POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

A portable EEG-based Brain-Computer Interface for Imagined Speech Detection: towards an Assistive System for Restoring Communication

MASTER THESIS IN BIOMEDICAL ENGINEERING

Cavallini Federico, 10619764

Advisor:
Professor Emilia Ambrosini

Co-advisors:
Professor Marc Van Hulle

Aurélie De Borman

Bob Van Dyck

Academic year:
2022-2023

Abstract: Pathologies such as strokes and amyotrophic lateral sclerosis can lead to the loss of verbal communication while leaving cognitive abilities intact. Brain-Computer Interfaces (BCIs) based on imagined speech (IS) paradigms offer a potential solution to restore verbal communication, allowing for a more natural interaction. These systems aim to decode what a person imagines saying without any corresponding body movements. The primary challenge with IS lies in the absence of a measurable motion to use as ground truth independent from brain signals, making the design of experimental paradigms critical. Previous studies have approached this challenge with varying success, using both invasive and non-invasive techniques to decode imagined words within a limited vocabulary. The most diffused set-up is 64-channel clinical-grade EEG; in this study we investigate IS using a portable EEG device featuring only 8 channels. We focused on detecting whether a subject is imagining saying a word or not, framing it as a binary classification task. Two paradigms were compared to analyse how different instructions influence the subject's imagination process and to address the challenge of the lack of ground truth. We exploited the gained insight to design offline experiments achieving an average prediction accuracy of 65.3% over six subjects, with peaks of 74% and 86.6% for two of them. Finally, we transferred the model to online settings providing neurofeedback to the users. In two out of three subjects, online feedback accuracy surpassed chance level, reaching 77.8% and 88.9% during the final stages of the session, demonstrating a learning process synergically involving both the user and the BCI system.

Keywords: BCI, Imagined Speech, Neurofeedback, EEG, Portability, Machine Learning

1. Introduction: Brain Computer Interface

«Something like a giant invisible diving bell holds my whole body prisoner», but not his mind. However, when «my diving bell becomes less oppressive, [...] my mind takes flight like a butterfly». This is how Jean-Dominique Bauby describes his own condition in the autobiographical novel *The*

Diving Bell and the Butterfly [1], where he narrates his experience after that a cortico-subcortical stroke led him to a complete motor paralysis but with intact sensory and cognitive functions. Making that diving bell lighter is the aim of Brain Computer Interfaces (BCIs): to create an alternative communication pathway between the brain and an external word circumventing the impaired or lost natural functionalities.

«A BCI provides its user with an alternative method for acting on the world» [2]. It is a computer-based communication system that translates the brain activity into commands for an external device. Brain signals are acquired, analysed and translated into commands to operate an actuator to carry out a desired action or task. Since it only uses signals coming from the central nervous system (SNC) and not physiological brain peripheral output pathways (nerves and muscles), it differs from other assistive devices. BCIs can help people with severe motor disabilities to interact with the enviroment, communicate and enhance independence. In general, they can provide improvement of the quality of life of neurologically impaired patients suffering from various pathologies such as amyotrophic lateral sclerosis (ALS), stroke, brain/spinal cord injury, muscular dystrophy, etc. Other uses are also possible such as gaming, entertainment and military applications.

BCI systems can be classified according to different features. The primary classification revolves around the recording device used, distinguishing between invasive – *e.g.*, microelectrode arrays and electrocorticography (ECoG) – and non-invasive approaches – commonly using electroencephalography (EEG), magnetoencephalography (MEG), as well as functional near infrared spectroscopy (fNIRS) and functional magnetic resonance imaging (fMRI). Beyond this fundamental distinction, several other characteristics help describe a BCI:

- Synchronous (or cue guided)/Asynchronous (or self-paced) [3]: synchronous BCIs require users to adhere to a fixed repetitive scheme for switching between mental tasks. The phenomena to be recognized are time-locked to a cue. Asynchronous BCIs allow subjects to voluntarily decide when to transition from one mental task to the next.
- Active/Reactive/Passive [4]: active BCIs decode outputs directly from brain activity that users actively control, independently of external events. These often rely on asynchronous systems. Reactive BCIs derive outputs from brain activity elicited in response to external stimulation, such as paradigms based on P300 Event-Related Potentials (ERPs). Passive BCIs do not require voluntary user control but are used to monitor implicit information about the user's current state, such as emotional state or fatigue levels.
- Exogenous/Endogenous [5]: in exogenous BCIs an external stimulation generates a specific brain activity to be decoded, while in endogenous BCIs do not relay on any external stimulation and the system recognizes the specific mental states from self-generated brain signal. An example of exogenous BCI is SSVEP-based paradigms, where the flickering of a target at a certain rhythm induces brain oscillations at the same rhythm, making it detectable. Instead, endogenous BCIs can be employed for motor imagery paradigms.

Like any communication or control system, a BCI consists of input – which includes electrophysiological activity from the user –, output – involving device commands –, components for translating input to output and a protocol defining the governing communication rules [2]. Originally, it was the *subject learning* to voluntarily regulate their brain activity by means of neurofeedback; subsequently, *machine learning* techniques facilitated the identification of statistical brain state signatures during a calibration session. But the most effective modality to create a working system is to allow the interaction of the two adaptive controllers, the user and the system

[6]. The user needs to establish a good correlation between their intent and the signal features utilized by the BCI. Simultaneously, the BCI must accurately and efficiently select, extract and translate these features into device commands [2]. The general architecture of a BCI system is depicted in Figure 1.

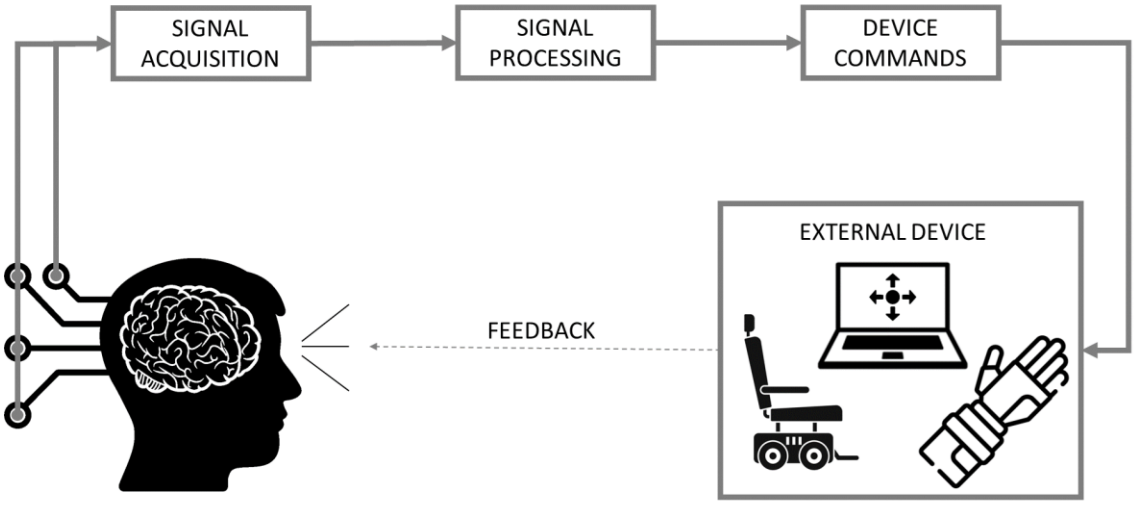

Figure 1: Basic design of any BCI system: brain signals are acquired, processed and translated into commands to control an external device. Success depends on the interaction of two adaptive controllers – the user and the system (adapted from [2]).

1.1 Imagined Speech BCI

Verbal communication is the natural way of human interaction. Various conditions, such as stroke and ALS, can result in the loss of speech articulation (anarthria), while preserving language skills and cognitive abilities [7]. Speech BCIs have the potential to offer individuals with paralysis a means of rapid communication by deciphering neural signals generated during speech-related tasks (attempted speech, silent speech, imagined speech, inner speech) into text or sound [8]. Two main strategies for decoding speech can be identified: discrete and continuous decoding. Continuous speech BCIs aim to predict a spectrogram that can be converted into audible speech, providing a more natural and flexible output. In contrast, discrete decoding involves generating output in the form of text, words, phonemes, or vowels within a fixed set of possibilities. Continuous decoding allows for a broad range of vocabulary sizes, as it can capture the nuances of natural speech. Discrete decoding, on the other hand, limits the degrees of freedom, which reduces the ability to scale to larger vocabulary sizes but simplifies the prediction process.

Different speech-related modalities are investigated and tested. Beyond the performed speech modality which can be studied only in healthy subjects, the attempted speech modality is also used to build BCI systems to restore basic communication in patients who cannot produce intelligible speech but can vocalize in the form of grunts and moans. These BCIs have achieved partial communication restoration through invasive recording techniques, such as ECoG [9] and the recording of spiking activity via intracortical microelectrode arrays [8]. Additionally, other speech modalities that do not involve acoustic output have been investigated [10]:

- silent speech (or mimed speech): speakers are directed to mimic the movements involved in typical speech production while deliberately preventing the release of audible sound. Silent

speech can be assessed through the observation of articulatory gestures using motion-capture technology, various imaging methods or by tracking muscle activity through electromyography (EMG).

- imagined speech (IS): similar to silent speech but it also involves the inhibition of articulatory movements. Subjects are asked to imagine speaking but without moving any body part. In this context, imagined speech closely resembles the first-person motor imagery of speaking, wherein individuals are encouraged to experience the sensation of speaking rather than mere internal dialogue. Since imagined speech occurs without any actual articulatory movements, it necessitates examination only at the neural level.
- inner speech: among the different definitions of it, Vygotsky's model [11] characterizes inner speech as an internalized thought process focused on abstract meanings. Unlike imagined and silent speech, inner speech lacks phonological neural elements and the conversational dynamics of external dialogues. Consequently, the study of inner speech, even from a neural perspective, poses greater challenges.

Of particular interest are the last two modalities, which can be executed similarly by both healthy individuals and paralyzed patients. This enables an easier transfer of information from healthy subjects experiments to the development of rehabilitation and restoration applications [10]. However, while performed and attempted speech generate audio signals that can be measured, and in silent speech movements can be tracked through motion sensors or EMG electrodes, in contrast, imagined and inner speech can only be observed at neural level, without an independent signal serving as a ground truth. This lack of a reference signal adds complexity to the challenge of utilizing imagined speech for BCIs. Nevertheless, many researchers are addressing this paradigm through different techniques for its transferability from healthy subjects to patients.

Although invasive techniques like ECoG have demonstrated higher classification accuracy in decoding imagined speech, EEG, being non-invasive, is the most commonly employed method for its investigation. EEG does come with inherent limitations, including lower signal-to-noise ratio (SNR) compared to ECoG, the presence of artifacts, and restricted spectral and spatial resolution [12]. However, it maintains a good temporal resolution. The primary advantages of this technique are its low cost, safety and portability [13], [14]. These advantages can be further enhanced by employing a portable setup with a reduced number of channels using dry electrodes [14].

1.2 Intent and thesis outline

This thesis aims to take initial steps towards developing a non-invasive imagined speech (IS) BCI using an 8-channel lightweight and portable EEG setup: Mentalab Explore+. Such a system holds the potential to facilitate communication restoration for individuals who have lost the ability to articulate speech. The ultimate project goal is to create a system capable of decoding four directional commands (“UP”, “DOWN”, “LEFT”, “RIGHT”) and one confirmation command (“SELECT”). This system would enable communication for controlling a mobile assistive device or a graphical interface. To be fully functional and practical, the system must operate in online settings as an asynchronous, endogenous, active BCI.

The first necessary step is to detect whether the user is performing some speech imagination or not. This is the objective of this thesis. In order to address the detection problem, limited command vocabulary (only “LEFT” and “RIGHT”) is used and to cope with the huge issue of lacking a ground truth signal, a cued system (synchronous) is employed. This exploratory research focuses

exclusively on healthy subjects, and the paradigm is initially tested in offline settings. Recognizing the potential benefits of closing the BCI loop through neurofeedback, an online experiment providing feedback to users is then also conducted.

In Section 2, a general overview of EEG-based BCI systems is followed by an extensive and critical analysis of the state-of-the-art of IS BCIs. This analysis is used to define an appropriate taxonomy for IS paradigms and establish realistic performance benchmarks. Indeed, the section identifies issues in the performance reporting process and proposes strategies for more accurate and fair reporting. Section 3 provides the description of the implementation choices made for this work, including data collection, paradigm design, signal inspection and classification pipeline development. The subsequent three sections describe three consecutive phases of the work. Chapter 4 presents a pilot study conducted on a single subject. It was a long iterative process aimed at gaining confidence with the tools necessary for implementing the BCI system and to address challenges in the IS paradigm. Important aspects like the experimental protocol and model architecture are explored. Chapter 5 details the main experimental procedure of this thesis, focused on developing a model capable of offline IS detection using data from six different healthy subjects. EEG signals are examined, and model performance is reported and compared with existing state-of-the-art. In chapter 6, it is described the implementation of an online paradigm, which was tested on three healthy subjects. During those experiments, IS is detected in real-time by an adaptive model and feedback about the predictions is provided to the user to close the BCI loop, facilitating user adaptation and simultaneous learning. In the last sections, conclusions about the whole work are drawn and the bibliographic references are reported.

2. Literature review

2.1 EEG-based BCI pipeline [12], [14], [15]

The aim of BCIs is to extract a command from the brain signals to operate a device. The necessary signal processing pipeline depends on how the problem is defined and which acquisition technique is employed. The aim of this thesis is to develop a discrete IS BCIs based on an EEG recording system. Hence this will be the focus of the following review: the next subsections are meant to give an introduction about the main methods employed in a pipeline for EEG processing within a discrete BCI system. An extensive review can be found in [15].

The challenge of a discrete BCI can be framed as a classification problem where a prediction about the class of the performed trial should be given starting from the brain signal, in this study, EEG. The general structure of a BCI signal processing pipeline is depicted in Figure 2.

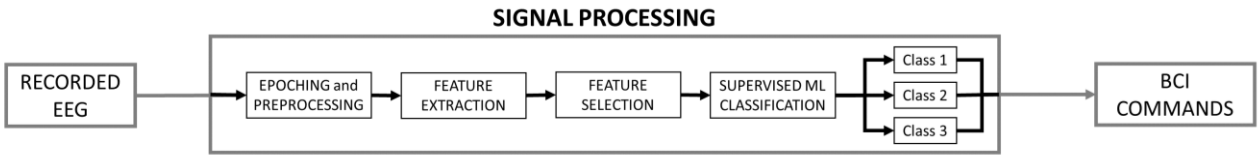

Figure 2: EEG-based BCI signal processing pipeline. The EEG is cutted into epochs, pre-processed and then features are extracted from it. The most relevant features are selected and used by machine learning (ML) algorithms for the final classification providing the output of this block. Then it can be translated into commands and be used to operate a device in a BCI system as shown in Figure 1.

2.1.1 Epoching and pre-processing

For EEG signal, the pre-processing phase can be crucial to increase the quality of the signal and reduce the effect of artifacts due to noise. It can come from various extrinsic sources: the environment (*e.g.* power line noise) or possible experimental artifacts (which can be partly reduced by adopting standard operating procedure, but are often not completely removable, such as the signal slow drift due to changes in the impedance of electrode contact). Or, the noise source can be intrinsic due to biological artifacts such as ocular movements artifacts, eye blinking artifacts, muscular artifacts or cardiac artifacts. To cope with these artifacts various algorithms have been proposed and used. They can be grouped as:

- Regression methods: the artifact effect on the EEG signal is estimated using the amplitude relation of a reference signal such as the ECG or the EOG.
- Blind source separation (BSS) method: the use of these methods is based on the hypothesis that the recorded signals are the result of the summation of distinct original signals coming from different sources. The term "blind" refers to the fact that, differently from regression methods, with these methods the reference signal is not needed.

  Indipendent Component Analysis (ICA) extracts component maximizing their independence. It is based on statistical algorithms, uses linear transformations to work and is built on the hypothesis that the original signals are statistically independent. The maximal number extracted independent component is limited by the number of EEG available channels. It is often used to identify and remove eye blinking and heartbeats artifacts.

  Principal Components Analysis (PCA) and Canonical Correlation Analysis (CCA) are other BSS methods which are more used to reduce the data dimensionality and to identify specific (noisy) patterns (typically at fixed frequencies) in the signal.
- Wavelet Transform (WT) is used for spectral decomposition of the signal. Moving to the time-frequency domain, it can localize features related to certain frequencies in the time. This method is proper to cope with artifact only if their frequency content is very specific and different from that of the signal.
- Filtering methods:
  - Frequency filtering: as WT, this method can be useful only if the noise has a specific frequency content not overlapped with the EEG. Usually notch filters are used to remove the power line noise; high pass filters to remove the effect of slow signal drifts; and low pass to limit the frequency content to a range where useful EEG information is present.
  - Adaptive filtering: using a reference signal considered the source of the noise, an estimation of its effects on EEG channels is estimated and subtracted to clean the signal from artifacts. Similarly, to regression methods, the reference signal is needed.
  - Wiener filtering: based on statistical technique, this method uses a linear time invariant filter estimated without the need of a reference signal. Nevertheless, the computation methods can make it difficult to use.

Creating epochs out of the EEG signal is then a needed step for cued BCIs where the prediction is based on the EEG signals measured in a fixed time interval. The length of this time window depends clearly on the paradigm and stimulation strategies adopted. A common strategy is to use "markers" during the EEG recording: the time when a specific stimulus is given to the subjects is saved so it is

possible to associate and synchronize it to the EEG signal. Hence, the EEG signal can be cut around this time instant to obtain the "epoch" related to the specific stimulus, which typically represents the onset cue for the task. This epoching step can be performed after the preprocessing phase, or before on the raw EEG. In this second case, the pre-processing steps should be performed directly on the cut signal, but it should be considered that some of these methods can have bad performances on short time windows or create border artifacts. Hence two solutions are possible: for some of these methods (such as ICA), the parameters can be fitted on the whole EEG recording and then applied on the single epochs. For other methods, such as frequency filters, epochs can be firstly cut a bit broader taking into account the possible border effects, and then, after the pre-processing steps, a second crop can be performed to get rid of the excess signal corrupted by border artifacts.

