M and F, the subjects consistently reported significantly different workload and difficulty levels for the four conditions, supporting the suitability of the experimental design to study cognitive states.

One important question is whether the EEG decoder is really decoding cognitive state or motor activities. Some positive evidences supporting the cognitive state can be found on the analysis of regression models (*c.f.* Figure 5.5). The analysis shows that α band is highly discriminant

[5] Another latency being ignored here is the spectral filtering as forward and backward processing was used.

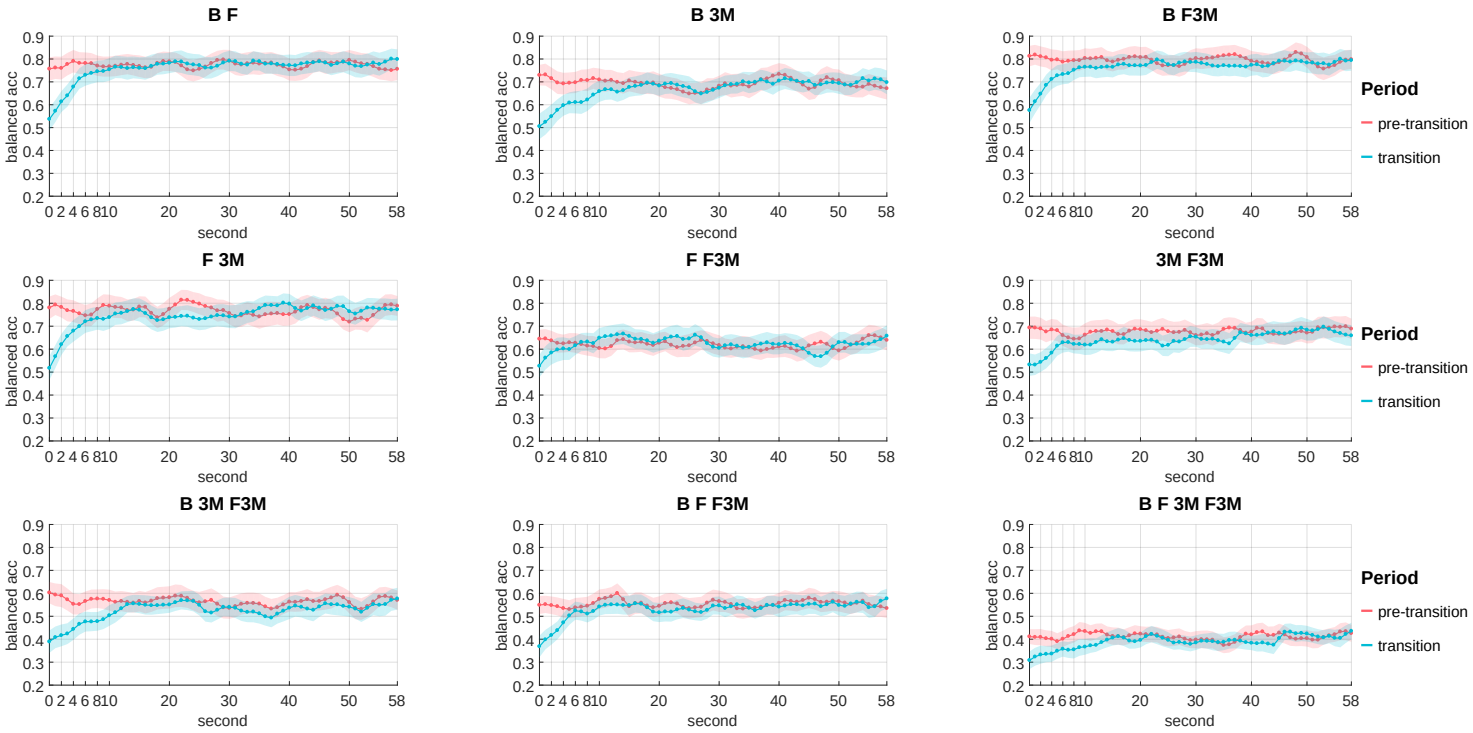


(a) Tested on the discrete phase

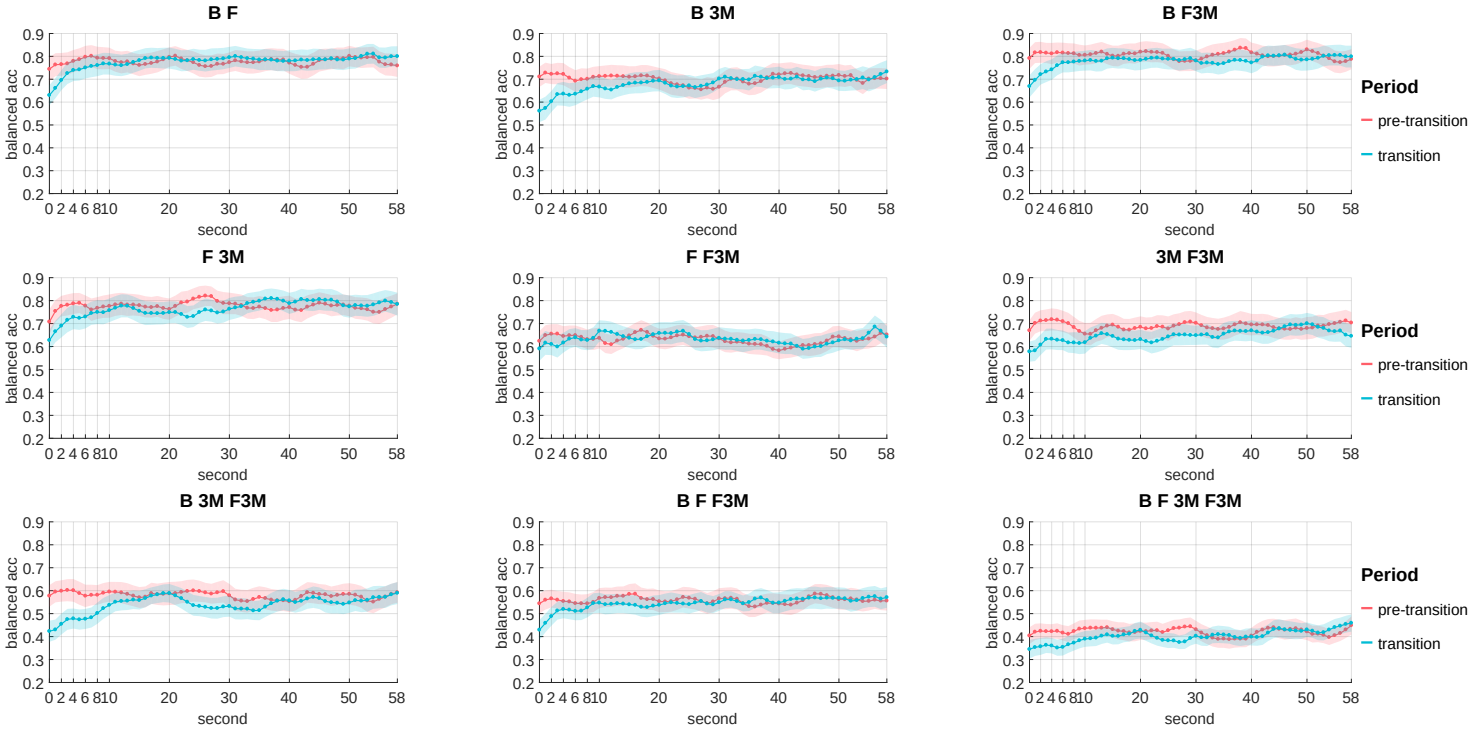


(b) Tested on the continuous phase

Figure 5.6 – Transition periods. Each curve represents the test accuracies of one of the defined periods, where the conditions were used as the labels. Each dot of test accuracy was averaged across 48 recordings tested on either (a) the discrete or (b) continuous phases.



(a) Forward processing



(b) Forward and Backward processing

Figure 5.7 – Estimation of total and intrinsic delay by windowed analysis. (a) is based on forward processing and thus represents the total delay. (b) uses forward and backward processing which can show the intrinsic delay. The accuracies were averaged across all subjects and folds.

over the left hemisphere and also the θ band around FCz for both F and 3M compared to B. As each task requires a different hand to operate, it likely implies that the pattern is not exclusively linked to motor actions but also conveys cognitive information. Although the α bands over the left hemisphere are not very similar between F and 3M compared to B, the strong α power modulations on the right hemisphere on the F and F3M tasks were consistent with Chapters 2 and 4 in which subjects performed the flying task with their right hand as opposed to this chapter. Consequently, the α bands at CP4 and C4 seem to reflect the perceived difficulty instead of motor activities in a visuomotor task. Another possibility would be reflecting the process of orientation.