2.1.2 Feature extraction

Features can be extracted from the cut epochs in very different ways, according to the characteristic behaviour of the brain signal in the specific conditions. The aim is to maximize discriminability between epochs belonging to different classes. Before starting with feature engineering, it may be useful to inspect data to find where and how EEG is different among classes.

Although the possible combinations and ways of combining or presenting features are limitless, it can be useful to define a taxonomy related to the domain of different feature extraction methods:

- Time domain: some features can just be extracted straight forward from the EEG epochs measuring some statistical metrics (such as amplitudes at certain time instants with respect to the cue, root mean square along the epoch, variance...). PCA and ICA can also be applied to extract information from data, reducing the dimensionality. Autoregressive (AR) models can be used to analyse non-stationary signal as EEG: for example, it is possible to tune an AR on the EEG and use the coefficient as features for the classification problem. Also other extensions such as adaptive AR models with Kalman filter can be used.
- Frequency domain: spectral features can be often useful to discriminate different mental states from EEG. Functional bands in EEG show a certain robustness and consistency among people. Fast Fourier Transform and Welch's method can be an option to obtain relevant information from the signal in form of spectra. Power associated to specific functional bands can also be compared within each other. For example, in [16], the mental state of subjects is predicted by using also the feature called engagement defined as:

$$E = \frac{\beta}{\mu + \vartheta} \qquad \text{Equation 1}$$

  (where β, μ and ϑ stand for the power in the relative frequency bands). Nevertheless, since EEG is highly non-stationary and changes of the frequency content can be a key information to discriminate different classes, time-frequency features can be more effective.
- Time-frequency (TF) domain: the modification in spectral content is fundamental and techniques such as Short-Time Fast Fourier Transform (STFFT) and Wavelet Transforms (WT) – continuous or discrete – can be used to extract useful features.
- Spatial domain: using the spatial distribution of the signal over the scalp and filtering in the spatial domain (*i.e.*, combining the EEG channels) in an optimal way can lead to extract very well discriminable features. A commonly used technique is Common Spatial Patterns (CSP), first proposed for EEG in [17]. This algorithm creates spatial patterns which maximize the variance of the signal for one class, while minimize it for the other one and vice versa. The spatial patterns

can then be translated into spatial filters to apply on EEG: from the filtered signals (combination of the starting EEG channels), it is possible to extract significative features such as the variance (possibly in logarithmic scale) [18]. The expansion to multiclass problem was addressed and implemented in [19].

This kind of algorithm can be associated to frequency filtering to obtain patterns which also consider the spectral component and differentiability in each channel: Common Spatio-Spectral Patterns (CSSP) [20]. Also, another strategy proposed by [21] to optimally detect Event-Related Desynchronizations/Synchronizations is Filter Bank CSP (FBCSP): a filter bank of band-pass filters is firstly applied to EEG signals and then in the second stage CSP algorithm is used to filter in the spatial domain. Each pair of band-pass and spatial filter creates a new double-filtered signal, containing specific information for the band-passed frequency range. From that signal it is then possible to extract features such as the power or other statistical measures (the root mean square, the variance...) from each epoch [22].

- Riemannian geometry domain: lately some studies have used features extracted with Riemannian approaches spearheaded by the use of covariance matrices [23]. Spatial covariance matrices can synthetically represent information coming from an epoch using the relationship between channels. Compact and meaningful representation can be obtained in this way and, since they are symmetric positive definite matrices, Riemann geometry can be applied. Also, CSP feature extraction algorithms were claimed to be representable in Riemannian geometry [24].

2.1.3 Classification

Finally, the last step needs to provide a classification output when fed with extracted features. To reduce the dimensionality and increase the model's robustness, the most relevant and discriminable features are usually selected through correlation-based methods. For classification then classical machine learning (ML) algorithms can be used: support vector machine (SVM) with non-linear kernel is often achieving better performances but sometimes, other algorithms can be preferred for the purpose of explainability (linear models) or to minimize the relevance of hyperparameter tuning (regularized linear discriminant analysis – RLDA). Deep learning (DL) approaches have also been adopted and appreciated for their ability to directly extract hidden features from the EEG signal without the need of a feature engineering process. Nevertheless, two problems are limiting the use of DL architectures in this field, especially in the most explorative ones. First, the difficulty to explain the predictive process, making them partly suitable for product-oriented application but hardly usable for understanding and explaining brain mechanisms. Second, the need of massive datasets, which is hard to achieve given the high variability in data, inter- and intra- subject and session.

2.2 State-of-the-art: EEG-based BCI for imagined speech

Over the past decade, the IS paradigm has garnered significant attention, with studies exploring both continuous and discrete IS BCIs. EEG-based research has predominantly concentrated on discrete decoding, addressing the classification of vowels, phonemes and complete words. These studies have employed various strategies in their implementation choices, covering experimental setup, paradigm, subject instructions, feature extraction and classification models. As a result, diverse combinations have emerged, leading to very different final performance outcomes. Table 1 presents the most relevant EEG-based studies on discrete IS BCI, highlighting their key paradigm implementation strategies, including experimental setup, protocol, pre-processing, classification techniques and reported performance metrics. A more comprehensive overview, encompassing additional studies, can be found in Appendix A.

Table 1: State-of-the-art review (for a complete overview see Table A1 in appendix A)

Ref.	Set-up	Protocol	Pipeline	Performance
[18]	Clinical data quality. 64 channels. Sampling rate: 1000Hz.	Subjects imagine speaking 12 given English words without moving articulators or making sound. A 13th "Rest" class serves as a control. Each block includes randomized trials for each word and the "Rest" class. Trials consist of a 2s audio instruction, a randomized 0.8-1.2s fixation, and a 2s imagination period, repeated 4 times. A 3s relaxation follows each block. 7 subjects participated in 22 blocks, totalling 88 repetitions per word.	The signal undergoes band-pass filtering (0.5-40Hz) and is segmented into 2s epochs from each trial's start. Binary classification of "Rest" vs "Imagine" epochs, keeping the dataset balanced, utilizes spatio-frequency features based on CSP. Input features for the classifier include logarithmic variances of the first and last three CSP components. Classification employs a SVM classifier.	Reported performances are the average accuracy obtained from a 10-fold cross-validation procedure. Hyperparameter tuning process of the SVM is not clearly explained. In IS detection (binary classification performed 12 times, one per each word vs "Rest" condition, chance level: 50%) the reported average accuracy along all the 12 words is 80.7%. In multi-class classification problem, considering all the 13 classes (chance level: 7.7%), they achieve an accuracy of 40%.
[25]	They use the public dataset by [26]. Ag-AgCl cup electrodes. 6 channels. Sampling rate: 1024Hz.	They use the public dataset by [26]. IS of 6 Spanish words ("up", "Down", "Left", "Right", "Forward", "Backward") and vowels. The word to be imagined is played as an audio and shown on the screen. Then the subjects imagine that word in correspondence of an audio "beep" reproduced 3 times within a window of 4s. 4s rest is then allowed. 15 subjects. 40 IS trials per word.	Only 0.5s of the 4s available are used. Different strategies are tried, in particular a Convolutional Neural Network (CNN) taking as input raw EEG data achieves best classification performances.	A nested cross-validation procedure is used to tune hyperparameters of the architecture and then report performances in a robust way. The achieved accuracy in binary classification between each word pair (50% chance level) is 66%.
[27]	Clinical quality experiment. 65 channels. Sampling rate: 1000 Hz.	Inner-speech task of six Japanese words: "Up", "Down", "Left", "Right", "Forward", "Backward". The word is shown on the screen for 4s, then it is substituted by a fixation point for a second. As it disappears the subjects imagine saying the word, repeating it for 4s. 16 subjects. 10 trials per word (40s recording per word).	Signals are band-pass (1-120Hz) and notch filtered. ICA is used to remove eye blinks and CCA to reduce EMG sources of noise. The time window to be used for creating epochs is the whole 4s repeating period. Power spectral density is computed (using Welch method) per each channel and the powers in 12 frequency bands (10Hz wide, non-overlapping) are then extracted. The 12x65 powers are fed into a SVM to classify the epochs to the specific word.	Reported performance of the model are obtained using a 20-fold cross-validation scheme. Hyperparameter tuning process of the SVM is not clearly explained. The obtained accuracy for the multi-class classification problem (chance level: 16.6%) is very high: 83% using all electrodes available, 81% and 75% using only specific subsets of electrodes corresponding to pre-frontal cortex and frontal pole. The relevance of frontal electrodes and high gamma frequency bands in the decoder, make the authors speculate that "the high accuracies in the tasks were caused not only by EEG components but also by EMG artifacts".

2.2.1 Paradigm taxonomy

Experimental protocols are composed of common consecutive phases which can have different modalities and durations. Firstly, it is essential to determine the speech modality to be used among those identified in section 1. In this case, it is IS. The main experimental protocol phases are:

1. *instruction* – the presentation of the word to imagine. In multimodality experiments, this phase is also used to communicate which speech modality should be used to perform the task;
2. *preparation* – a preparatory period when the subject waits for the task to begin;
3. *go-cue* – indication that the subject should start the task;
4. *imagination* phase;
5. *resting* time.

Preparation and *resting* time can be very useful to increase the quality of data and minimize the presence of artifacts due to movements and eye blinking during imagination periods. Nevertheless, considering the same number of IS trials, the total experiment time would increase if adopted.

For the *instruction* about the prompt to be imagined, different modalities can be used:

- audio: the instruction is given through an audio recording of the word [18], [28]–[30]. In this case the auditory cortex is elicited, potentially influencing the activation of the brain also during the following imagination period [29].
- visual:
    - the written prompt is shown on the screen [27], [31], [32]. This type of instruction could cause a certain fatigue due to the accommodation process that the eye needs to perform to read prompts of different lengths. So, it might be uncomfortable for protocols with many different instructions and high number of repetitions.
    - symbols are used, in particular with arrows or oriented triangles, to define the word to be imagined [33], [34].
    - questions with a closed answer [34].
- combined audio and visual instruction about the word can be presented at the same moment [26], [35].

Cueing for the imagination onset is probably the most critical part. Indeed, it is the only source of information available to define the performing period, the ground truth. Since no visible and measurable behaviour (*i.e.* movement) should happen during IS paradigm, it is fundamental that the subjects perfectly understand when they should start the imagination phase and try to be as consistent as they can to the cue. In some cases, the go-cue is merged with the instruction: the subject may be asked to perform the imagination just after the instruction phase finishes. The modalities used to give the go-cue are:

- audio:
    - a "beep" sound is giving the cue for imagination onset [26].
    - the subject starts the imagination phase as the acoustic instruction finishes [28].
- visual:

- appearing/disappearing of figures different than those used for instruction; those can be a fixation point [27] or cross [18], [30].
- color changing of a fixation dot: [33] uses it to indicate the end of the imagination period.
- the subject starts the imagination phase as the visual instruction (written word or symbol) disappears from the screen [31]–[33], [35].

▪ rhythm completion: the subjects attend a series of acoustic stimuli ("beep" sounds) with a fixed rhythm; after the last stimulus is given, the subjects have to start the imagination phase at the moment when the subsequent "beep" would happen [29], [34]. This method tries to remove the presence of a potential response evoked by a sensorial stimulus used as cue (auditorial or visual).

Finally, when designing the *imagination* phase, a crucial consideration is whether subjects should imagine saying the prompt only once at each cue, with the possibility to use one instruction followed by multiple single-repetition go-cues to optimize acquisition time [18], [26], [29], [30], [36], [37]. Or, alternatively, subjects can be instructed to continuously repeat the prompt for a longer time [27], [28], [31]–[35]. Employing continuous repetition is beneficial to have a coherent period during which the activity remains consistent, aiding in understanding the ongoing brain activation. However, to detect a single event it is needed to use a protocol where the subject is instructed to imagine speaking the word at a specific time, only once. In this case, one should keep in mind that the brain activity will consist of both the IS activity and the sensorial response to the go-cue. On the other hand, in a continuous repetition protocol, the cue effect is present only at the beginning and not throughout the entire imagination period. With single word imagination strategy, it's important to note that the time required for all the preparatory steps (preparation, instruction and rest) is allocated for a short imagination time: just the time needed for one single trial. To optimize recording sessions, one possibility is to have more cues following the same instruction: the subject performs IS repetitively, according to a bunch of repetitive cues. This approach reduces the overall time spent for instructions while still enabling single-word imagination [18], [26], [30].

2.2.2 Accuracy and performance reporting pitfalls

The accuracy of a classification method is critical in determining the performance and reliability of a BCI system [38]. Discrete IS BCI studies use to report their performance in terms of achieved classification accuracy. Since not many data samples are available, it is common to report a cross-validated accuracy score: to guarantee the robustness of the reported values, accuracy is computed on k-fold splits by dividing data into a train and a test set for k different times. Then, the k scores are averaged and reported. Given the different number of classes in different studies, it is important to consider the proper chance level to judge the achieved performance.

A fair comparison of different protocols and classification methods is very hard to accomplish due to all the differences among the paradigms (different chance levels, different devices used, different prompts which the subjects are asked to imagine speaking, difference in the modality and the timing of instruction and cuing). Moreover, some issues biasing the reported performances have been identified: the cross-validation (CV) methods to tune hyperparameters of the models can cause an inflation of the reported performances whether one does not pay attention enough at avoiding information leakage between training and testing data sets. In [39] it has been shown that in classification models based on a SVM classifier, the tuning process of the hyperparameters should not be performed with the same CV procedure used to report model accuracy. Instead, a nested CV is needed to guarantee generalizable and robust results. Classification accuracy of the same SVM

classifier on BCI data can increase from 63% to 76% if a non-nested CV is used to tune hyperparameters and report performance [39]. The same was also proven with neural network classifiers, where the increase was proven from 60% to 90% [38]. It can be found also in this aspect the reason for such a wide range of the reported accuracies in literature, spanning – even for the same public dataset – from values just over chance level to highly accurate models. Also in IS BCI studies, the hyperparameter tuning process is not always performed through a nested CV, resulting into not completely reliable reported accuracy. Among the analysed studies some of them use models where hyperparameter tuning is not requested because they are estimated in a Bayesian framework (*e.g.*, RLDA [30]). Only [25] reported the use of a nested CV to optimize the neural network architecture and to report the achieved performances (66% in binary classification). The biggest part of the other analysed studies is not clearly explaining the hyperparameter choice, or they declare to use non-nested CV procedure to optimize them and report the model performance. Those accuracy results may be inflated and not generalizable.

The main limitation for applying nested CV is surely the high computational cost which increases drastically with the number of splits used and hyperparameter combination investigated. A heuristic strategy to avoid data leakage is to use reasonable parameters without tuning them to achieve better results as performed in [36]. Suboptimal models would be created in this way, but the obtained perfomances are more reliable.

The second critical aspect in performance reporting was identified in [40]: it is fundamental to deal with the high temporal non-stationarity of EEG signal to ensure that the observed discriminability of data is not mistakenly attributed to the mental state when it may actually be influenced by time-related features. EEG has relevant similarity (resulting in high correlation) between signal segments close in the time to each other. When classifying EEG segments into different states, it's crucial to consider this, particularly in experimental protocols where multiple segments for the same class are obtained from adjacent time windows. Indeed, in [40] this aspect was investigated for a passive BCI aimed at distinguishing the mental states. In that case, under specific conditions, the use of a random splitting strategy to create train/test sets inflated the classification accuracy by 25%. A possible strategy to cope with this problem is to use block-wise splitting strategies, where EEG segments coming from the same block (same time period) and belonging to the same class are kept together in the same split, either in the train or in test set. In this way, the information related to the time proximity of two consecutive time intervals cannot be used to classify the EEG epoch in the correct class, but only the task-related features will be exploited by the model.

[41] applied a similar analysis to an active BCI paradigm developed by [42] where they analysed the brain response evoked by images shown to healthy subjects. They claim and show that high performances reached by [42] are mainly coming from the employed block design where all stimuli of a given class are presented together, in the same time period; instead, the model fails with a rapid-event design, where stimuli of different classes are randomly intermixed and do not share temporal proximity features. «The block design leads to classification of arbitrary brain states based on block-level temporal correlations that tend to exist in all EEG data, rather than stimulus-related activity» [41]. This same problem may be also present in an IS discrete BCI whether the protocol does not contemplate randomization of the instruction order as in [32] or if the protocol employs multiple repetitions following the same instruction as in [18], [30]. For sure randomization of the instruction order is fundamental for a fairer evaluation of the classification performances. Moreover, when performing k-fold CV for model evaluation, paradigms where more repetitions are performed would require an instruction-wise data splitting strategy keeping all the adjacent repetitions

together, in the same split. Without this attention, it can be expected to achieve better classification performance which are partly based on the time-related signal characteristics shared by closer epochs. This is also shown in [28]: IS detection accuracy (imagination period vs resting period; chance level: 50%) goes from 96% to 58% when inter-class time distance and intra-class time distance (*i.e.* the time distance between epochs belonging to different classes and to the same class) are respectively different or equal. When they are different, the model can exploit temporal features for the classification, resulting in very good performance: indeed those temporal features are strong and shared by all the epochs belonging to the same class. Instead, if the epochs of both the classes are uniformly coming from the same period of time, the only features to use for the model's decision are related to IS, making the classification more difficult and the accuracy drop.

3. Methodology

3.1 EEG collection

3.1.1 Instrumentation

All data in this thesis are recorded with Mentalab systems. For the pilot study the device Explore was used; for the other experiments the more recent version, Explore+ was employed [43]. Both versions share most of the core features; the more recent version simply ensures a more stable and higher-quality signal. Mentalab Explore+ is an 8-channel research-grade EEG recording device. It has 9 electrodes: one of them is used as ground for electrical recording of the other 8 channels. It is also equipped with motion sensor, recording 9 signals (3 orthonormal linear accelerations, 3 angular accelerations and 3 angular orientations). It is provided with Bluetooth communication which streams the signal to a receiver up to 10m.