The first analysis of transition (*c.f.* Chapter 5.5.4) clearly demonstrates the existence of transition periods after a task-change. The latency was further estimated with and without the effect of post-processing (*c.f.* Chapter 5.5.5). With the effect of post-processing, the total latency ranged around six to twelve seconds before the decoder can reliably provide accurate information about the cognitive state in the evaluated tasks. While without the effect of post-processing, the intrinsic latency is likely to be between four and eight seconds.

It is worth noting that, in Chapter 5.5.4, the stable periods have very similar results regardless of N. One may expect the test accuracy with $N = 0$ in the stable period to be lower, since the first few samples would actually come from the transition period and were more likely to be misclassified. Even the transition is confirmed, twelve seconds out of 180 or 300 seconds are rather small. This small amount of data from the transition period was unlikely to largely affect the built model and accuracies. Hence, the high accuracy with $N = 0$ does not contradict the hypothesis.

The estimation of latency in the transition periods can reflect the fact that the neural signature, and in turn the decoder, needs time to reliably reflect the cognitive state. Although the employment of forward and backward moving averaging eliminated the extrinsic latency from the decoder, the intrinsic latency can still be contributed by two factors. One factor is the behavioral reflection. It is plausible that the subject needs time to gather information about the new task and its difficulty, effectively delaying changes in her cognitive state. The second factor is from the onset that the subject gathered the necessary information to the moment that the decoder can reliably decode. 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 state estimation of users.

5.7 Conclusion

The proposed modeling of decoders still functions in the case of different tasks. However, the decoding accuracy still needs to be rigorously validated with a closed-loop experiment as Chapter 4.

The neural correlates of the flying task are consistent with Chapters 2 and 4 even using a different hand for controlling, suggesting that the decoders are more likely to capture cognitive states instead of motor activities, at least not those from hands.

For the second question regarding the latency, a complete transition analysis was performed. The transition period was confirmed and the latency excluding the contribution from signal processing was also estimated. This latency in the designed protocol is between four and eight seconds. This discovery suggests that designers of HMI systems should take this information into account when designing an interactive system. Moreover, this probably explains why the online decoding accuracy in Chapter 4 was much lower than the offline accuracy. In the online decoding, four waypoints formed a decision set which took roughly twelve seconds to pass. Twelve seconds are not much longer than the estimated four to eight seconds latency which already excluded the latency from post-processing. This argument, however, needs further examination, because the way to induce different difficulty levels is different and the change of difficulty level in Chapter 4 was in a small amount.

6 Conclusion

6.1 Contribution

This thesis has presented analyses based on three protocols, with a total of 123 recordings and involving 49 subjects, to prove the feasibility of building a cognitive state decoder based on EEG signals in order to fine-tune the human-machine interaction loop. The contributions are summarized below.

6.1.1 Regulating Cognitive States

The fundamental interaction of this thesis is based on the Yerkes-Dodson curve and the framework of challenge point, both showing an inverted U curve [YD08, GL04]. A user should reach her best state when the task difficulty level–highly linked to arousal state–is at the sweet spot. In both Chapter 2 and 4, two sets of online decoding experiments were conducted and were compared with presumed ground-truth conditions, in order to see how well a cognitive state decoder based on EEG signals can perform.

This thesis tried to identify and reach the sweet spot in two settings and both had a few cases where using an EEG decoder outperformed the presumed ground-truth conditions. Chapter 2 employed a bi-directional regulation with a continuous feedback. The subjects could always observe that the difficulty–radius of the waypoint–is changing all the time, either becomes easier or more difficult. In the ground-truth condition, it is clear that the distributions of radius mostly approximate an inverted U. In the case of EEG condition, a few subjects were able to yield similar shapes, and one subject even highly outperformed the ground-truth condition in terms of the task performance. The relatively low decoding accuracy was probably the reason that not many inverted U curves could be found. Another issue in this protocol was the fast-changing difficulty level–also orally confirmed by many subjects– which likely could also be intertwined with the issues of decoding latency being identified in Chapter 5.

Chapter 4, on the other hand, emphasized one-directional regulation from the left side of the

curve which is suitable for learning, while regulating from the opposite can be suitable for critical applications [FCSS19]. The difficulty this time only changed at a specific event which gave a rather stable difficulty level over around ten seconds. With an improved decoder, the offline accuracy was largely improved. Although the online decoding accuracy was not as high as the offline analysis, five and seven subjects out of thirteen in the two online sessions were able to yield similar behavioral results compared to the ground-truth conditions. In one session, one subject even outperformed the ground-truth condition with an EEG decoder which has a little bias towards more difficult levels. This probably implies that the subject, although being instructed, was still trying to stay in the comfort zone, and still have not reached the sweet spot. However, with an EEG decoder, the subject could not avoid leaving her comfort zone.

In sum, both chapters showed that implementing an EEG decoder inside the HMI loop has positive effects on different aspects. With the proposed decoding method, a few subjects could already benefit from this setting even when compared to the behavioral conditions.

6.1.2 Dynamics of Cognitive States

This thesis investigated the temporal dynamics of cognitive state when the objective difficulty level is changed. The result undermines the typical assumption of defining ground truth in cognitive state computing.

Typically, the ground truth of cognitive states is based on whether it is before or after an event or a stimulus, leading to an assumption of instant reaction from that the measured signals. It was plausible in some cases, especially when the signals directly measure the response of the central nervous system. For example, error-related potentials can reflect the awareness of error in less than one second [CSM14]. However, in Chapter 5, I had shown that, for the designed protocol, an intrinsic latency is observable in a scale of a few seconds, while excluding the effect from signal processing. As a result, there must be a certain kind of delay, either due to physiological adaptation or the reaction time of the subject.

This finding can have several impacts. First, designers of HMI system should take this latency into consideration when designing a system, especially when precise timing or quick interaction is required. Second, researchers working on cognitive state computing should carefully define ground truth. If it is a typical single-trial BCI study with a short duration in each trial, the transition period may largely impact the quality of decoding. In the conducted study, a trial lasted at least three minutes, which is long comparing to a few seconds that the transition effect may be neglected with sufficient training data.

Lastly, efficient HMI requires fast decoding of the current cognitive state of the subject. In this thesis, the used decoders basically have a total latency less than around 10 seconds.[1]

[1] The latency includes the length of the sliding window as well as all the temporal filtering in pre- and post-processing.

However, as pointed out in Chapter 1.1, some EEG-based studies still rely on tens of seconds in decoding cognitive states, although high efficiency is not strongly required in their applications [ABD+16a, WRB+17]. A high latency definitely can cause an issue in a cocktail party scenario, where EEG signals are used to decode to whom the user is attending [BDFB16]. In the report, the authors still used at least twenty seconds as a trial. That means, if the decoder is to be used on-the-fly to enhance a hearing-aid system, the user will face a delay of up to twenty seconds with a perfect decoder.

6.1.3 Decoding Framework

This thesis conducted a fair comparison between different data-driven decoders in Chapter 3, and the proposed two-stage decoder outperformed all the other conventional BCI decoders. The previously conducted survey covered plenty of methods [LCL+07, LBC+18]. Although a direct comparison of different methods is not their main purpose, the comparison was based on different sets of subjects and protocols. The discrepancies can make the comparison less reliable, as most studies did not involve dozens of subjects. On the contrary, this thesis evaluated different sets of methods, including commonly used ones, on the same set of data. The experimental results suggest two crucial factors to improve decoding accuracy. First, performing regression to the perceived difficulty levels is better than classifying. Second, post-processing with a low-pass filter on the regression responses, and likely class labels as well, is highly recommended as long as the latency is acceptable.