For the ground reference, a wet sticker electrode on cleaned hair-free skin area (the forehead) is used to guarantee a clean and stable signal. For the others, it was chosen to use conductive polymeric electrodes and, to further increase the quality of the signal, a little amount of conductive gel is applied. The set-up is not dry, but due to the use of only 8 channels and a minimal amount of gel on electrodes, it is quicker to prepare and less bothersome for the subjects. Thanks to its Bluetooth connectivity, Mentalab Explore+ is a completely portable EEG amplifier. The light weight and the absence of cables connected to the laptop improves the subject's comfort.

Sample frequency is set to 500Hz.

3.1.2 Electrode positioning

Since the device only allows for 8 EEG channels, the implementation choice about the channel positioning is critical and needs to be optimized. A functional overview of the brain areas more involved in speech and IS activity is surely needed to identify the optimal locations for electrodes. Nevertheless, the translation from brain areas to scalp EEG channel position is not so straight forward, hence an analysis of the most important EEG channels is needed to identify the best channels to detect IS. An analysis of the classification performances in literature was carried on to find the pool of electrode positions which may better fit the problem. The final selection of positions is influenced by this analysis and the signal optimization procedures. In some cases, it may be challenging to obtain a clean EEG signal due to the cap/head shape, and the available EEG cap is one size only. It's not guaranteed that all positions can be easily optimized, as the cap's adherence to the head may vary in different scalp regions. As a result, certain positions that provided good signal

quality with the used cap were chosen, even if the initial analysis had suggested the use of slightly different positions.

Clinical neurology identifies the main speech-related brain areas on the left hemisphere in most of the people: the dominance of the left hemisphere is indeed present almost in the totality of right-handed population (95-99%) and in above 70% of left-handed population [44]. The first investigations about speech related areas come, as often happened in neurology, from the analysis of pathological conditions: lesions in specific left brain areas causing aphasia in patients, suggested to Paul Broca and Gustave Dax left lateralization of language functioning [45]. Lateralization was also extended to language comprehension by Wernicke who found that a lesion in the left temporal lobe could be associated to language comprehension deficit [46], [47].

Considering the Brodmann brain areas division, for the highest conceptualization functional layers of speech, the cortex areas mainly involved in speech production are found to be [48]:

- «Area 4 (primary motor cortex): control of the larynx and the tongue. Perception of sounds in which the tongue is predominantly involved (in collaboration with area 6).
- Area 44/46: articulation of sounds and their assembly into syllables.
- Area 45/44 (also known as Broca's area): elaboration of syntactic relations between words.
- Area 47/45: processing of semantic relations between words.
- Areas 9, 8, and 46: comprehension and production of sentences.
- Areas 1, 2, and 3 (lower parts): perception and production of speech sounds.
- Supramarginal and Angular gyri: analysis of meaning (animate or inanimate gender nouns), syllabic analysis of words, word formation.
- Areas 41/42 (primary auditory) and area 22 (also known as Wernicke's area, secondary auditory, connected to Broca's area through the Arcuate fasciculus): elaboration and categorization of sounds. Association of a sequence of sounds to a concept.
- Area 21: retrieval of information related to syntactic properties of words that will be processed by area 45/44 of the frontal lobe.
- Superior planum temporale (Sylvian parietal-temporal area): it transforms auditory input into motor input; conjunction of information from and to the auditory area; laryngeal control». [48]

However, IS cannot be simply described as speech without the motor components [49]. There are two main hypotheses about the similarity of neural correlates of overt and imagined speech: either they are considered to share the main neural processes and activations [50] or they are claimed to be completely separated activities [51]. In the middle a third hypothesis was proposed by [52]: overt and imagined speech neural activities are similar at lexical level, while differ at phonological levels. In general, it is hard to have an agreed model for IS, also considering the difficulty to define it in a coherent and consistent way, and for the lack of behavioural and measurable independent metrics. Studies about the similarities and differences between overt and imagined speech investigate these hypotheses, trying to find the main streams of information happening in the brain for IS. The recording modality used for these kinds of studies are mainly MRI [49] and ECoG [29]. Anyways, the passage from MRI brain voxels to the scalp position is not that immediate. Instead, on the other side, the passage from ECoG to EEG may be more straightforward; however, different sources of

noise and signal quality may be misleading. Hence, to identify the most informative EEG channel locations, one should be mainly driven by the EEG analysis directly.

ECoG analysis by [29] has shown that during inner speech the activation in the right hemisphere is more focused in specific areas, while in the left one it is more diffused. It looks like the right hemisphere only has some specific areas related to IS, *i.e.* the sensory-motor areas and the auditory cortex, while in the left hemisphere the activated areas are more broad. They analysed two different modalities of imagination: subjects were asked to imagine themselves speaking a specific word, or to imagine listening to it. The most activated regions result the fronto-parietal sensorimotor ones showing a similar behaviour as the one found in [53] via MRI. The most activated regions in the right hemisphere results those related to sensory tasks, the auditive cortex, as if a sensory predicted copy would be created as an expectation by the forward models in the brain. Hence, the role of the auditory (temporal) cortex is relevant on both hemispheres.

Some studies used full-scalp EEG to identify the electrode positions which are more relevant for decoding IS. Topographical maps of variance or correlation with the target on the scalp are often used, as well as channel ranking based on the discriminability of different classes. Pearson correlation with classification targets is used in [35] to rank the extracted features and the respective EEG channels (from a 64-channel pool, standard 10-20 placement system). The most relevant electrodes are found in central locations (FC6, C5, C3, CP1, C4), in the temporal locations (T7 on the left and FT8 on the right), around the auditory cortex (CP3, CP5). Hence, similarly to [29], they have found a strong lateralization to the left hemisphere (7 out of the 10 best electrodes are on the left hemisphere) where the best electrodes are quite spread out. On the right hemisphere instead the most informative locations are localized in the auditory cortex. Notice that in this work the instruction about the prompt to be later imagined is given both with a text stimulus and with an auditory utterance.

In [18], classification was performed by using only selected subsets of electrodes and the accuracies compared to the full-scalp set-up. As a result, they have shown no significative difference between perfomances using all 64 electrodes or only the cortical group of the 10 left-sided electrodes covering Broca's and Wernicke's area or of the 10 symmetric electrodes covering the auditory cortexes. Also here, the lateralization is shown to be relevant, as well as the importance of the auditory cortices on both hemispheres. To be exhaustive, here's the two used subset of electrodes performing not worse than the whole scalp:

- Broca's and Wernicke's areas: AF3, F3, F5, FC3, FC5, T7, C5, TP7, CP5, P5.
- Auditory cortex: FT7, FT8, FT9, FT10, T7, T8, TP7, TP8, TP9, TP10.

Finally in this paper, they also performed single electrode IS detection achieving accuracy higher than chance level. The results are shown in topomaps where one can identify the electrodes performing best, and supposedly providing more information: on the left hemisphere the best electrodes are FT9, FT7, F7, FC5, while on the right hemisphere FT10. Also in this study, the instruction about the word to be imagined is auditive so the activity of the auditory cortex may come from an extension of the sensorial perception, possibly representing a later-stage activity in the auditory cortex. Nevertheless, [37] shown that by imagining to pronounce a word (without hearing it), the subject internally hears it. This could partly explain the relevance of the auditory cortices in IS decoding.

For the current study with Mentalab Explore+, 8 electrodes and the ground position had to be chosen. A left-sided dominance seems the best choice to maximize the IS related information. Nevertheless, the relevance of the two-sided auditory cortices also should not be neglected. The final choice is to have three electrodes on the right hemisphere covering auditory cortex (including the channel FT10, one of the most relevant in [18]) and five on the left. Three left channels are covering the auditory cortex symmetrically to the right side and additionally the last two channels are employed for IS related areas in the frontal cortex, in this way they could get signal from Broca's and related areas. Posterior location, more related to Wernicke's area, are a bit neglected due to the small number of channels, although some activity from those centres may be recorded through the electrodes covering the posterior part of left auditory cortex.

The selected electrodes are:

- right hemisphere: auditory cortex: TP8, T8, FT10.
- left hemisphere: auditory cortex: TP7, T7, FT9; frontal cortex: FC5, F7.

Nevertheless, the channels T7, T8 (just above the ears) and F7 resulted difficult to be optimized to get good signal quality during experimental sessions. This is probably due to the shape of the cap which is not able to guarantee enough pressure over the scalp in those locations which are at the extremities of it. Hence, their positions were slightly changed respectively into: C5, C6, F5.

For the ground, the position Fpz was chosen: it is a hair-free location so the use of a sticker electrode is possible and could guarantee a reliable and stable signal which is fundamental have clean measurments from all other sensors. The final EEG montage implemented during the experiments is shown in Figure 3.

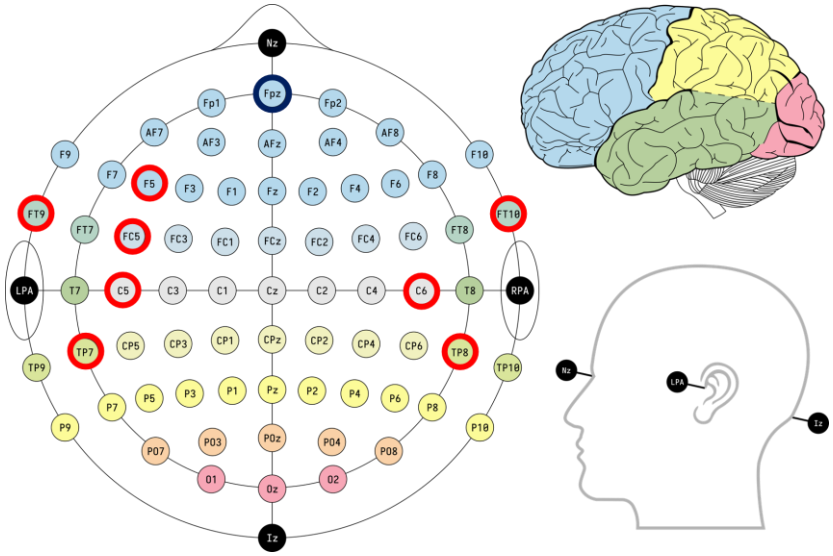

Figure 3: Final electrode position choice. The ground channel (Fpz) is encircled in blue, the other eight channel position in red.

3.2 Experimental paradigms

As it has been reviewed in the state-of-the-art analysis, there are different possible implementation choices about the experimental paradigm to use. At first, the kind of imagination should be addressed. In all the experimental sessions the subjects are instructed to perform IS by imagining

saying the words as if they were pronouncing it without moving any articulator. It is specified that the imagination process should not be about imagining hearing or reading the word in their minds, but to imagine pronouncing it.

Since the focus of this study is about detection, the vocabulary of words to be imagined is very reduced and a control resting class is also considered. The selected prompts to be imagined are the English words "LEFT" and "RIGHT". For the control task, it was chosen to give to subjects the same sensorial stimulation as it is given in the IS tasks. This is needed to avoid that the classification is based on evoked reaction to visual stimulus more than on the actual processes of speech imagination. The control task will be addressed as "NONE".

For each cue, the subjects are requested to perform only one single repetition of the IS task, they are not requested to continuously repeat the word. The single-repetition modality is selected to work in a condition more similar to the natural way of communicating.

Two different paradigms were implemented and compared along the pilot study. The previous characteristics are common to both, but they differ in the way the instruction about the word to be imagined and the go-cue are given. In particular, the aim is to identify how imagination strategy can vary according to the used paradigm and how the onset timing changes. In IS, the lack of a measurable ground-truth is a big challenge and the different strategies can optimize the adherence of the subject perfomances to the desired behaviour. Giving the cue in a way that the subjects can start their imagination phase exactly at the precise time at every trial is fundamental and these two paradigms are designed to investigate this aspect.

3.2.1 Color-changing cues

The first implemented paradigm is more like those present in literature, in particular it is inspired by [18], [30] and [33]. It is divided into blocks, each of them lasting few minutes. In each block a certain number of trials is performed. The number of blocks and trials depends on the specific experiment and will be referred when each of them will be explained. However, the trial is the building block of this paradigm and it is always the same. Each trial is associated to one task ("LEFT", "RIGHT" or "NONE") and the order used to present different classes within a block is randomized.

A white fixation cross is shown on the screen for the whole duration of a trial untill the final resting phase. The trial starts with the *instruction* about the task to be performed, it is shown on the screen for 2s. The instruction is composed of a triangle oriented to the left or to the right for the two imagination tasks ("LEFT" or "RIGHT"); the triangle is surrounding the fixation cross and after 2s disappears. For the "NONE" task, only the fixation cross is shown.

Then, three repetitions in a row for the same task (previously instructed) are cued: each repetition is composed of a *preparation period* of 1.5s (with a random variability of ±0.2s) where the white fixation cross is shown. Each *preparation period* is followed by a *performance period* of 2s where the fixation cross has turned green. As soon as the fixation cross turns from white to green, the subjects imagine saying the word previously instructed. If the subject was instructed for the "NONE" task, no particular activity should be performed but the cross changes color anyways. Then, the *preparation period* of the subsequent repetition follows, indicated by the cross turning white again. After the three repetitions, an *ending period* of 1.5s (with a random variability of ±0.2s) is indicated by the white fixation cross.

Finally, the trial is over: the fixation cross disappears and the screen turns black for 1.5s (with a random variability of ±0.2s) for the *resting period*. Then a new instruction is shown, as a new trial begins. This paradigm is shown in Figure 4.

The ending period avoids the apparition of a strong ERP after the last repetition. By using it, it is guaranteed that also the conclusion of third repetition will not be different from that of the other two.

In this paradigm, subjects are asked to minimize the blinking during the periods where the cross is green (*performance period*) and in general to try to exploit the *resting period* to blink if needed.

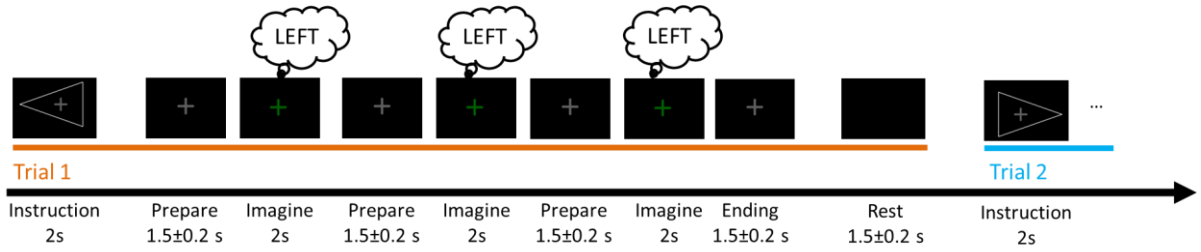

Figure 4: *Color-changing cues* paradigm. One trial is composed of three consecutive singularly cued repetitions of the same word. The subject should imagine saying the word when the fixation cross turns green (as depicted by the thought bubbles). All the phases have a specific duration.

3.2.2 Online feedback

Based on the same *color-changing cues* paradigm, also some experiments with real-time feedback are implemented. The main structure is the same but a trained IS detector is available. This detector is able to process the live streaming EEG from the recording device and give a prediction about each repetition, whether a word was imagined or not. During the *resting period*, the model computes the three predictions (one per repetition). Their correctness is later shown on the screen: "OK" feedback is given whether the model was able to correctly detect the presence (or the absence) of IS in the corresponding repetition, "X" feedback if the model prediction is wrong. The timing of each phase is shown in Figure 5. The only differences with the previous paradigm are the presence of the feedback phase and that the *resting period* can be longer to allow the model to compute the prediction. Indeed, if, due to computational burden, more time is needed, the *resting period* is further elongated.

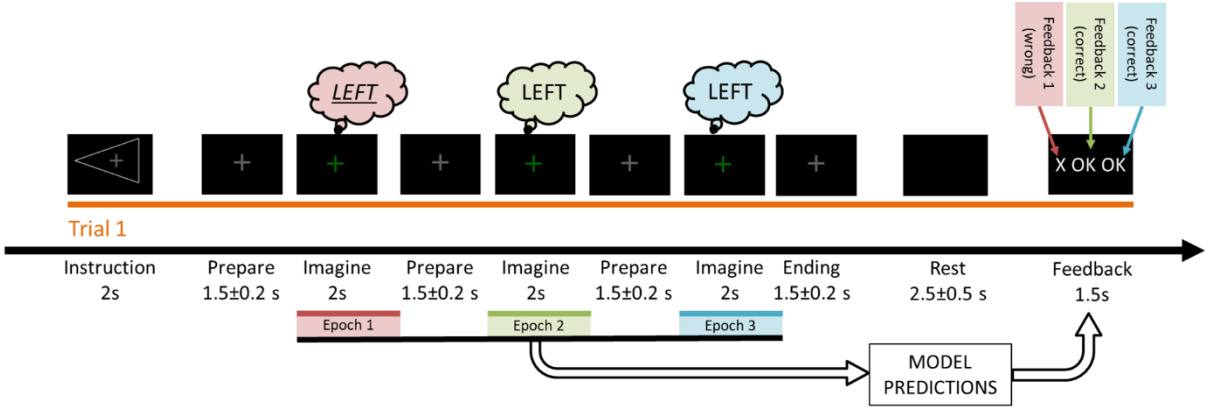

Figure 5: Online feedback paradigm. The resting period is elongated to allow the model to perform the necessary processing to predict. Then visual feedback about the ability of the model to

correctly classify the trial as "IMAGINE" or "NONE" is given. The feedback information whether the prediction was correct or not. In this case the first repetition was not consistent with the training set (represented in italic) and the model was not able to correctly predict it: hence an "X" feedback is given.

3.2.3 Sliding cues

The *sliding cues* paradigm is implemented to help the subject to be more consistent about the imagination timing onset. The instruction and the cue are given together, always in form of oriented triangle ("LEFT" and "RIGHT") or absence of triangle ("NONE") surrounding a cross.