This thesis further conducted solid validations of the proposed two-stage decoder. The decoder had not been only validated on one data set but also in two other different experiments, including both online and offline validations. The high accuracies in both offline analyses suggest that the decoder is applicable when the subjects were doing different tasks, although more variations on the tasks are still worthy of investigation. Furthermore, as shown in Figure 4.8, misclassification of offline analysis in Chapter 4 mostly happened at a difficulty level that is around the boundary between the two classes. This implies that the regression captured the ordinal trend of the perceived difficulty level from EEG signals. However, either a better way to find the threshold for classification is needed or there were tiny differences between the features around that level. The online accuracy, although still needs to be improved, it is still better than the chance level.

6.1.4 Neural Correlates

This thesis provides new insight into the neural correlates of the perceived difficulty level. Previous literatures usually suggest or believe that either the engagement, attention, or similar index can properly reflect the corresponding cognitive states [PBB95, BLL+07, SM12, AG13, EFG16]. There exist statistical significances, but the performance is still questionable in real-time decoding. In Chapter 2, it has been shown that the data-driven approaches which only use linear combinations of spectral features outperformed the engagement and attention

indices that are non-linear. Moreover, the prominent locations of those indices are usually not clearly specified, or use the sum of power from all available EEG electrodes.

This thesis, on the contrary, identified rather consistent spectral features over two different tasks, three protocols, and totally involving forty-nine subjects. The features are (1) the α bands in the right hemisphere, specifically around C4 and CP4, and (2) the θ bands around Fz, FCz, and Cz. These findings do not necessarily overthrow the observations of the engagement and attention indices, but provide more concrete information as it was obtained from different tasks and subjects. Additionally, since one task required the subject to use their right hands while the other to use left hands for steering the simulated drone, it is rather plausible to declare these features are reflecting difficulty-related states at the grand-average level.

6.1.5 Publications

During the preparation of this thesis, the following articles, listed in chronological order, are published or submitted:

- F. I. T. Dell'Agnola, **P.-K. Jao**, A. Arza Valdes, R. Chavarriaga, J.d.R. Millán, F. Dario, and D. Atienza Alonso, " Machine-Learning Based Monitoring of Cognitive Workload in Rescue Missions with Drones," IEEE Transactions on Affective Computing, under review. This paper is not included in the thesis.
- **P.-K. Jao**, R. Chavarriaga, F. I. T. Dell'Agnola, A. Arza Valdes, D. Atienza Alonso, 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, under review. This paper essentially corresponds to Chapter 5.
- **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. This paper is adapted in Chapter 2 and incorporated in Chapter 3.
- **P.-K. Jao**, R. Chavarriaga, and J. d. R.Millán, "Using Robust Principal Component Analysis to Reduce EEG Intra-trial Variability," in 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2018. This paper is incorporated, although not fully developed, in the thesis.

6.2 Future Works

The conducted studies have demonstrated the feasibility of building an EEG decoder to decode cognitive states. Online decoding experiments also showed that the results were comparable to behavioral conditions, and sometimes can be even better. However, online decoding accuracy is still to be improved. One can explore different decoders with a cross-day validation, which

reflects the real scenario. Another option is conducting online and offline experiments on the same day, which may be a too heavy experiment for the subjects. The new experiment should also carefully consider the identified latency issue in Chapter 5, to see if accuracy can be improved. On the other hand, mutual learning between the subject and decoder should also be considered. It has been shown that with more training time or proper learning rates, the overall performance can be raised [MVS+17, PTS+18]. Even though in the conducted experiments, the subjects were not supposed to actively modulate their brain signals to have the desired decoding outcome. It does not obey the nature of humans. Some subjects must have tried to learn the behavior of their decoders. Therefore, future experiments should also consider the case of explicitly regulating their cognitive states as the study of Faller *et al.* [FCSS19] and evaluate the behavioral outcome.

Another direction is establishing a more solid method to define functional task difficulty of the designed task, and consequently defining three regions, Easy, Sweet Spot, and Hard. Defining the sweet spot can be a challenge, but once properly defined, another bi-directional experiment with a three-class decoder to avoid quick oscillations of difficulty level can be conducted to see if there is more benefit than one-directional regulation.

In the case of inducing difficulty levels with different tasks, an online experiment is still needed to verify the decoding accuracy.

The thesis is more prone to the questions of the engineering side while the questions of the neuroscience side are also worthy of investigation. Future work can involve a deeper analysis of the neural correlates.

One direction is probing into the subject level on which areas are involved and their temporal patterns rather than the accuracy analysis around the onset. There can be some correlation between EEG signals and the aiming error which is the planar distance between the aiming cross and the center targeted waypoint, as the error on a tracking task has correlation [Coh16].

Another direction can be examining potential correlations between the EEG-decoded cognitive states and other physiological signals. A perfect correlation is unlikely to be found. However, if a high correlation is identified, it can lead to other interesting questions. For example, what is the causality between the two and what kind of extra information from the uncorrelated part can be augmented for decoding the cognitive states?

The least but not the last direction is identifying the actual cognitive states related to the perceived difficulty level. They can be states arousal, attention, stress, engagement, error awareness *etc.*

Bibliography

[ABD+15] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Alfredo Colosimo, Ilenia Graziani, Jean-Paul Imbert, Géraud Granger, Railene Benhacene, Michela Terenzi, Simone Pozzi, and Fabio Babiloni. Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks. In *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2015.

[ABD+16a] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Alfredo Colosimo, Stefano Bonelli, Alessia Golfetti, Simone Pozzi, Jean-Paul Imbert, Géraud Granger, Raïlane Benhacene, and Fabio Babiloni. Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment. *Frontiers in Human Neuroscience*, 10:539, 2016.

[ABD+16b] Pietro Aricò, Gianluca Borghini, Gianluca Di Flumeri, Alfredo Colosimo, Simone Pozzi, and Fabio Babiloni. Chapter 10 - A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. In Damien Coyle, editor, *Brain-Computer Interfaces: Lab Experiments to Real-World Applications*, volume 228 of *Progress in Brain Research*, pages 295–328. Elsevier, 2016.

[ACZG08] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE Int. Joint Conf. on Neural Networks*, 2008.

[ADT12] Mathieu Andrieux, Jérémy Danna, and Bernard Thon. Self-control of task difficulty during training enhances motor learning of a complex coincidence-anticipation task. *Research Quarterly for Exercise and Sport*, 83(1):27–35, 2012.

[AG13] Marvin Andujar and Juan E. Gilbert. Let's learn!: enhancing user's engagement levels through passive brain-computer interfaces. In *CHI Extended Abstracts on Human Factors in Computing Systems*, 2013.

[APGvG10] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010.

[BAV+14] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[BDFB16] Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5):402–412, 2016.

[BHL+18] Elena Boto, Niall Holmes, James Leggett, Gillian Roberts, Vishal Shah, Sofie S Meyer, Leonardo Duque Muñoz, Karen J Mullinger, Tim M Tierney, Sven Bestmann, Gareth R. Barnes, Richard Bowtell, and Matthew J. Brookes. Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, 555(7698):657, 2018.

[BLL+07] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, 78(Supplement 1):B231—-B244, 2007.

[BPL03] Roland Brunken, Jan L Plass, and Detlev Leutner. Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1):53–61, 2003.

[BPP85] Olivier M. Bertrand, François M. Perrin, and Jacques Pernier. A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 62(6):462–464, 1985.

[BRS+14] Cathrin M Buetefisch, Kate Pirog Revill, Linda Shuster, Benjamin Hines, and Michael Parsons. Motor demand-dependent activation of ipsilateral motor cortex. *Journal of Neurophysiology*, 112(4):999–1009, 2014.

[BWS96] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42(3):361–377, 1996.

[CBB17] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

[CL11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Coh16] Michael X Cohen. Midfrontal theta tracks action monitoring over multiple interactive time scales. *NeuroImage*, 141:262–272, 2016.

[CPF16] Mickael Causse, Vsevolod Peysakhovich, and Eve Florianne Fabre. High working memory load impairs the processing of linguistic stimuli during a simulated piloting task: an ERP and pupillometry study. *Frontiers in Human Neuroscience*, 10(240), 2016.

[CRBP11] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(6):1052–1063, 2011.

[CSM14] Ricardo Chavarriaga, Aleksander Sobolewski, and José del R Millán. Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*, 8, 2014.

[DB08] Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2008.

[DCA18] Fabio Dell'Agnola, Leila Cammoun, and David Atienza. Physiological characterization of need for assistance in rescue missions with drones. In *IEEE Int. Conf. on Consumer Electronics*, 2018.

[Dic05] Michele D Dickey. Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development*, 53(2):67–83, 2005.

[DLWK11] Moritz Dannhauer, Benjamin Lanfer, Carsten H Wolters, and Thomas R Knösche. Modeling of the human skull in eeg source analysis. *Human Brain Mapping*, 32(9):1383–1399, 2011.

[DMH+14] Sven Dähne, Frank C Meinecke, Stefan Haufe, Johannes Höhne, Michael Tangermann, Klaus-Robert Müller, and Vadim V Nikulin. SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, 86:111–122, 2014.

[DSBMP15] Ian Daly, Reinhold Scherer, Martin Billinger, and Gernot Muller-Putz. FORCe: Fully Online and Automated Artifact Removal for Brain-Computer Interfacing. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(5):725–736, 2015.

[EFG16] Kate C Ewing, Stephen H Fairclough, and Kiel 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*, 10:223, 2016.

[FCSS19] Josef Faller, Jennifer Cummings, Sameer Saproo, and Paul Sajda. Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task. *Proc. of the National Academy of Sciences*, 116(13):6482–6490, 2019.

[Fow94] Barry Fowler. P300 as a measure of workload during a simulated aircraft landing task. *Human Factors*, 36(4):670–683, 1994.

[Fri97] Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[FWM+08] Mehrdad Fatourechi, Rabab K. Ward, Steven G. Mason, Jane Huggins, Alois Schlögl, and Gary E. Birch. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *7th Int. Conf. on Machine Learning and Applications*, 2008.

[GCM11] Gangadhar Garipelli, Ricardo Chavarriaga, and José del R Millán. Single trial recognition of anticipatory slow cortical potentials: the role of spatio-spectral filtering. In *IEEE Int. Conf. on Neural Engineering*, 2011.

[GEF+15] Uwe Graichen, Roland Eichardt, Patrique Fiedler, Daniel Strohmeier, Frank Zanow, and Jens Haueisen. SPHARA-a generalized spatial fourier analysis for multi-sensor systems with non-uniformly arranged sensors: Application to EEG. *PloS one*, 10(4):e0121741, 2015.