Again, this paradigm is divided into blocks composed of a number of trials. Each trial is just a single repetition of the word to be imagined. The cues (represented by smaller white crosses) slide with a slow constant velocity along the screen to minimize the possible ERP due to abrupt changes in the field of view or unexpected accelerations. A thicker white fixation cross is shown in the middle of the screen. During the whole block, the subjects are asked to keep the gaze to the center of it. From the right side a continuous stream of smaller white crosses slides. Some of them are surrounded by a grey oriented triangle associated to the relative imagination task, other ones are not and are instead associated to the "NONE" task. The subjects are supposed to start performing the instructed task in the exact moment when the sliding cross overlaps with the fixation cross. In case of "NONE", the subjects should just go on fixating the central cross. In Figure 6a and Figure 6b two subsequent phases of the same block are shown. In Figure 6a the subject is waiting for the first cue of the block which will be for a "RIGHT" trial. In Figure 6b the subject has just finished the task "LEFT" and is waiting for the next cue, with a "NONE" instruction. Each cue is equally spaced from the previous and the subsequent, with a 4s inter-trial interval. The distance between two cues (sliding crosses) is 240 pixels, resulting in a sliding velocity of 80pixel/s.

Of course, with this paradigm there is no specific distinction between instruction, preparation and imagination time; but on the other side, the possibility to see the cue coming allows the subjects to be more precise about the onset of the imagination performance. About blinking, the subjects are generally instructed to minimize it during the blocks, but no specific moments within the single blocks could be indicated as more appropriate.

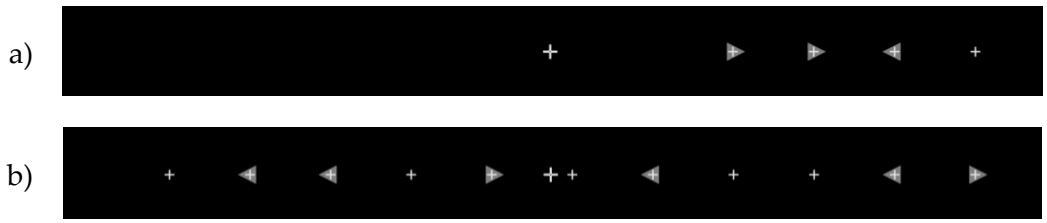

Figure 6: *Sliding cues* paradigm. In a) the block has just started and the subjects is waiting for the first cue ("RIGHT") to slide over the fixation white central cross. In b) a "RIGHT" task has just passed and the next one will be a "NONE" cue.

Note that the main focus of this study is about IS detection, hence for most of the next analyses, data related to the two classes "RIGHT" and "LEFT" will be often put together in the class called "IMAGINE".

3.3 Data inspection

3.3.1 ICA for eye blinking detection

Artifacts due to eye blinking were identified and managed differently according to the specific paradigm. Artifact correction was never applied; instead, when the issue is addressed, epochs rejection strategy is preferred for its simplicity. Identification of these artifacts is anyways crucial.

Although eye blinks can be easily identified by visual inspection of the EEG signal, an automatic method is needed. One way would be simply setting a threshold on the min-max range spanned by EEG voltage signal within an epoch. This could be enough to identity the presence of an eye-blinking artifact. The optimal threshold could be optimized in each session and for example in figure Figure 7a a good value to identify an eye-blinking artifact could be 170μV. Nevertheless, this method does not allow to be sure that the origin of such a big span in voltage measurement comes from an eye blink or another artifact. Instead, ICA is applied: by visually analysing independent components over a short fraction of the measurement it is possible to identify which indipendent component is related eye blinking (*e.g.* in Figure 7b it is the third component ICA002). It is then possible to define a threshold to apply only on that component to identify the presence of an artifact due to eye blinking. As it can be seen from the image, that one is the only component receiving information from the signal source related to eye blinking. Hence, by applying a threshold on the min-max range spanned by that independent component during an epoch, it is easy to define whether the analysed epoch is corrupted by eye blink or not. This is the method used to identify corrupted epochs, and the threshold tuning is manually performed on each session.

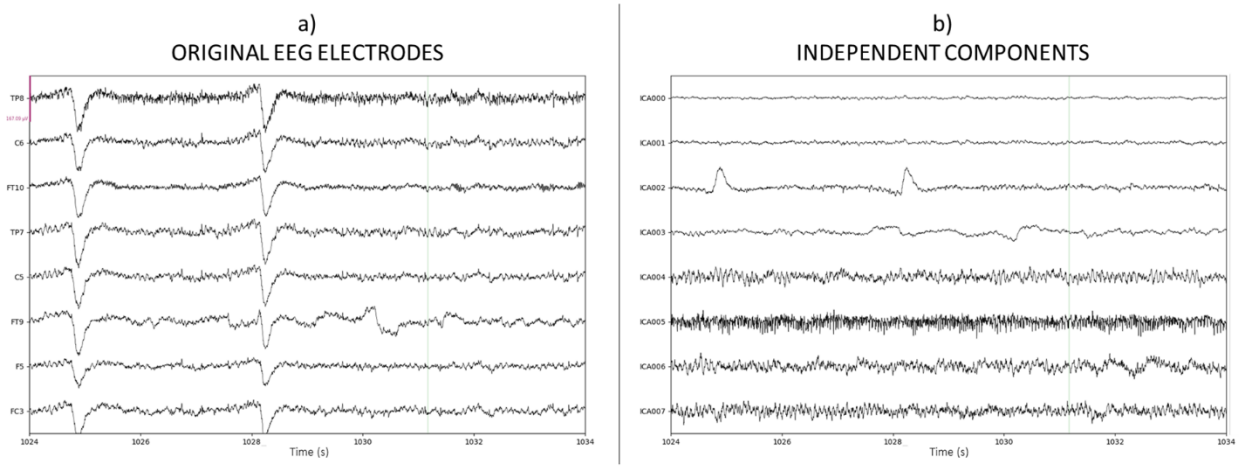

Figure 7: ICA decomposition for eye blink detection. In a) the 8 EEG channels are represented (μV). The same time window is also shown in b), where the 8 independent components are shown. In this time window there are two eye blinks (around 1025s and 1028s). In b) they are visible on all the channels, while in a) it is clear that all the contribute is focused in component ICA002.

3.3.2 ERP visualization

Event related potentials (ERPs) are stereotyped electrophysiological response of the brain to a stimulus. They consist of a phase- and time-locked sequence of positive and negative signal deflections happening at specific times with respect to the instant of the stimulus. This kind of response can be evoked by specific visual stimuli and exploited to build reactive BCIs as originally

proposed in [54]. In IS paradigms, ERPs have not been exploited to discriminate different classes: indeed, they are usually exploited to analyse sensorial brain responses in exogenous BCIs.

In this study, ERPs are visualized and analysed with respect to the go-cue. It is investigated how their timing changes by changing the way to present the cue. Discriminability between ERP of different classes is also investigated.

Since ERPs are highly stereotyped signals, time- and phase- locked, a typical strategy to represent them is to perform time alignment of signal windows with the same length. Time 0 is usually set when the cue is given. Once the time windows are aligned, the average per each time point along the number of available signal windows is computed. So, a final signal, of the same duration as the initial ones is obtained. Only oscillations consistent along the blocks in terms of time distance from the cue and sign are finally visible. Indeed, all the non-consistent oscillations are reduced and considered noise. Averaging multiple aligned time windows increases the visibility of the consistent signals by increasing the SNR by $\sqrt{N}$ where N is the number of trials averaged.

In the current study, ERPs are visualized on each channel, class-wise. Raw signal is band-pass filtered (0.5-5Hz), and epochs are cut around the time when the go-cue is presented with a baseline correction. To highlight some signal features that will be later discussed, starting time of the time window (always 2s long) changes according to the analysed paradigm: for *sliding cues* paradigm the considered time window is from 1s before the cue, up to 1s after the cue (-1s – 1s), while for *color-changing cues* paradigm, the analysed time window is composed of the 2s after the go-cue. The baseline is composed of the first 0.5s considered. To perform this analysis, epochs containing eye blinking artifacts were dropped.

To further improve SNR and find the most relevant characteristic of the ERPs, instead of computing the time-locked average, it is possible to apply Singular Values Decomposition (SVD) on the EEG filtered epochs [55]. SVD is applied on the matrix composed by the concatenated epochs, resulting into the dimension $N_{time} \times N_{trials}$, where N_{time} is the number of time samples available in each epoch, and N_{trials} the number of available epochs. This procedure is used to find relevant signal patterns; then, based on the most explicative components, it is possible to reconstruct the signal and visualize the identified characteristics. In this study, only the first component was used to reconstruct the signal, because in all analysed cases the first resulting eigenvalue is at least the double of the second, meaning that its relevance is the double of the second one. SVD is applied on the same time windows described before.

3.3.3 TF analysis

Numerous biological systems display rhythmic patterns and temporal dynamics. These temporal structures in signals can be revealed and quantified using techniques like fast Fourier transform (FFT) and other spectral analysis methods. However, it's important to note that FFT operates under the assumption of signal stationarity, implying that spectral and other features of the signal are expected to remain constant over time. While FFT provides a perfect representation of a signal, irrespective of its temporal dynamics, its power spectrum is readily interpretable only for signals that exhibit stationarity. When non stationarities are present – as in EEG – the power spectrum becomes blurred, and the relevant information is often encoded in the phase spectrum, making it challenging for visual interpretation. To address these non-stationarities in signals, time-frequency (TF) analyses come into play. In these analyses, the power spectrum is computed over short time windows, recognizing that the signal is approximately stationary within these smaller, sliding time intervals. In EEG, this assumption typically entails using time windows spanning a few hundred

milliseconds. This approach allows to capture and understand the dynamic changes in signals coming from the brain [56].

For TF analysis and for the next section about TF classification, only two classes are considered: "IMAGINE" and "NONE". If the eye blinking artifact rejection procedure creates an unbalanced dataset (different number of rejected epochs for one class), the most represented class is randomly down sampled. For TF analysis different algorithms have been tested (Hilbert transform, Stockwell transform, STFFT), but the method which shown more interesting results is the representation based on Morlet Wavelet Transform. In particular, the algorithm used follows the formulation by [57], [58].

A complex Morlet wavelet is defined as the product of a complex sine wave with a Gaussian window whose variance σ influences the resolution of the analysis. The wavelet is convolved in time with the EEG signal, resulting into a complex signal whose power and phase are the features of interest of each time point. The spectral precision is increased with a wider Gaussian, while a narrower Guassian increases the time resolution at cost of a decreased spectral resolution. A typical approach is to use a variable length of the Gaussian with the frequency. To create a multiscale TF representation of the instantaneous power, this convolution process is performed by using Morlet wavelets representing different frequency on the same signal. The final representation is a 2D map of the power values (one per each time point for each frequency analysed), with the frequencies on the y-axis and the time on the x-axis. In this work, the Gaussian window is adapted to the frequency: the variance σ is the time needed for 10 complete cycles of the analysed frequency. The frequencies are 28 bins linearly spaced from 8Hz to 120Hz (4Hz distance).

For the representation, the power values (always positive), are normalized with respect to a baseline period using a z-score scaler. In this way it is possible to show positive or negative modification of the power of the EEG signal, at the different frequencies. The baseline is taken as the 0.5s anticipating the cue for the imagination onset. Once computed the TF representations for all the single epochs, those belonging to the same classes are averaged to show the mean TF representation of the signal in a specific condition (the different classes).

3.3.4 TF classification maps

In the IS paradigm, the absence of a ground truth signal requires a deep analysis of the data to find the most relevant signal windows usable for classification. Considering the relevance of spectral features, a TF analysis in terms of classification perfomances can be useful [59]. Binary classification of epochs into the two classes "IMAGINE" and "REST" is run on subsequent narrow time windows, using only specific frequency bands to identify the most discriminable spectral features.

The investigated frequency bands have 4Hz width, are non-overlapping and are spanning from 4Hz to 80Hz. In total, 20 frequency bands are analysed. Time window width is adapted to the considered frequency: their duration is 10 times the period of the central frequency of the relative band. *E.g.*, for the first band (4-8Hz), the used time window is lasting $T = 10 \cdot \frac{1}{6Hz} = 1.67s$. This results in longer time windows for lower frequencies and shorter duration for high frequencies. The time windows are shifted to make the shortest time windows to overlap by 50%: the last frequency considered is 80Hz, then the shortest time window is $T_{min} = 10 \cdot \frac{1}{80Hz} = 0.125s$. The time shift is then 50% of this time window: 0.0625s. Note that for the last spectral band they will be overlapped only by 50%, while for the lower frequencies, the overlap will be more.

After having pre-processed EEG signal (band pass filter 0.5-90Hz and notch filter at 50Hz), one classification pipeline (filtering, epoching, classificator training and prediction evaluation) is created

for each couple time window-frequency band. The used classifier architecture is composed of a CSP module which extracts the logarithmic variance of the first 3 components of each epoch; the three variance values are given as input to an SVC (Gaussian kernel, C=1).

The following procedure is performed on each time-frequency pair to obtain $N_{frequency} \times N_{time\_windows}$ classification performance scores (considering the number of frequency bands $N_{frequency} = 20$ and $N_{time\_windows}$ the number of window analysed depending on the length of the total time range). Each score refers to a specific frequency band (F) and time window (T):

1. the pre-processed EEG signal (8 channels) is band pass filtered in the considered band F;
2. epochs with the specific width are cut around the central time T;
3. 5-fold CV (implemented with trial-wise strategy for *color-changing* paradigm – for extensive explanation see section 3.4.3) is used to divide the epochs into balanced train and test set for 5 times. For each of the 5 folds, the classifier is trained and then used to predict the test set labels (binary classification between "IMAGINE" and "REST"). 5 accuracy scores on the test set are obtained;
4. the 5 accuracy scores are averaged to obtain one performance score for each T-F pair.

Once $N_{frequency} \times N_{time\_windows}$ average scores are obtained, they are visualized on a 2D plot: central time T of the epochs on x-axis and frequency band F on y-axis. The accuracy scores are color coded. Minimum displayed score is chance level: 50%.

3.4 Classification model and evaluation

3.4.1 Architecture

Many alternative model architectures were tested and their performances were compared on the available data in pilot study: architectures based solely on spectral content; feature extraction through wavelet decomposition; Riemannian features; time-frequency features obtained with STFFT. The final pipeline is based on an FBCSP algorithm, resembling those proposed in [21], [22]. This algorithm is complemented by a feature extractor that relies on ratios between functional frequency-band powers, drawing inspiration from [16]. The extracted features are then subjected to supervised dimensionality reduction, where those most correlated with the target in the training set are selected. These selected features are subsequently fed into an SVC. The whole model architecture is depicted in Figure 8.

3.4.1.1 Epochs

The classifier is built to merge information coming from all the 8 EEG channels together. It works on single epochs of 2s cut around the onset of the imagination phase directly from the raw un-processed EEG signal (sampling rate: 500Hz). The exact time window around the marker time is a hyperparameter to be tuned. A 1s elongation before and after the epoch is also used to avoid border effects, so in total the preprocessing steps are performed on 4s epochs (input: 2000 time points, 8 channels)

3.4.1.2 FBCSP

The epochs are fed into a filter bank of 33 band-pass filters. The bands are 50% overlapped as in [22], their range increases as the frequency increases and they span on the whole spectrum from 0.5Hz to 90Hz avoiding the frequency in the interval 48-52Hz to cancel the influence of the power line noise at 50Hz. The 33 bands are:

- up to 44Hz they are 4Hz wide (except for the first one to cut out slow drifts): 0.5-4Hz, 2-6Hz, 4-8Hz, 6-10Hz, 8-12Hz, 10-14Hz, 12-16Hz, 14-18Hz, 16-20Hz, 18-22Hz, 20-24Hz, 22-26Hz, 24-28Hz, 26-30Hz, 28-32Hz, 30-34Hz, 32-36Hz, 34-38Hz, 36-40Hz, 38-42Hz, 40-44Hz;
- one broad band 4-44Hz;
- from 40Hz to 80Hz they are 8Hz wide, but all frequency content from 48Hz to 52Hz is deleted to avoid power line noise: 40-48Hz, 44-48Hz, 52-56Hz, 52-60Hz, 56-64Hz, 60-68Hz, 64-72Hz, 68-76Hz, 72-80Hz;
- two last broader bands to cover up to 90Hz: 80-90Hz and 40-90Hz.

After this filter bank, epochs are cropped by 1s on both sides, resulting into 2s filtered epochs of 8 channels each (33 frequency bands, 1000 time points, 8 channels). CSP algorithm is then applied on each of the 33 bands: 33 different sets of patterns are obtained. From each, the first two and the last two components (the two maximizing the variance for first class and for the second class) are kept resulting into a dimensionality reduction of 50%: from 2s signals, 8 EEG channels in 33 frequency bands to 2s signals (1000 time points) on 4 components in 33 frequency bands. From each component in each frequency band, 5 statistical features are computed: signal power, variance, peak-to-peak amplitude, zero crossing rate and sum absolute value [22]. The number of extracted features is 660 (coming from 5 features per signal, 4 signals per frequency band, 33 frequency bands).

3.4.1.3 Functional-band powers ratios

Besides, the 4s epochs are also filtered by 4 band-pass filters. The four bands represent the 4 functional brain waves frequencies:

- ϑ: 4-7Hz.
- μ: 8-13H.
- β: 14-35Hz.
- γ: 30-80Hz.

Epochs are cropped by 1s on both sides to remove border effects, and the power of the resulting signals is then estimated (4 power values per each of the 8 channels). Two features per channel are then computed [16]:

- engagement E (defined in Equation 1).
- γ/β ratio (where γ and β stand for the power in those frequency bands).

3.4.1.4 Feature selection

The 660 features coming from the *FBCSP* block and the 16 features representing the *functional-band powers ratios* are concatenated and their amplitude normalized on the training set with a standard scaler. A univariate feature selection step is implemented: the features are ranked according to their ANOVA F-value with respect to the target vector and the first K are selected. The number of selected features K is a hyperparameter to be tuned.

3.4.1.5 Classifier

Finally, the K features are fed into a SVC with radial basis function kernel (which in literature has shown best results).