[GL04] Mark A Guadagnoli and Timothy D Lee. Challenge Point: A Framework for Conceptualizing the Effects of Various Practice Conditions in Motor Learning. *Journal of Motor Behavior*, 36(2):212–224, 2004.

[GMSW13] Mateusz Gola, Mikołaj Magnuski, Izabela Szumska, and Andrzej Wróbel. EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects. *Int. Journal of Psychophysiology*, 89(3):334–341, 2013.

[GTH+08] David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P.N. Rao. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2008.

[Har78] Fredric J Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proc. of the IEEE*, 66(1):51–83, 1978.

[HDP+18] Matar Haller, Thomas Donoghue, Erik Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra. *bioRxiv*, 2018.

[HHA18] Jamison Heard, Caroline E Harriott, and Julie A Adams. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, 48(5):434–451, 2018.

[HS88] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

[HVE06] Ulrich Hoffmann, Jean-Marc Vesin, and Touradj Ebrahimi. Spatial filters for the classification of event-related potentials. In *Proc. of the European Symposium on Artificial Neural Networks*, 2006.

[JCM18a] Ping-Keng Jao, Ricardo Chavarriaga, and José del R. Millán. Analysis of EEG correlates of perceived difficulty in dynamically changing flying tasks. In *IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2018.

[JCM18b] Ping-Keng Jao, Ricardo Chavarriaga, and José del R Millán. Using robust principal component analysis to reduce EEG intra-trial variability. In *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2018.

[JZBJ02] M Javorka, I Zila, T Balharek, and K Javorka. Heart rate recovery after exercise: relations to heart rate variability and complexity. *Brazilian Journal of Medical and Biological Research*, 35(8):991–1000, 2002.

[Kli99] Wolfgang Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2–3):169–195, 1999.

[KOL15] Jun-Su Kang, Amitash Ojha, and Minho Lee. Concentration Monitoring with High Accuracy but Low Cost EEG Device. In *Int. Conf. on Neural Information Processing*, 2015.

[LBC+18] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[LCL+07] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for EEG-based computer interfaces. *Journal of Neural Engineering*, 4:R1–R13, 2007.

[LJY17] Yuan-Pin Lin, Ping-Keng Jao, and Yi-Hsuan Yang. Improving cross-day EEG-based emotion classification using robust principal component analysis. *Frontiers in Computational Neuroscience*, 11:64, 2017.

[LSC04] Steven Lemm, Christin Schafer, and Gabriel Curio. Bci competition 2003-data set iii: probabilistic modeling of sensorimotor/spl mu/rhythms for classification of imaginary hand movements. *IEEE Transactions on Biomedical Engineering*, 51(6):1077–1080, 2004.

[MAC15] Mark Mulder, David A Abbink, and Tom Carlson. Introduction to the special issue on shared control: applications. *Journal of Human-Robot Interaction*, 4(3):1–3, 2015.

[MBJS96] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence J Sejnowski. Independent component analysis of electroencephalographic data. In *Advances in Neural Information Processing Systems*, 1996.

[MBL+00] Scott Makeig, T Bell, TW Lee, TP 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.

[MC10] José del R Millán and Jose M Carmena. Invasive or noninvasive: Understanding brain-machine interface technology [conversations in bme]. *IEEE Engineering in Medicine and Biology Magazine*, 29(1):16–22, 2010.

[Mil04] José del R. Millán. On the need for on-line learning in brain-computer interfaces. In *IEEE Int. Joint Conf. on Neural Networks*, volume 4, 2004.

[ML11] James SP Macdonald and Nilli Lavie. Visual perceptual load induces inattentional deafness. *Attention, Perception, & Psychophysics*, 73(6):1780–1789, 2011.

[MMM14] Filip Melinscak, Luis Montesano, and Javier Minguez. Discriminating between attention and mind wandering during movement using EEG. In *Proc. 6th Int. Brain–Computer Interface Conf.*, 2014.

[MO18] Pierre Morel and Others. Gramm: grammar of graphics plotting in Matlab. *Journal of Open Source Software*, 3(23):568, 2018.

[Mor08] Christine S Moravec. Biofeedback therapy in cardiovascular disease: Rationale and research overview. *Cleveland clinic journal of medicine*, 75(2):S35, 2008.

[MRMH11] Matthew W Miller, Jeremy C Rietschel, Craig G McDonald, and Bradley D Hatfield. A novel approach to the physiological measurement of mental workload. *Int. Journal of Psychophysiology*, 80(1):75–78, 2011.

[MVS+17] Jan Saputra Müller, Carmen Vidaurre, Martijn Schreuder, Frank C Meinecke, Paul Von Bünau, and Klaus-Robert Müller. A mathematical model for the two-learners problem. *Journal of Neural Engineering*, 14(3):036005, 2017.

[NSKDB16] Laura Naumann, Matthias Schultze-Kraft, Sven Dähne, and Benjamin Blankertz. Prediction of difficulty levels in video games from ongoing EEG. In *Int. Workshop on Symbiotic Interaction*, 2016.

[PB37] Wilder Penfield and Edwin Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, 60(4):389–443, 12 1937.

[PBB95] Alan T Pope, Edward H Bogart, and Debbie S Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(1–2):187 – 195, 1995.

[PBLM11] Serafeim Perdikis, Hamidreza Bayati, Robert Leeb, and José del R Millán. Evidence accumulation in asynchronous BCI. *Int. Journal of Bioelectromagnetism*, 13:131–132, 2011.

[Per18] Michael Eric Anthony Pereira. *Neural correlates of performance monitoring during discrete and continuous tasks*. EPFL, Lausanne, 2018.

[PTS+18] Serafeim Perdikis, Luca Tonin, Sareh Saeedi, Christoph Schneider, and José del R Millán. The cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users. *PLoS Biology*, 16(5):e2003787, 2018.

[PVAG+14] Peter Putman, Bart Verkuil, Elsa Arias-Garcia, Ioanna Pantazi, and Charlotte 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*, 14(2):782–791, 2014.

[RSAG09] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009.

[RSdM+16] Noortje H Rijken, Remko Soer, Ewold de Maar, Hilco Prins, Wouter B Teeuw, Jan Peuscher, and Frits G J Oosterveld. Increasing Performance of Professional Soccer Players and Elite Track and Field Athletes with Peak Performance Training and Biofeedback: A Pilot Study. *Applied Psychophysiology and Biofeedback*, 41(4):421–430, 2016.

[RWAH16] Michal Rapczynski, Philipp Werner, and Ayoub Al-Hamadi. Continuous low latency heart rate estimation from painful faces in real time. *23rd Int. Conf. on Pattern Recognition*, 2016.

[SDH82] William M Savin, Dennis M Davidson, and William L Haskell. Autonomic contribution to heart rate recovery from exercise in humans. *Journal of Applied Physiology*, 53(6):1572–1575, 1982.

[SM12] Daniel Szafir and Bilge Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, 2012.

[SS09] Mathew Salvaris and Francisco Sepulveda. Visual modifications on the p300 speller BCI paradigm. *Journal of Neural Engineering*, 6(4):046011, 2009.

[SSJS16] Sameer Saproo, Victor Shih, David C Jangraw, and Paul Sajda. Neural mechanisms underlying catastrophic failure in human–machine interaction during aerial navigation. *Journal of Neural Engineering*, 13(6):066005, 2016.

[Tho82] David J Thomson. Spectrum estimation and harmonic analysis. *Proc. of the IEEE*, 70(9):1055–1096, 1982.

[Tre95] Sven Treitel. Spectral analysis for physical applications: multitaper and conventional univariate techniques. *American Scientist*, 83(2):195–197, 1995.

[VDA+05] Timothy Verstynen, Jorn Diedrichsen, Neil Albert, Paul Aparicio, and Richard B Ivry. Ipsilateral motor cortex activity during unimanual hand movements relates to task complexity. *Journal of Neurophysiology*, 93(3):1209–1222, 2005.

[Vig97] Ricardo Nuno Vigário. Extraction of ocular artefacts from eeg using independent component analysis. *Electroencephalography and Clinical Neurophysiology*, 103(3):395–404, 1997.

[VPV11] Roland Vocat, Gilles Pourtois, and Patrik Vuilleumier. Parametric modulation of error-related ERP components by the magnitude of visuo-motor mismatch. *Neuropsychologia*, 49(3):360–367, 2011.

[VRE+17] Gaël Varoquaux, Pradeep Reddy Raamana, Denis A Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, 2017.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987.

[WGG06] Yijun Wang, Shangkai Gao, and Xiaornog Gao. Common spatial pattern method for channel selelction in motor imagery based brain-computer interface. In *Annual Conf. of the IEEE Engineering in Medicine and Biology*, 2006.

[WRB+17] Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, Peter Gerjets, and Martin Spüler. Online EEG-Based Workload Adaptation of an Arithmetic Learning Environment. *Frontiers in Human Neuroscience*, 11:286, 2017.

[Wri08] Rex A Wright. Refining the Prediction of Effort: Brehm's Distinction between Potential Motivation and Motivation Intensity. *Social and Personality Psychology Compass*, 2(2):682–701, 2008.

[YD08] Robert M Yerkes and John D Dodson. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482, 1908.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

PING-KENG JAO

Brain-Machine Interface | Data Scientist | Digital IC Designer

@ pingkeng.jao@gmail.com in linkedin.com/in/ping-keng-jao Geneva, Switzerland Taiwan

EXPERIENCE

École Polytechnique Fédérale de Lausanne, CNBI

Doctoral Assistant

Feb 2016 – Jan 2020 Geneva, Switzerland

- Working on decoding cognitive states to dynamically optimize task difficulty.
  - design and build protocols with Unity (C#) and Python.
  - record EEG and EOG signals mainly with a Biosemi system.
  - process EEG and EOG signals with MATLAB.
  - conduct closed-loop experiments with real-time decoders.
- Acted as a teaching assistant for "Brain-Computer Interaction" and "Data Analysis and Model Classification" classes, each for two semesters.
- Supervised 2 Master-level projects.
- Published 2 IEEE conference papers, working on 3 IEEE journal papers.

Academia Sinica, Music and Artificial Intelligence Lab.

Full-time Research Assistant

Feb. 2013 – July 2015 Taipei, Taiwan

- **Brain-Computer Interface based Sound Source Separation**: Developed a MATLAB prototype to control the volume between left & right channels.
- **Sound Source Separation**: Improved 2 dB in source-to-distortion ratio by convolutional sparse coding in a setting of score-informed monaural audio.
- **Dictionary-based Music Genre Retrieval System**: Achieved state-of-the-art performance and 8X acceleration with MATLAB by a screening method.
- Organized a reliable and high-performance computing environment based on NAS.
- Contributed to calibration of EEG signals in a cross-day setting.
- Published 5 IEEE/ACM papers, 1 IEEE sponsored paper, and 1 workshop paper within 2.4 years.