The feature extraction steps were implemented as an Sklearn transformer class (sklearn.preprocessing.FunctionTransformer), taking as input the epoch(s) and transforming it(them) into a feature vector of length 676. In this way, it was possible to implemente the whole model as an Sklearn pipeline class (sklearn.pipeline.Pipeline) [60].

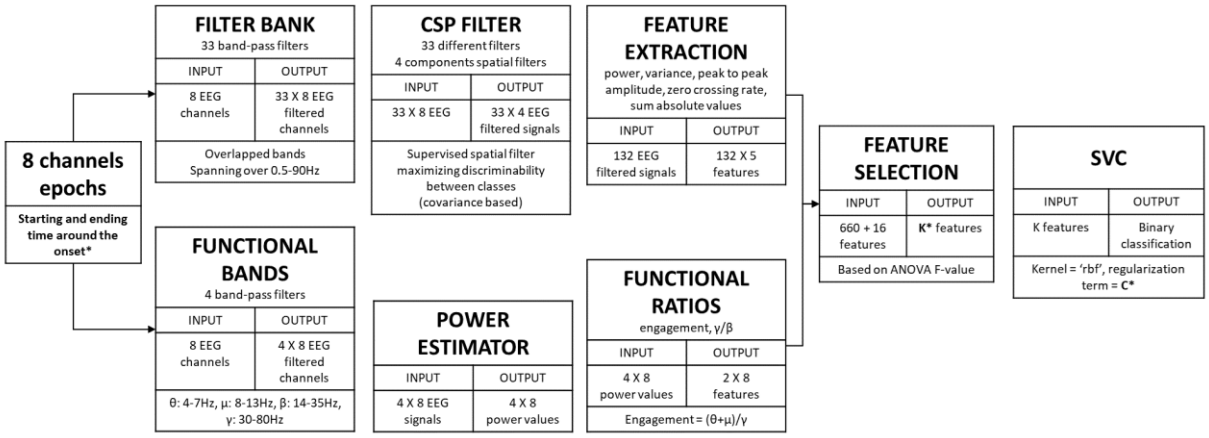

Figure 8: Model architecture. All the building blocks are depicted together with their input and outputs. The hyperparameters to be tuned are indicated with an asterics.

3.4.2 Hyperparameter tuning

The hyperparameter tuning is a critical phase which can easily introduce biases due to information leakage and inflate performance. It is very important to identify them and define a correct optimization strategy. The hyperparameters to be tuned are:

- the time window used to create the epochs. Starting time, duration and baseline correction could be influent factors. To reduce the degrees of freedom, the duration is fixed to 2s (as proposed in [18]) and the 0.25s interval just preceding the epoch start is used for baseline correction. In this way only the starting time with respect to the onset cue remains as an optimizable variable.

  The optimal value is surely depending on the specific paradigm and on how and when each subject perceives the cue and performs the task. For the pilot study (chapter 4), an extensive analysis is performed to find the best time window for each paradigm. The found value will be used when evaluating the performances of other subjects in the whole *corpus* experiment based on the same paradigm (chapter 5). For the sessions with online feedback (chapter 6), the starting time of epochs will be tuned on data coming from the previous offline session of the same subject, under the hypothesis that imagination strategy and cue perception remain the same for the same subject.

- the number of selected parameters (K).

- the regularization term C and the kernel γ parameter of the SVM. They have a complementary role. Beyond being strongly related within each other, the effect of the variability of these parameters has a strong dependence on the number of selected parameters K [61]. In order to reduce the burden of finding the optimal combination, only the parameter C is left to be tuned, while, as suggested by Sklearn toolbox guidelines [62], γ is fixed to:

$$\gamma = \frac{1}{K \cdot variance(X)} \qquad \text{Equation 2}$$

(where X are the training samples features).

The optimization procedures for the time windows are described before. Instead, the optimal combination of the two residually variable parameters K and C will be investigated through a grid-search in two phases: data from the pilot case will be used to tune hyperparameters of the models used to report classification performances on the whole *corpus* experiments. Then, a subject-specific grid-search will be performed to optimize models to be used for online sessions. Note that to evaluate performances in pilot study, these hyperparameters are not optimized, but reasonable values are selected (C=1, K=40). In this way, data leakage due to wrong optimization procedures is avoided. A nested CV would have been a better solution to properly find optimal hyperparameters. But its implementation was not possible, due to the high computational cost.

When optimal combinations are investigated, a grid-search with a 5-fold CV (trial-wise splitting procedure) is used to test the 21 combinations obtained making the two hyperparameters vary in the following ranges:

- K = {20, 40, 60, 80, 110, 140, 200}.
- C = {0.5, 1, 5}.

3.4.3 Cross-validation strategies and metrics

To ensure the robustness of the performance estimates used to evaluate the system, a 10-fold CV procedure is employed. In this procedure, each epoch is included nine times in the training set and once in the testing set. The final accuracy of the model is determined by averaging the classification accuracies of the model predictions over the ten testing sets. However, it's important to note that the splitting strategy used to create the ten folds requires special attention, as it can impact the results. For the evaluation of offline sessions, two splitting strategies are implemented and compared:

- *random split*: all the repetitions are assigned in a pseudo-random way between the train and the test set. This strategy can be applied to both implemented paradigms: the *color-changing cues* and the *sliding cues*. In the case of the *color-changing cues* paradigm, it may allow for example the presence of two of the three consecutive repetitions in the train set, while leaving the third repetition in the test set.
- *trial-wise split*: in the *color-changing cues* paradigm, the three consecutive repetitions following the same instruction are kept together in the same set, either in the train or in the test set. Clearly, this strategy is not applicable to the *sliding cues* paradigm where multiple repetitions are not employed. It is used to avoid the model to base the prediction on the shared temporal features which are not related to IS. Indeed in this way the train and the test splits of the same fold do not share temporal features related to EEG non-stationarity.

As it is very common in BCI applications, the metric used to evaluate the prediction performance on each split, and then averaged, is the classification accuracy. It's essential to evaluate this accuracy in comparison with the chance level for the specific problem. While identifying the chance level might seem straightforward in balanced datasets, it's important to establish confidence boundaries to confirm that the model outperforms a random classifier by more than mere chance. A robust chance level model, validated with simulations, was introduced in [63]. According to this model, for a balanced classification problem, the confidence limits (α=1%) can be estimated as:

$$\tilde{p} \pm \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}\, z_{1-\frac{\alpha}{2}} \qquad \text{Equation 3}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the standard normal distribution (if α=1%, $z_{1-\frac{\alpha}{2}} = 2.58$), n is the total number of available data (summing all classes) and $\tilde{p}$ is the expected accuracy of an unbiased estimator (in this case $\tilde{p} = 0.5$ since the classes are balanced). For the offline sessions in this study, in the whole *corpus* experiments, it is the case to set $n = 432$ (216 repetitions per class). Hence, for having results better (confidence α=1%) than a random classifier, the accuracy to overcome is 56.1%.

4. Pilot study

The first phase of the study was dedicated to gaining a comprehensive understanding of the problem and delineating the most effective strategies for both the experimental paradigm and the classifier. This entailed a thorough exploration of the two paradigms, with a specific focus on pointing out the fundamental differences in the subject's interpretation of the cue for the imagination onset. Given the primary challenge common to all imagination paradigms – the absence of a measurable and independent ground truth signal – a significant portion of the investigation concentrated on the temporal aspects of brain responses. The aim was to analyse how the timing of IS was influenced by the chosen paradigm, seeking ways to enhance the consistency and reliability of the timing aspect.

4.1 Materials: experimental procedure

In the pilot study, a single subject (male; age: 23; right-handed; without neurological complaints) was examined across multiple sessions, each featuring slight variations in the experimental paradigm to identify their major criticalities. This iterative process led to the definition of the two definitive experimental paradigms described in section 3.2.

The participant received information about the procedures and objectives of the experiments and provided written consent before taking part. Experiments were pre-approved by the Ethical Committee of UZLeuven, Belgium.

The major analyses are conducted on two sessions separated by a 2-week interval. In the first one, the *color-changing cues* paradigm was used. It consisted of 14 blocks. In each block, 5 trials per each of the 3 instructions ("LEFT", "RIGHT", "NONE") are performed in randomized order. It resulted in 15 repetitions of each word per block, totalling 210 repetitions for each class. The entire experimental session had a duration of 55 minutes. During the third block, the subject reported an inability to complete the task, and as a result, this block was excluded from the analysis, leaving 195 repetitions per class. To create a balanced dataset for the detection problem, specifically between the "IMAGINE" and "NONE" classes, down-sampling was implemented. "LEFT" and "RIGHT" trials were halved through a random selection procedure creating the class "IMAGINE". It is important to note that down-sampling was applied to trials and not repetitions, meaning that either all three repetitions following the same instruction were retained or removed together.

For the second session, the *sliding cues* paradigm is used. It consisted of 21 blocks. In each block 10 trials per each of the 3 instructions are performed in randomized order. It resulted in a total of 210 single repetition trials per class. The total time is the same (55 minutes). Also in this session, to create a balanced dataset between the "IMAGINE" and "NONE" classes, down-sampling was applied. Just a random stratified strategy was implemented since no problems about multiple repetitions are present.

4.2 Blinking artifacts

To visualize and process data correctly, the first step needed is identifying the corrupted epochs to avoid artifacts to spoil the analysis or influence classification performances. Indeed, if the artifacts occur consistently in one class more than another, if not dealt with, their presence may be a source of information captured by the classifier to give a prediction. The algorithm employed for the detection of epochs affected by eye blink artifacts is elaborated upon in the section 3.3.1. Table 2 provides a breakdown of the percentage (along with the corresponding number) of corrupted epochs in each session, divided by class.

Table 2: Eye blink artifacts rate (and total number). Comparison between the two paradigms considering different classes.

	COLOR-CHANGING CUES	**SLIDING CUES**
NONE	3% (5/195)	8% (17/210)
LEFT	2% (4/195)	10% (21/210)
RIGHT	1% (2/195)	9% (18/210)
TOTAL	**2% (11/585)**	**9% (56/630)**

No relevant disparities among the different classes in terms of eye blink artifact occurrence can be observed. Notably, the paradigm employing *color-changing cues* exhibited a considerably lower eye blink rate compared to the paradigm with *sliding cues*.

The subject was aware of the importance of mitigating eye blink artifacts and was instructed to minimize them, particularly during task performance periods (when instructed to either "IMAGINE" or follow a "NONE" command). In the *color-changing* paradigm, the inclusion of well-defined resting periods following each trial facilitated the subject to limit eye blinks within the designated time windows for task execution. The subject was allowed to blink during these resting periods, and this allowance helped to mitigate the need for strenuous efforts to suppress eye blinks. Conversely, in the *sliding cues* paradigm, no specific resting period was designated. The subject was generally encouraged to reduce eye blinks throughout the blocks. However, given the lack of dedicated time slots for blinking within a 2-minute block, it was understandable that the subject occasionally needed to blink, leading to more frequent occurrence of eye blinks during the processed epochs.

Given these results, the decision was made to implement an epoch rejection strategy exclusively for the *sliding cues* paradigm. Indeed, to guarantee a certain consistency, applying the same strategy to the *color-changing cues* paradigm would necessitate discarding the entire set of three repetitions associated with the same instruction, even if only one of them was corrupted. Although an artifact correction strategy might have been preferable in this case, as discussed in the section 3.3.1, its computational demands could prove too heavy for a real-time application which is the final aim of this work. In the *sliding cues* paradigm, after the rejection of corrupted epochs, it became essential to restore dataset balance. This was achieved through a random downsampling of the two most represented classes, ultimately obtaining a total of 189 single repetition trials from each class.

4.3 Creating the ground truth: time windows comparison

For IS, subjects are asked to perform imaginary speaking *without moving any body part*. Hence, there is no way to know whether the subject consistently performs the task, neither about the modality of imagination, neither about the timing. The only available tool is self-assessment, which is not objective. This is why an extensive analysis of the neural response to the onset cue is performed. The difference between the two paradigms will be investigated with the means of different tools:

- ERP representation which may reflect the perception of the stimulus and how it triggers the subject to start the imagination task (described in section 3.3.2).
- TF representation, to identify which are the event-related synchronisation/desynchronisation (ERS/ ERD) which may be related to IS and when they happen with respect to the cue (section 3.3.3).
- TF classification maps and epochs optimization, to pinpoint the most discriminative time windows and frequency bands for the detection problem and see if they change between the two paradigms (section 3.3.4).

4.3.1 Results

Figure 9 and Figure 10 show the response evoked by the go-cue in the two different paradigms respectively by applying a time-locked average (Figure 9) and by showing the first component reconstruction after SVD (Figure 10). For clarity, only the most relevant channels are represented here; the complete set of electrodes is represented in appendix B (Figure B1 and Figure B2).

In time-locked average (Figure 9), for the *sliding cues* paradigm, a negative inflection of the signal can be seen. On the other hand, oscillations with specific and consistent timing are not easy to be identified for the *color-changing cues* paradigm.

More consistent characteristics can be identified from the analysis of the first SVD component reconstruction (Figure 10). In both paradigms, for most of the channels, the first component represents a double bouncing signal. The components in the two paradigms are very similar but with a difference in the timing: for the *sliding cues* paradigm, an anticipation of 1s with respect to the *color-changing cues* paradigm is evident and it is highlighted by the visualization with a 1s shift between the two paradigms. A clear and consistent difference between classes is not really evident in none of the two representation modalities, hence it would be difficult to use them for discriminability purposes.

In Figure 11, the TF representation of the "IMAGINE" class epochs is shown for the two different paradigms. The same four electrodes as above are shown for simplicity; in appendix B, in Figure B3 and Figure B4, all the electrodes are represented and a comparison between the "IMAGINE" and "NONE" classes is also performed.

The main phenomenon which can be identified with a certain consistency in both the paradigms is a desynchronization (ERD) of components in the γ frequencies band. Indeed, a negative inflection of the power in γ spectral band can be identified: in the figure, it is highlighted by a dashed black box. It is particularly interesting analysing its timing with respect to the cue event: it is happening almost at the same moment as the cue is given for the *sliding cues* paradigm, while for the *color-changing cues* paradigm it starts about 1s after. The similarity in the nature of the event and the shared frequency range may suggest that it represents the same brain activity related to IS. Indeed, it is

happening only for the "IMAGINE" class, while it is absent for the "NONE" class (for this comparison between classes, refer to Appendix B).

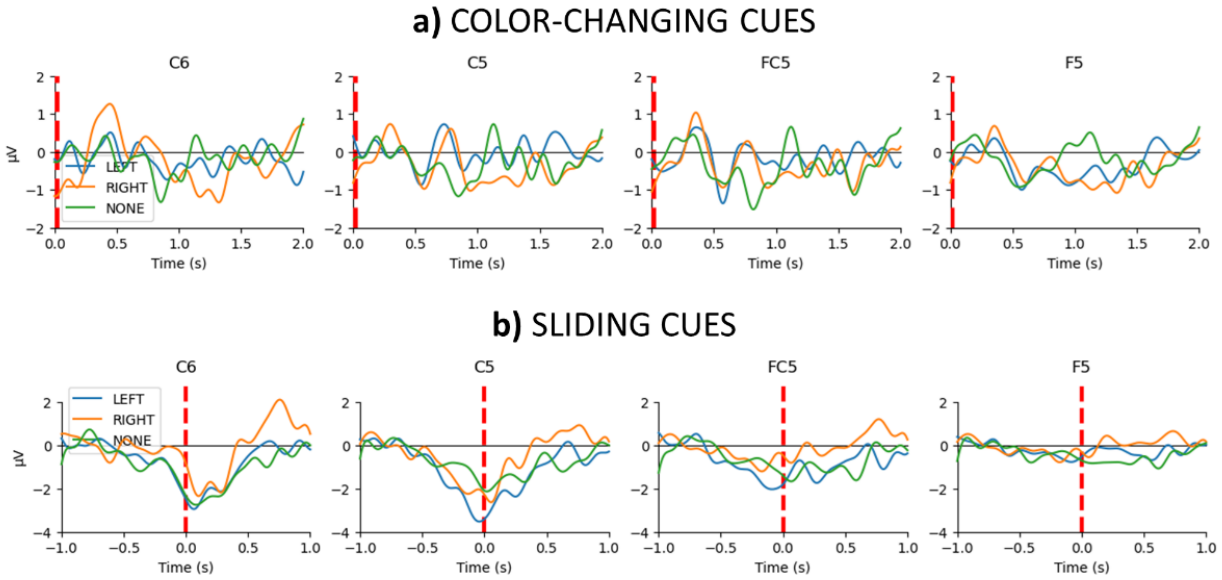

Figure 9: Time-locked average ERP. The go-cue is always given at time 0s (dashed red line). In a) they are represented for the *color-changing cues* paradigm, in b) the *sliding cues* paradigm. In both cases, a 2s window is represented, but they are visualized with a 1s shift.