Ministry of Defense, Army Special Force Command

Company Chief Counselor, Second Lieutenant

Oct 2011 – Sep 2012 Taiwan

- Completed airborne training.

EDUCATION

Ph.D. in Electrical Engineering

Mar 2020 (Expected)

École Polytechnique Fédérale de Lausanne, CNBI Switzerland

- Thesis: Decoding Cognitive States under Varying Difficulty Levels.

M.S. in Electrical Engineering

Sept 2009 – June 2011

National Cheng Kung University, Smart EDA Taiwan

- Thesis: "An Effective Complex Obstacle-Avoiding Rectilinear Steiner Tree Construction Algorithm and its Application in Congestion Driven Routers", done in C/C++

B.S. in Electrical Engineering

Sept 2005 – June 2009

National Cheng Kung University Taiwan

- GPA: 90.55/100, Department Rank: 4/144.
  Done several IC design projects and completed algorithm-related courses.

Version:03/02/2020

SUMMARY

Experienced data scientist specialized in EEG and music signals. For the future, the main interest is, but not limited to, decoding cognitive state during conduction of auditory tasks and building relevant BMI systems. Electrical engineering background contributes to skills of hardware development.

STRENGTHS

Machine Learning | Signal Processing | Brain–Machine Interface | EEG Analysis | Music Signal Analysis | Digital IC Design & Automation

Adamant | Fairness | Hard-working | Independent | International Environment | Responsible | Team Player

PROGRAMMING

MATLAB	7 yr	Data analysis
C/C++	4 yr	IC, Protocol, Algorithm
Python	3 yr	Protocol
LaTeX	7 yr	Papers, CV
Verilog	2 yr	CPU, Display Controller

LANGUAGES

Taiwanese	C2	C1
Mandarin	C2	C2
English	C1/C2	C1/C2
French	A2/B1	B1

MOST PROUD OF

Being recognized
as an outstanding reviewer of Journal of Neural Engineering.

Courage I had
to complete 5 times of parachuting even I consider myself having acrophobia.

Increasing ski level
as a novice to simple black level within a season.

PUBLICATIONS

Journal Articles

P.-K. Jao, R. Chavarriaga, F. I. T. Dell'Agnola, A. Arza Valdes, D. Atienza Alonso, and J.d.R. Millán, "EEG Correlates of Difficulty Levels in Dynamical Transitions of Simulated Flying and Mapping Tasks," *IEEE Trans. on Human-Machine Systems*, under major revision.

F. I. T. Dell'Agnola, **P.-K. Jao**, A. Arza Valdes, R. Chavarriaga, J.d.R. Millán, F. Dario, and D. Atienza Alonso, "Machine-Learning Based Monitoring of Cognitive Workload in Rescue Missions with Drones," *IEEE Trans. on Affective Computing*, under preparation.

P.-K. Jao, R. Chavarriaga, and J.d.R. Millán, "Learning to Fly with a BMI Instructor," under preparation.

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, p. 64, 2017.

P.-K. Jao, L. Su, Y.-H. Yang, and B. Wohlberg, "Monaural music source separation using convolutional sparse coding," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2158–2170, 2016.

P.-K. Jao and Y.-H. Yang, "Music annotation and retrieval using unlabeled exemplars: Correlation and sparse codes," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1771–1775, 2015.

G. G. Lee, M.-J. Wang, B.-H. Chen, J. Chen, **P.-K. Jao**, C. J. Hsiao, and L.-F. Wei, "Reconfigurable architecture for deinterlacer based on algorithm/architecture co-design," *Journal of Signal Processing Systems*, vol. 63, no. 2, pp. 181–189, 2011.

Conference Proceedings

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'l. Conf. on Systems, Man, and Cybernetics*, 2018.

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'l. Conf. of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2018.

P.-K. Jao, P.-I. Chen, and Y.-H. Yang, "Disk jockey in brain-a prototype for volume control of tracked instrument during playback," in *Proc. Int'l. Works. Brain-Computer Music Interfacing*, 2015.

P.-K. Jao, Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, "Using robust principal component analysis to alleviate day-to-day variability in EEG based emotion classification," in *37th Annual Int'l. Conf. of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2015.

P.-K. Jao, Y.-H. Yang, and B. Wohlberg, "Informed monaural source separation of music based on convolutional sparse coding," in *IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing*, IEEE, 2015.

P.-K. Jao, C.-C. M. Yeh, and Y.-H. Yang, "Modified LASSO screening for audio word-based music classification using large-scale dictionary," in *IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing*, IEEE, 2014.

C.-C. M. Yeh, **P.-K. Jao**, and Y.-H. Yang, "Awtoolbox: Characterizing audio information using audio words," in *Proc. of the 22nd ACM Int'l. Conf. on Multimedia*, ACM, 2014.

P.-K. Jao, L. Su, and Y.-H. Yang, "Analyzing the dictionary properties and sparsity constraints for a dictionary-based music genre classification system," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, IEEE, 2013.

Patents

G. G. C. LEE, H.-Y. Lin, C.-F. Chen, and **P.-K. Jao**, *Method for merging the regions in the image/video*, US Patent 8,948,510, Feb. 2015.

REFERENCES

On request.

RECREATIONS

Computer Games | Reading | Scuba Diving | Sketch | Ski | Snorkeling | Hiking | Ukulele