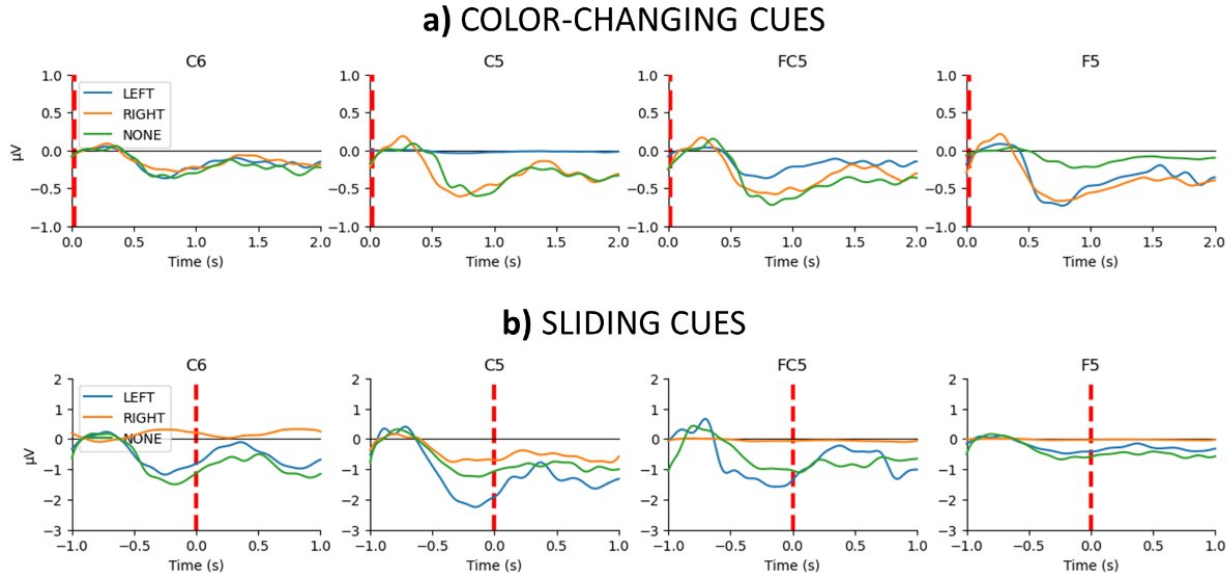

Figure 10: SVD first component ERP. The go-cue is always given at time 0s (dashed red line). In a) they are represented for the *color-changing cues* paradigm, in b) the *sliding cues* paradigm. In both cases, a 2s window is represented, but they are visualized with a 1s shift. Note that this time shift highlights the similarity in the signal shape between the two paradigms. The double bouncing signal starts as the go-cue is given for *color-changing* paradigm, while the signal is anticipated to 1s before the go-cue in the *sliding* paradigm. However, the signal morphology is the same.

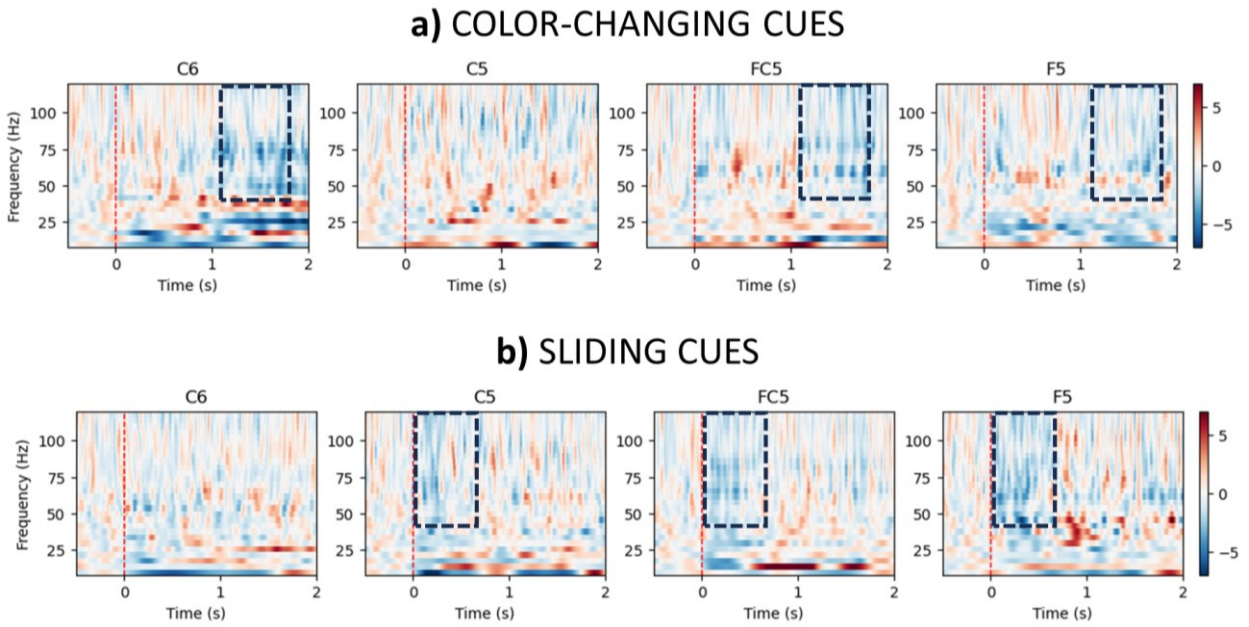

Figure 11: TF representation of the "IMAGINE" epochs for the two paradigms. The go-cue is always given at time 0s (dashed red line). In a) the *color-changing cues* paradigm, in b) the *sliding cues* paradigm. Negative values (blue) represent a power reduction in the corresponding frequency band; positive values, a power increase. Dashed boxes encircle the identified ERD (reduction of the signal power) happening in γ band at different times for the two paradigms, but with similar features. Note: in the "NONE" epochs the ERD is not visible (appendix B Figure B3 and Figure B4).

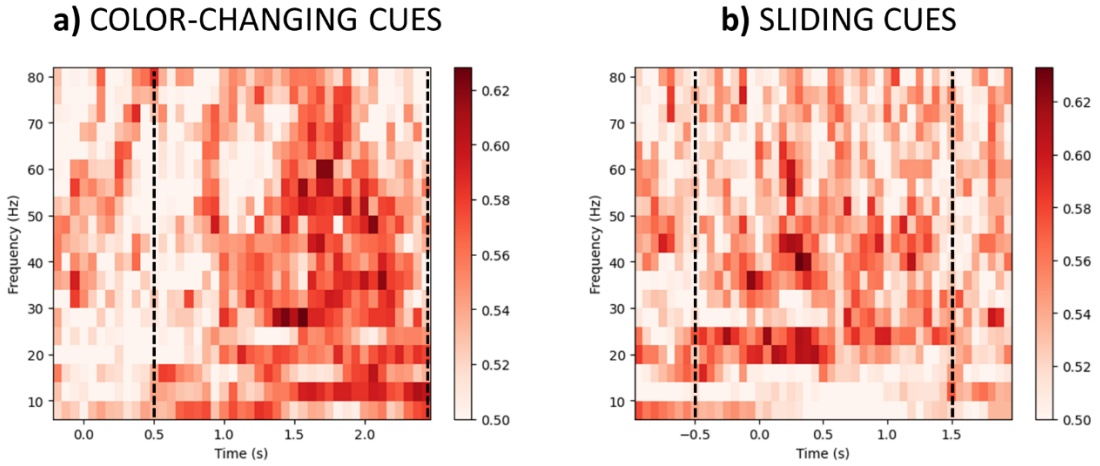

Figure 12: TF classification accuracy maps for the two paradigms. In both the plots time 0 refers to the moment when the go-cue is given. Red intensity level encodes the accuracy starting from 0.5 (50%) up to the maximal level reached in the single plot (scale on the right). The dashed lines represent the extremities of the optimal time window identified via CV (values reported in Table 3). Notably, the optimal time windows visually correspond to areas which are achieving better performances.

Finally, Figure 12 shows a map of the TF classification performances in adjacent narrow time windows, considering the contribution of specific frequency bands. The best performances in IS detection are obtained at different time instants according to the paradigm. Visually, from the two plots in Figure 12 a mismatch of about 1s can be found in the best performing areas. The most informative frequency bands appear those within the γ range (30-70Hz) and β range (14-35Hz).

Considering a more quantitative analysis, a 1s mismatch is also found in the search of the optimal time window in the two paradigms: Table 3 reports the accuracy values (5-fold CV) of the classification model described in section 3.4 when taking as input 2s long time windows starting in different instants with respect to the go-cue. The most informative time window for each paradigm is shown on Figure 12 as dashed black lines and coincide with the areas which could be visually identified as more discriminable: they are 1s apart, for *sliding cues* paradigm it starts 0.5s before the cue for task onset is given; for *color-changing cues* paradigm, it starts 0.5s after the cue is given.

Table 3: Time-window optimization. The CV accuracy scores achieved for each examined time-window (the time is considered with respect to the go-cue). The best for each paradigm is signaled by an asterisk. They are 1s apart.

Time window (s)	5-fold CV	
	Color-changing	**Sliding**
-0.75 – 1.25	59.5%	61.0%
-0.5 – 1.5	57.2%	**61.8%***
-0.25 – 1.75	60.0%	60.8%
0 – 2	59.7%	58.3%
0.25 – 2.25	64.9%	57.5%
0.5 – 2.5	**65.1%***	55.3%

4.3.2 Discussion: the time delay between the two paradigms

The negative wave found in ERP grand average visualization notable in *sliding cues* paradigm was also found in [64], and they claimed it to be related to speech and IS. It covers similar brain regions, but the timing is different than what was observed in the current study; nevertheless, this difference in timing may be due just to the difference in the employed paradigm. On the other side, they did not consider a control class where the stimulus is presented but the subject should not perform any specific task, as it was done in this work. Notably, we identified the same signal behaviour for "NONE" class too, opening the possibility that it may be driven by the sensorial evoked potential and not directly related to IS.

From the representations of first SVD components of ERPs, a difference between the two paradigms about how the cue for onset is perceived could be identified: indeed, a very similar signal pattern (composed of two consecutive oscillations) can be found in the two paradigms, but with a clear time shift by 1s (well enhanced by the shift in the plotted time windows). Still, since this pattern is found both for cues giving the onset for imagination and for cues related to "NONE" epochs, it cannot be directly linked to the act of performing IS. It is more likely related to the perception of the stimulus or in general to an intent-evoked response: the cue may cause a specific neural activation even though no particular action is required but the task is commenced.

The same 1s time shift was also identified in γ band frequency desynchronization, with a clear distinction between "IMAGINE" and "NONE" classes suggesting that this ERD could be related to IS brain processes. Changes in the activity in γ range were also found in [36], and also in that case it was claimed that they may be related to IS brain activity [65]. Finally, the 1s shift is qualitative visible also in classification performances from Figure 12, and it was quantitively measured through the results reported in Table 3 about the optimal time window to detect IS. This further confirms that the employed paradigm determines a 1s shift not only about the perception of the cue, but also on the moment when IS is performed and this information must be exploited for detection purposes.

This difference linked to the time shift was expected: in the paradigm with *sliding cues*, the subject can easily predict the time instant when he will be requested to perform IS and start it exactly in the correct instant. Also the signal preceding this moment can be used for classification purposes as preparatory time: indeed the subject is expecting the cue knowing when and what he is asked to perform and ready to go as soon as the cue is given. Instead, for *color-changing cues* paradigm, the subject will not be able to properly perform IS at the instant when the cue is given, since a longer reaction time is needed when the cross changes color. Nevertheless, a higher detection accuracy is obtained for the optimal window in *color-changing* paradigm: it suggests that the subject could keep the consistency among trials in the response and in the timing also in the *color-changing cues* paradigm.

4.4 Paradigm selection and hyperparameter tuning results

Using the two optimal windows for each paradigm, classification was addressed via a 10-fold CV to identify which of the two paradigms obtains better results in terms of detection accuracy. In this classification problem, the epochs corrupted by eye blink were discarded only for the *sliding cues* paradigm. The datasets were balanced between "NONE" and "IMAGINE" classes by downsampling the more represented class. 189 epochs per class are considered for the *sliding cues* paradigm, and 195 for the *color-changing cues* paradigm. The model used is described in section 3.4.1. For this analysis only the time window was adapted, while the other hyperparameters were kept fixed to reasonable values (previously defined in section 3.4.2). The two classification accuracies obtained are 66% (by using the trial-wise splitting strategy; instead, 72% with random strategy) and 63% respectively for the paradigm with *color-changing cues* and with *sliding cues*. Both the paradigms surpassed chance level (50%) and the respective upper confidence (α=1%) limits (56.5% and 56.6% – they are different for the slightly different number of data).

For the core analysis of this work, it has been decided to go on with the *color-changing cues* paradigm. Surely for having shown better discriminability properties, but the main reasons are its comparability with the state-of-the-art and the possibility to investigate the effect of the splitting strategy: indeed, by applying the two different strategies, two different levels of accuracy were achieved. Finally, also for the subjects, it can be a less fatiguing protocol: indeed, for the *sliding cues* paradigm an effort must be conducted to minimize the movements of the gaze during the experiment.

Once the *color-changing cues* paradigm was selected, the model hyperparameters were optimized: so, the obtained values will be used for the evaluation of the whole *corpus* experiments (described in chapter 5). This analysis was conducted via a grid-search as explained in section 3.4.2. For the time window optimization, the results are coming from Table 3: the best window would be 0.5-2.5s, but it was preferred to use the preceding one (0.25-2.25s) since the difference in perfomances is minimal

and it is closer to the go-cue. For the two other hyperparameters, the results of the grid-search are reported in Table 4. The best combination is K=200 and C=1.

Table 4: Hyperparameters tuning. CV accuracy score of each combination. The best is signaled by an asterisk.

K \ C	**0.5**	**1**	**5**
20	54.6%	54.6%	54.9%
40	56.9%	60.3%	59.5%
60	59.7%	62.8%	61.3%
80	59.2%	64.4%	61.5%
110	58.7%	62.8%	63.3%
140	60.7%	65.6%	62.8%
200	57.4%	**68.7%***	63.1%

5. Off-line imagined speech detection

The pilot study has demonstrated that IS detection can be performed with accuracies surpassing chance-level. To enhance the generalizability of these findings, subsequent investigations were conducted on a broader cohort of healthy subjects. The primary objective of this wider investigation is to ascertain whether the same detection model architecture can effectively operate across multiple subjects. Given the huge inter- and intra-subject variability of EEG signals, each session will be analysed by itself. Additionally, the impact of various data splitting strategies on the reported accuracy is explored, enabling a fairer comparison with the state-of-the-art performances. An examination of the influence of subjects' experience with the IS paradigm is undertaken by recording multiple sessions with the same participants. This investigation aims at elucidating if familiarity with the paradigm affects classification performance. In addition to these quantitative analyses, qualitative assessments are conducted to evaluate the strategies employed by different subjects when performing IS.

Furthermore, an investigation is made into the potential influence of confounding factors. Even though factors as eye blinks and head movements (influencing EEG signal by possibly adding some EMG component) are not directly related to IS, they can influence classification performance, possibly in a positive way if some consistency in specific classes is found. Hence it is important to evaluate their effect to be sure that the achieved accuracy scores are determined by decoding neural signal and not by capturing unvoluntary, residual artifacts. Head movement will be investigated by the analysis of the motion and orientation signals recorded by Mentalab Explore+.

5.1 Materials: experimental procedure and statistical analysis

The whole *corpus* experiments are performed on 6 subjects (2 females; aged between 23 and 30; without neurological complaints). All the participants received information about the procedures and objectives of the experiments and provided written consent before taking part. Experiments were pre-approved by the Ethical Committee of UZLeuven, Belgium

The experiment comprises 2 sessions of about 50 minutes, at least 1 week apart. All the sessions were held with the same protocol, the one based on *color-changing cues*. Each session is composed of 12

blocks (interleaved of 30s pause); in each block, 6 trials belong to the "NONE" class, and 6 trials to the "IMAGINE" class (3 "RIGHT" and 3 "LEFT"). The trials are randomly sorted. Each session results into 216 repetitions for class "NONE" and "IMAGINE" (divided into 108 per each of the two words, "RIGHT" and "LEFT"). The first session of subject 002 had to be discarded because of recording issues due to a bad connection of the ground electrode.

Each session is analysed by itself, no inter-session analyses are performed. The reported accuracies are obtained by averaging the 10 scores of a 10-fold CV. Both the splitting strategies are used and compared. In particular, the influence of the two splitting strategies on the obtained average accuracy is analysed by the Wilcoxon signed-rank test for paired samples. The classifier used is described in section 3.4.1 and its hyperparameters were tuned on the pilot study. Wilcoxon signed-rank test for paired samples is also employed to find significant differences between the first and the second session accuracy scores in each subject (only for the 5 of them with both sessions usable).

The sources of noise are analysed to determine if they can influence the model prediction capability: the eye blink rate in different sessions is reported and the 9 orientation/motion channels acquired (3 about linear acceleration, 3 about angular orientation and 3 about angular acceleration) are used to classify the epochs into the two classes. The used model is the same as for EEG processing, except for the block considering the ratio between functional bands, which is excluded. This model is tested with a 10-fold CV process (trial-wise splitting strategy).

Finally, the TF classification visualization is used on each subject to represent the different strategies employed by the subjects and how they differ.

5.2 Results

5.2.1 IS detection accuracy

The achieved CV accuracies for each session are reported in Table 5 and represented in Figure 13a; both the strategies for splitting procedure were tested and they are shown to obtain different performances. According to the Wilcoxon signed-rank test (Figure 13b), when using a 10-fold CV to evaluate the classification model, employing a random splitting strategy leads to significantly ($p<0.001$) different accuracy scores than using a trial-wise splitting strategy. In particular, the accuracy achieved when the model is evaluated with random created folds is higher (69.6±11.6%) than with trial-wise splits (63.8±13.2%).

Table 5: IS detection accuracy – splitting strategy comparison

Session	Sub 000		Sub 001		Sub 002		Sub 003		Sub 004		Sub 005		**Mean ±std**	
	RAND	T-WISE	RAND	T-WISE	RAND	T-WISE	RAND	T-WISE	RAND	T-WISE	RAND	T-WISE	RAND	T-WISE
1	70.8%	61.4%	58.8%	48.2%			62.5%	55.9%	61.6%	60.6%	86.3%	84.5%	**69.6% ±11.6%**	**63.8% ±13.2%**
2	74.3%	67.2%	77.8%	74%	53.2%	47.2%	68.3%	62.1%	61.8%	54.5%	89.6%	86.6%		

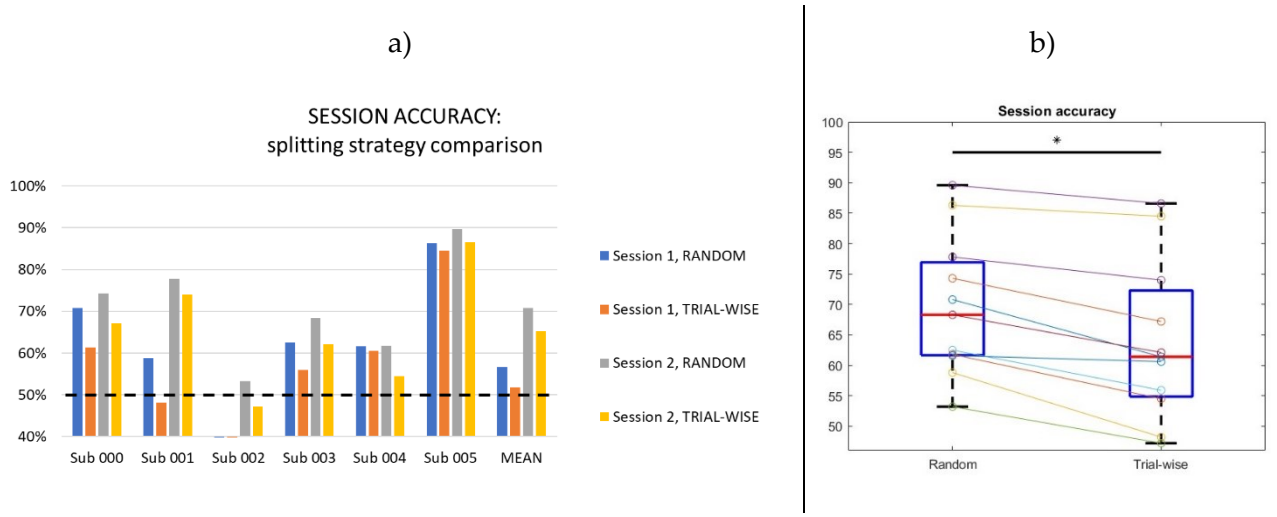

Figure 13: Splitting strategy accuracy comparison. In a) the detection accuracy achieved in different sessions are shown. The two different splitting strategy are represented for each session. b) represents the boxplots of the accuracy levels achieved in each session by using the two different strategies (paired samples); $p<0.001$.

From now on, only the trial-wise splitting strategy will be considered (Table 6 and Figure 14), since by using it, it is possible to train models only employing IS related information to classify the trials. For all the sessions, with exception for the second of subject 002, an accuracy above the chance level is achieved. Except for that subject, the confidence (α=1%) upper limit of chance level (56.1%) is surpassed by everyone, at least in one of the two sessions. As a final average along all the subjects 62.1% is reached in the first session, while 65.3% for the second session. The best performances were achieved for subject 005 who reached 84.5% and 86.6% in the two sessions; subject 001 with 74% in the second session; and subject 000, with 67.2% in the second session.

It can be noted that a trend of improvement from the first to second session is present in all subjects except for subject 004 (Figure 14b). Nevertheless, a statistically significant difference between the two conditions was not found.

Table 6: Trial-wise splitting strategy – IS detection accuracy. CV accuracy scores for each session. The values surpassing the upper confidence boundary are signaled by an asterisk.

Session	Sub 000	Sub 001	Sub 002	Sub 003	Sub 004	Sub 005	**Mean ±std**
1	61.4%*	48.2%		55.9%	60.6%*	84.5%*	**62.1% ±7.4%**
2	67.2%*	74.0%*	47.2%	62.1%*	54.5%	86.6%*	**65.3% ±9.9%**

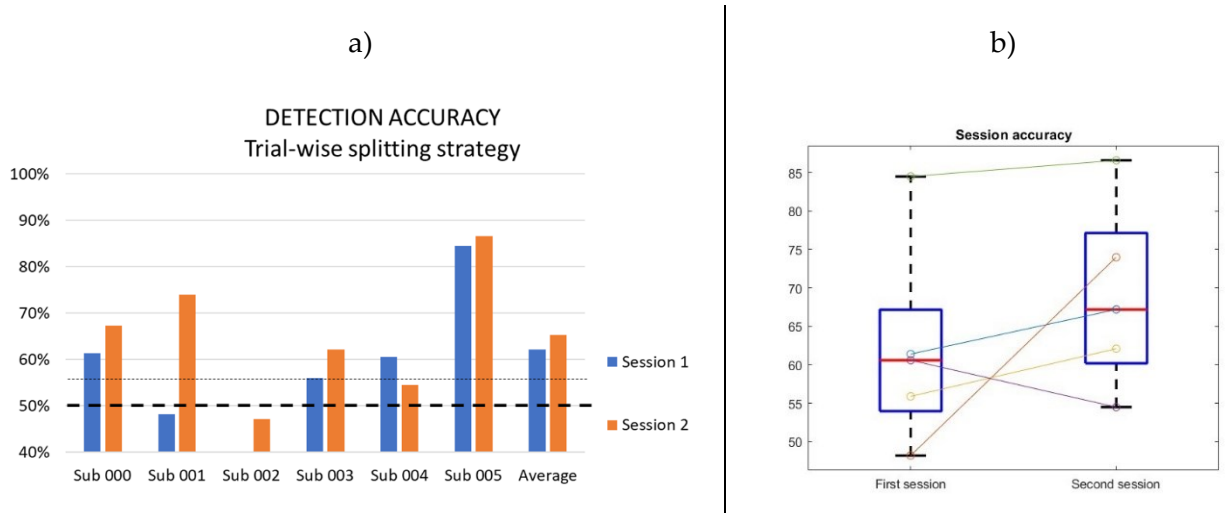

Figure 14: Trial-wise CV IS detection accuracy. In a) the accuracy achieved in each session are represented per subject. In b) the boxplot represents the comparison between the accuracy in the first and in the second session (no significant difference was found).

5.2.2 Eye blink artifacts

Eye blinks were detected using the ICA algorithm (see section 3.3.1). During the experiments, the subjects were asked to minimize the eye blinks, especially during the periods when a task was performed. Nevertheless, a high number of repetitions were corrupted by eye blinking during the first session. Hence, for the second sessions, the instruction to avoid blinking during task periods was emphasized, and subjects were able to reduce the rate with respect to the first session: as it can be seen in Table 7 in average the rate of corrupted epochs in "IMAGINE" class went from 6% to 3% and or "NONE" class from 15% to 5%.

Table 7: Eye blink rate. The percentage of corrupted epochs per class.

Session	Sub 000		Sub 001		Sub 002		Sub 003		Sub 004		Sub 005		**Average**	
	IMAGINE	NONE	IMAGINE	NONE	IMAGINE	NONE	IMAGINE	NONE	IMAGINE	NONE	IMAGINE	NONE	IMAGINE	NONE
1	10%	10%	1%	7%			15%	25%	1%	15%	2%	20%	6%	15%
2	5%	3%	0%	10%	3%	2%	1%	2%	7%	7%	2%	6%	3%	5%

When many epochs are corrupted by eye blinks – in the first session –, they are much focused in the "NONE" trials; instead, in the second session they tend to be more balanced between the two classes. A hypothesis could be that the decoder may be helped by an unbalanced presence of epochs corrupted by eye blinks in the two conditions. However, it looks like this is not the case: indeed, accuracy performances were kept almost the same, and even got slightly better from the first to the second session while the blink unbalance reduced.

5.2.3 Classification based on head movement signals

The classification accuracy achieved by using head motion and orientation signals are reported in Table 8.

Table 8: Head movements detection accuracy. The values surpassing the upper confidence boundary are signaled by an asterisk.

Session	Sub 000	Sub 001	Sub 002	Sub 003	Sub 004	Sub 005	**Average**
1	52%	52%	59%*	51%	52%	57%*	54%
2	53%	63%*	63%*	50%	51%	58%*	56%

Only few cases are above the upper confidence limit for chance level (marked by an asterisk), and it does not always correspond to good perfomances in classification accuracy coming from EEG. The maximum achieved accuracy is 63%, and average is 54% for the first session and 56% for the second. Both values are below the upper confidence limit for chance level. Although it does not have a clear correlation with the detection accuracies achieved by the EEG pipeline and the performances reached by using orientation data are low, it should be taken into consideration that the two best performing sessions for EEG-based IS detection (subject 001 session 2 and subject 005) are also among those achieving best results when the orientation signals are used (subjects 001 session 2, subjects 002 and 005).

5.2.4 TF classification maps

In Figure 15 the TF classification performances in narrow time window at different frequency bands are shown. For each subject the best session in terms of maximum achieved accuracy is presented.

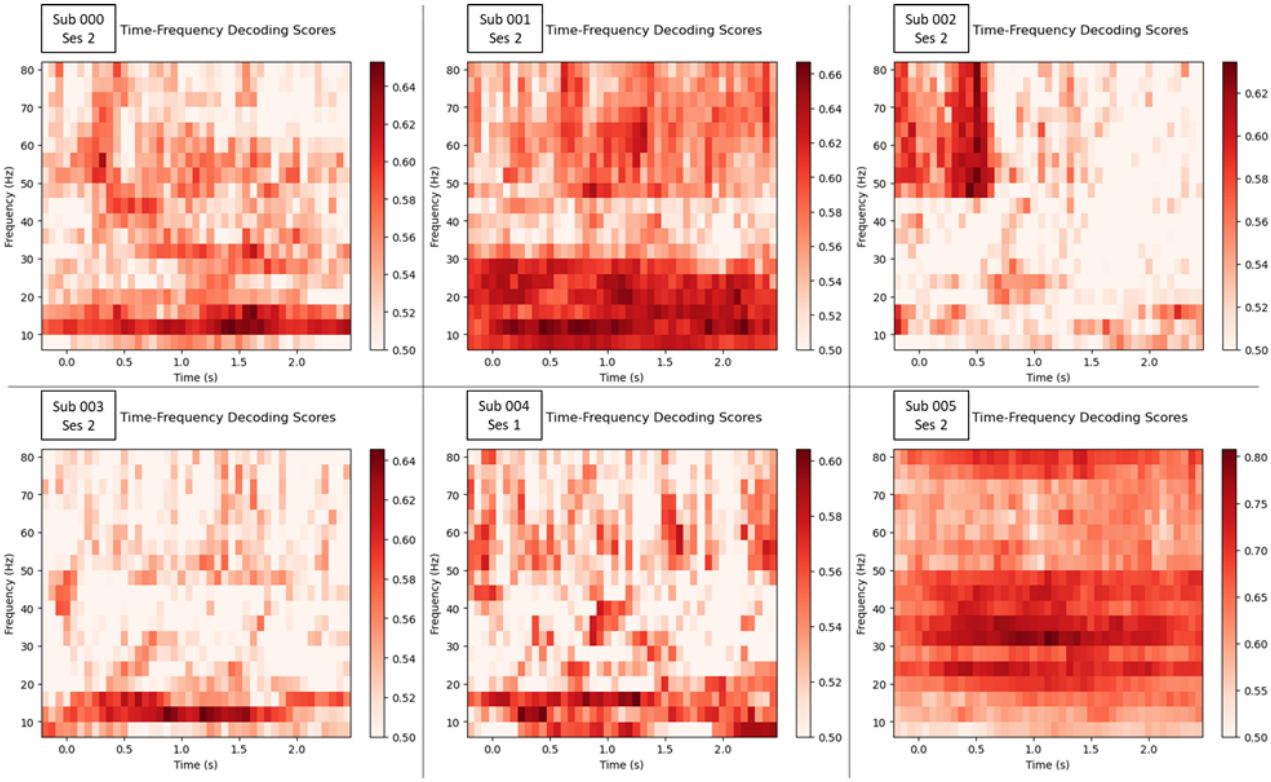

Figure 15: TF classification accuracy maps of the best session for each subject. The red scale encodes the accuracy obtained in each narrow time window by using the specific frequency band (4Hz interval).

For all subjects except for subject 002, the frequencies in β range show a high discriminability. For subject 001, 002 and 005, high-γ band has also huge relevance as classification power, while for other subjects it is less important, although it still contains some information. It is also interesting to note that the timing of best discriminable features is different: for subject 000 the peak in classification performance is happening at about 1.5s after the cue, for subject 003 between 0.75s and 1.5s while for subject 004 around 1.25s. For subjects 001 and 005, the whole time window from 0s to 2.5s is relevant for discriminability. It is interesting that for subject 002 (the only not surpassing chance level in the previous analysis), the relevant time window is strongly focused in the first 0.75s after the cue and in particular the most discriminative frequency range is the γ band. Notably, for this subject the accuracies obtained in this evaluation based on a very reduced pool of information are much higher than those achieved by the model using the whole information (62.8% as a peak in TF classification map against the 47.2% of the model presented in Table 6).

5.3 Discussion

5.3.1 Importance of a correct splitting procedure

The Wilcoxon test has shown statistical significance in demonstrating that the choice of splitting strategy for creating CV folds is crucial and has a significant effect on the reported accuracy. A fair procedure to create the splits for CV model evaluation is important, echoing the caution raised by [40]. It has been demonstrated that with a random splitting strategy, models can easily exploit the non-stationarity of EEG signals to achieve higher accuracy, as shown in [28]. Specifically, the three repetitions associated with the same instruction share certain features unrelated to the imagination process but arising from their temporal proximity. Hence, for the purpose of IS detection these time-related features are unsuitable for online classification. Their impact should be minimized in offline analysis to ensure a valid assessment of the model's ability to detect IS.

When multiple cued repetitions follow the same instruction, a trial-wise splitting procedure coupled with a random sorting of instruction within blocks – as employed in this study – removes the influence of these proximity features. Indeed, the model cannot rely on the temporal features of the epochs in the training set to predict an epoch of the testing set. Conversely, employing a random splitting procedure has been demonstrated to inflate model accuracy by exploiting this effect. Indeed, the prediction of one of the three consecutive repetitions may be mainly based on its time-related similarity to the other two repetitions contained in the training set. In this study, it was revealed an average accuracy difference of 5.8%, but overall, it's reasonable to presume that this effect becomes more pronounced when a greater number of repetitions related to the same instruction is performed (as in [18] where 4 repetitions follow the same instruction).

Of course, an additional method to address this issue is the randomization of the cue presentation order. Therefore, it is necessary to critically evaluate the performance reported in the state-of-the-art and exercise caution when implementing such paradigms, with particular attention to this aspect.

5.3.2 Comparison with the state-of-the-art

Not all the studies provide accuracy measures that are directly comparable to the results presented in this study. For instance, in [18], an average accuracy of 80% across subjects is reported for detection using both all 64 electrodes and only the 10 electrodes covering speech-related scalp regions. The experimental protocol in [18] bears similarities to the one employed in this study, albeit with one repetition more (4 repetitions per instruction as opposed to 3). It's worth noting that the CV fold splitting strategy used in [18] is random split creation which was shown to inflate accuracy

when associated to multiple repetition paradigms. In the current study, with one less repetition and 8 channels instead of 10, the average accuracy across subjects in the second session is found to be 69.6% when the random split strategy is used. Given that the presence of an additional repetition could contribute to further inflate reported accuracy by [18], in a fair comparison, the experimental setup based on Mentalab Explore+ presented in the current study achieved competitive performance in terms of IS detectability. However, since these scores are achieved by employing splitting strategies allowing the temporal features influence, it cannot be ensured how much these models – hence these accuracy scores – relate to IS or to temporal proximity features for classification.

On the other hand, considering studies where strategies comparable with our trial-wise procedure are employed, it is noteworthy to mention the study conducted by [28]. In their research, a detection accuracy of 58% is reported using a similar splitting strategy and a 64-channel clinical grade EEG system. In comparison, the model we proposed demonstrates a significant improvement, achieving an accuracy of 65.3%. For the IS detection problem, a better accuracy was achieved by [31] (83%) but in that case it was not used a single word IS, but the subjects were asked to continue imagining saying the word for 4s, making this paradigm different than the one investigated in the proposed study, and less usable in application-oriented modality. The only work analysed where a portable headset is used to cope with IS paradigm [32] reports a binary accuracy for decoding the words "YES" and "NO" of 58% (without any sorting randomization): this is much lower than the accuracy of 65.3% we reached in this study, demonstrating the contribution of our work.

To the best of our knowledge, there are no other studies in the current state-of-the-art literature that focus on IS detection using a portable device. Therefore, it becomes challenging to conduct a direct comparison between the results achieved with our setup and model, which attains an average detection accuracy of 65.3%. Considering the 70-75% accuracy threshold to have a usable system for online applications [2], it is not achieved for all the subjects. Nevertheless, some of them (subjects 005 and 001) are successfully reaching that level or at least approaching it (subject 000). This allowed to try an online implementation to see whether it results usable and how providing feedback to the user can influence its way of performing the task.

5.3.3 Non-IS artifacts influence

The blink rate and the head movements (analysed through motion and orientation signals) were analysed because they may represent some source of information which could possibly help the decoder to improve its classification capabilities, although they are not directly related to the main IS task. To implement a real assistive technology based on IS, the lonely contribution of IS should be estimated hence it is important to look at the influence of those other sources on the model and avoid them to inflate system evaluation.

The blink rate could be considered not to influence the core of classification because while the difference between the ratio of "IMAGINE" and "NONE" epochs corrupted by eye blinks is dicreasing from the first to the second session the classification accuracy is steady, with a slight improvement. So, it looks like the model is not making the decision based on that information source.

On the other side the prediction based on the motion and orientation signals representing the head movements are clearly much worse with respect to those coming from EEG suggesting that the main decisional processes are based on brain sources. Nevertheless, the sessions where the motion signals lead to better classification results (subject 005, subject 001 session 2, subject 002 session 2) are those where the γ frequency band is showing better discriminability according to the TF classification

maps. Clearly, undesired head movement come from residual muscolar activation, which generates electrical signal (EMG). This suggests a possible effect on model's ability to detect IS by the residual EMG signal, which could be captured by scalp electrodes mainly in γ band, as it was claimed in [27]. Although movements are maximally avoided (or at least, subjects are asked to minimize any movement), it is possible that some undesired, residual muscular activity is still present and it may influence the predictions of the model: indeed a complete suppression of articulator movement is not easy to achieve. For sure, a more detailed analysis is needed to reveal and quantify the real contribution of the undesired EMG signals which are not directly related to IS. To do it, a set-up using also channels measuring EMG is needed, and possibly a differential analysis could be performed to evaluate its effect.

5.3.4 Inter-subject variability and most discriminable features

Although no statistical difference can be found, an improvement of the performances of the decoder from the first to the second session was found in all the subjects except for subject 004. This could be given to a higher familiarity of the subject with the paradigm which can lead them to be more comfortable with the process of imagination during the second session. Inter-session variability is not investigated, but considering the difference in performances, for sure it would be found some difference in the imagination processes.

The TF classification maps obtained from different subjects clearly indicate that the most discriminative areas in the TF plan vary from one subject to another. Each subject appears to employ a unique strategy for IS, often with diverse timing. This diversity in strategies limits the feasibility of working with transfer-learning. It also introduces limitations and potential challenges when applying hyperparameter tuning strategies such as the one proposed in this study. Indeed, the model hyperparameters, including the optimal time window, were fine-tuned based on the pilot study. However, it's important to note that what works optimally for one subject may be suboptimal for others. For example, in the case of subject 002, who achieved very low accuracy with the complete model, the TF classification representation clearly shows that using a time window from 0.25s to 2.25s after the cue is not the optimal strategy. The most discriminative information in the TF plan appears to occur up until 0.75s after the cue. The subsequent time intervals contain less discriminable information. This observation may explain why the accuracy in that session fell below chance level, despite peak accuracies of up to 62.8% being achieved in the TF classification map.

Considering this high inter-subject variability, other strategies to optimize hyperparameters should be adopted and investigated: an example is the use of nested CV, which has, on the other side, the computational load limitations discussed in section 3.4.

In addition to providing insight into the strategies employed by subjects during the task, the TF classification maps can also offer valuable information about the most relevant spectral features for decoder creation. Although the plots are very different among each other, generally β band covers an important role. This observation aligns with previous research in the field, including the ECoG study by [29], which identified the β band as crucial for neural encoding of IS. Moreover, EEG studies, such as [18], have ranked the β band as the second most important frequency band for this purpose, following the γ band. The γ band has consistently emerged as the most informative frequency range for IS in other EEG studies like [27] and [66]. [66], for instance, employed a TF plan visualization to pinpoint regions with the most significant features but combined data from all their 4 subjects, making it challenging to distinguish individual subject strategies. Nevertheless, they noted the highest relevance in the frequency range above 70Hz. In our work, the TF classification

map also shows the γ range to be relevant. However, it's essential to highlight again that this band is particularly susceptible to muscular artifacts and may be associated with residual EMG activity happening during imagination due to unvoluntary facial movements. A more complex set-up, including EMG channels, is necessary to properly analyse and mitigate these influences.

5.4 Conclusion

In the offline experiments conducted across two sessions with a total of 6 subjects, the IS detection accuracy reached 65.3% on average across all subjects, with individual peaks of 86.6% and 74%. These results position the proposed setup and model within the range of state-of-the-art performances. The novelty of our study lies in two aspects: firstly, significant attention was dedicated to employing a fair CV splitting strategy, ensuring that the model predictions are based solely on IS information rather than exploiting EEG non-stationarity characteristics. Secondly, the biggest added value of our work lies in the use of a portable EEG cap with only 8 channels. The setup time is remarkably brief, taking just 5-10 minutes, and the device lightweight and portable design significantly improves the comfort for users.

Such good performances open the possibility to implement a system giving real-time feedback about the ability of the predictor to detect IS. In some subjects, the usability threshold of 70% [2] has already been surpassed in offline sessions, and it is nearly reached for the majority. Furthermore, the easiness of use of the device, connected via Bluetooth to the recording laptop, has enabled the developement of a real-time pipeline. The real-time model will be trained on the data coming from the same session, given the high variability which can occurr between sessions recorded in different days. Details on the implementation and performance of the real-time system will be provided in the following section.

6. Online feedback imagined speech detection

6.1 Previous studies

EEG has not been much used for IS paradigms in online frameworks. At the state-of-the-art, only two studies are available [67], [68]. In [67], two binary BCIs were developed: one for detecting IS of the word “NO” over a resting condition (detection) and another to decode the imagination of the words “YES” and “NO”. The employed paradigm required the users to continuously imagine the word (or rest) for 10s, and the classification is performed on the features extracted from those 10s of EEG signal. 64 EEG channels were used. In the online sessions, providing feedback to the user, accuracies of 76% and 69% were achieved respectively in the detection and decoding problems.

In [68] instead, single imagined repetition of the words is performed, similarly to the paradigm adopted in the current study. A ternary BCI was implemented having as classes a resting condition and two words. Users received feedback about the model predictions. The achieved average accuracy for the online sessions was 75%, with real chance level computed to 60%.

It's often suggested that same-session data is more valuable for BCI classifier training than data from previous days [69]. Indeed, both studies trained the model on data coming from the same day: in particular same-day data were oversampled in [67] to get more relevance with respect to data coming from previous sessions which were included in the training set. While in [68] they only used data coming from the same day. Both kept updating the training set after each block of the online session with the most recent available data to create an adaptive model.

6.2 Materials

6.2.1 Experimental procedure

Online feedback experiments were conducted only on three out of the six subjects involved in the whole *corpus* experiments (subjects 000, 003, and 005). The experimental protocol is detailed in the section 3.2.2 and is based on the *color-changing cues* paradigm. The model's architecture for online IS detection remains consistent with the one employed in offline sessions.

The entire session comprises 12 blocks, with each block containing 6 trials for the "NONE" class and 6 for the "IMAGINE" class (3 for "LEFT" and 3 for "RIGHT"). For each trial 3 consecutive repetitions are distinctly cued. In the initial 6 blocks, no feedback is provided to the user (the protocol is the same as for the offline sessions). After the sixth block, 6 online blocks follow. The model is trained using data from all the preceding blocks of the day and is used to detect IS in the subsequent block, providing feedback to the user. The model is retrained after each block, incorporating also data from the last block, resulting in an adaptive model that accumulates data from all previous blocks (the six offline ones and all the performed online). Model training occurs before each online block and takes no longer than 90s. During this time, the user is informed about the ongoing training process and the accuracy scores from previous blocks. In total, 6 blocks (equating to 36 trials, or 108 single repetitions per class - "NONE" vs "IMAGINE") are conducted without feedback (offline blocks), and other 6 blocks with feedback (online blocks). The first 6 take approximately 22 minutes, while the 6 online blocks require about 50 minutes to complete (due to the increased inter-block pause for training, and to the elongation of each trial for providing feedback).

6.2.2 User-specific hyperparameter tuning

As previously mentioned (section 5.3.4), the optimal hyperparameters may vary between subjects. Consequently, a dedicated tuning phase was conducted for the 3 subjects participating in the online feedback session. Leveraging data from their most successful offline session, the model hyperparameters, including the starting instant of the time-window, the regularization term of the SVC (C) and the number of features (K), were fine-tuned via a grid-search process, as detailed in section 3.4.2. Initially, the time window is optimized, and subsequently, 21 possible combinations of the two parameters are tested. The best-performing combination is then used for the online session.

6.3 Results

6.3.1 Online IS detection accuracy

In the online sessions, feedback is only provided during the last 6 blocks. The accuracy results for each of these 6 blocks (with 18 repetitions per class) are presented in Table 9. The upper confidence boundary (significance α=1%) of a random classifier predicting the trials of one block is 70% due to the low number of repetitions in each block [63].

For subject 000, the initial block starts at the chance level and exhibits a gradual improvement, eventually reaching peak accuracies of 83.3% and 77.8% in blocks 10 and 12, where the model is significatively better than random. For subject 005, in the first four online blocks, the real-time feedback was not properly working; instead in the last two blocks, the real-time feedback was given correctly 88.9% of the times (significatively better than random). Finally, for subject 003, the model never got much far from the chance level, reaching a peak at the last block of 63.9%.

For subjects 000 and 005, the model was able to provide real-time feedback significantly better than random and with usable and trustable outcomes for the user. They both display a common pattern.

The feedback provided in the initial blocks appears to be meaningless, but later, after a different number of blocks, it begins to deliver correct real-time predictions.

Table 9: Online detection accuracy divided in blocks. In each blocks the support is 18 epochs per class. 1% confidence upper boundary for the random classifier is 70%: the blocks with performances better than random are indicated with an asterisk.

Subject	BLOCK 7	BLOCK 8	BLOCK 9	BLOCK 10	BLOCK 11	BLOCK 12
000	50%	55.5%	69.4%	***83.3%**	61.1%	***77.8%**
003	52.8%	47.2%	55.6%	61.1%	55.6%	63.9%
005	50%	61%	50%	50%	***88.9%**	***88.9%**

6.3.2 Post-hoc offline analysis

On the data acquired during online sessions, different models were later evaluated, during an offline analysis. The model architecture was kept the same, while it was changed the origin of the training set. In Figure 16 there are reported the accuracies achieved on each block for the first group of models which are not properly usable in online settings:

- Leave-One-Block-Out (LOBO): the model is trained on the trials coming from all the blocks except one (the training set is made of 11 blocks). The reported accuracy is obtained on the specific left-out block.
- LOBO-only-without-feedback (LOBO-WO): the same as the LOBO strategy, but only the first 6 blocks (where feedback is not provided) are considered. So, the training set is composed of 5 blocks.
- LOBO-only-with-feedback (LOBO-W): the same as above, but only the last 6 blocks are considered (only those where feedback is given).

In Figure 17, the second group of models is reported. They are usable in online settings:

- Cumulative: it is the model used in the online blocks (the last six) to provide feedback. The training set includes all the data (both without and with feedback) recorded before the tested block. The training set size increases at each block.
- Cumulative-only-feedback: the model is trained on all the trials coming from the blocks preceding the test block but only starting from the seventh block. It is only trained on trials where feedback was provided. For this reason, its predictions can only be given starting from the eighth block. The training set size increases at each block, but is much lower than the cumulative model (6 blocks less).
- Adaptive buffer 1- (or 3- or 6-) block: the model is trained on the 1 (or 3 or 6) block(s) preceding the tested one. It only uses the most recent data, and for all tested block, the model is trained on the same number of blocks (1, 3 or 6).
- Fixed: the model is trained on the first 6 blocks (the ones without feedback) and is used to evaluate each of the 6 online blocks. The model never changes.

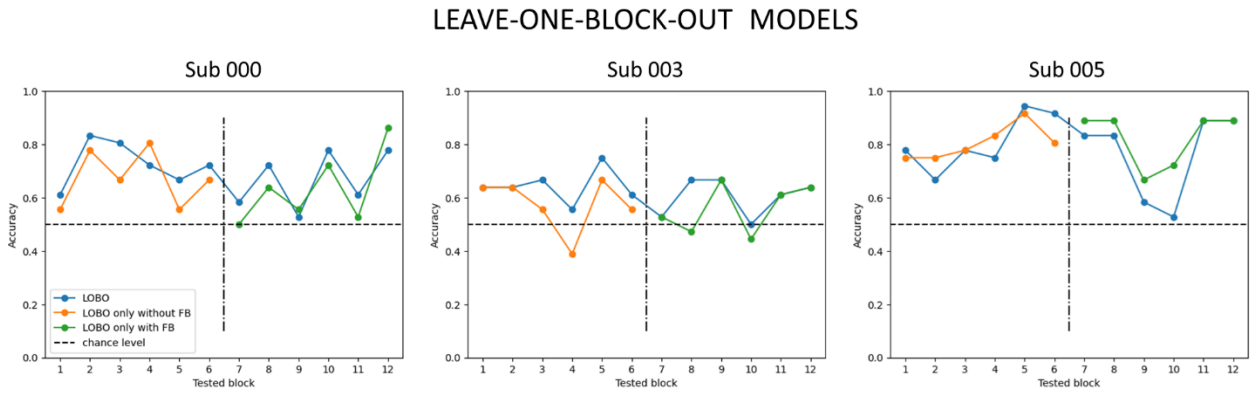

Figure 16: Block-wise accuracy obtained in post-hoc analysis by LOBO models for different subjects.

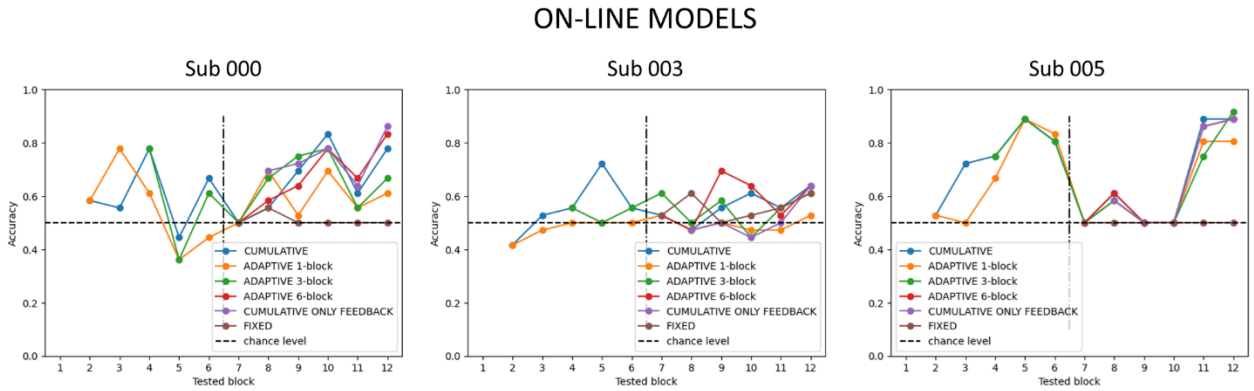

Figure 17: Block-wise accuracy obtained in post-hoc analysis for single subjects. Horizontal dashed line is chance level; vertical dashed-dotted line shows when the session switched from offline to online (after the line feedback was provided). All these models are potentially usable in real-time settings. The one used during the experiment is represented by the blue line (cumulative), but in this plots also the possible predictions for the first offline blocks are shown.

It's worth noting that the LOBO model, despite having a significantly larger training dataset compared to the LOBO-W or LOBO-WO, does not yield superior performance. Moreover, a common trend emerges, particularly for the online blocks: the same blocks where LOBO performs relatively better are also the blocks where the LOBO-W model exhibits better performance.

On the other hand, when considering models suitable for online applications, it's particularly striking that, in all three subjects (with a specific focus on subjects 000 and 005), the fixed model barely surpasses the chance level. For the two subjects who achieved good results, the cumulative-only-feedback model appears to be the most effective, or at least at the same level of the cumulative model which has many more available data to train on. Notably, accuracy for the seventh block is not provided for the cumulative-only-feedback model since the first online block is only used for the initial training of the model and prediction cannot be performed due to the absence of preceding trials with feedback.

6.3.3 TF classification maps

Finally, the TF classification maps are used to show the strategy employed by the users for imagination. In Figure 18, the maps depict the first 6 blocks (without feedback) and the last 6 blocks (with feedback) of each subject. The total number of repetitions per class is the same for both the groups. However, these maps do not reveal any distinctive patterns, aside from a notable difference between the two sets of blocks (trials without and with feedback) even within the same subject. Despite the relatively small time gap (data are collected on the same day, with only a few minutes in between), there are limited discernible similarities between them. Furthermore, it's apparent that, when feedback is provided, the regions in the TF plan that have a greater influence are more focused and less spread around; on the other hand, the maximum accuracy obtained decreases.

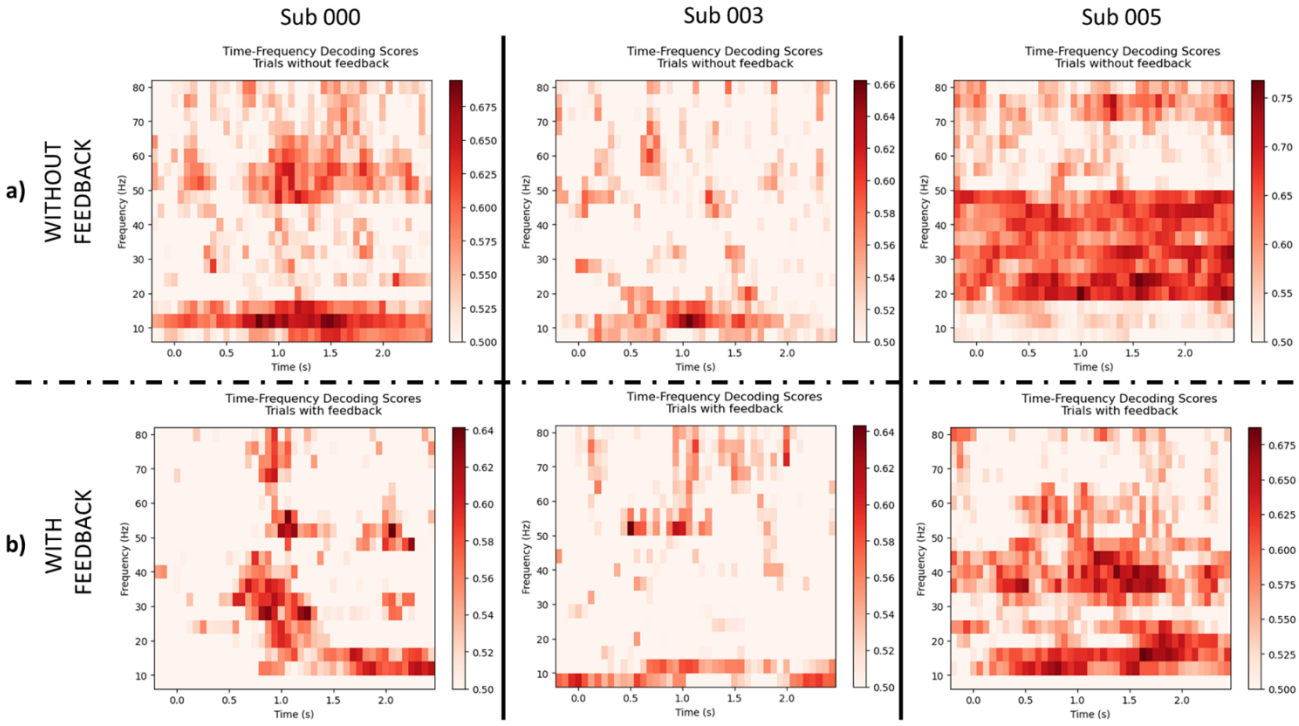

Figure 18: TF classification maps for the 6 blocks without (a) and with (b) feedback.

6.4 Discussion

6.4.1 Online IS detection feasibility: a bilateral learning process

The ability to analyse data in real-time and provide feedback to the user is a critical step toward creating a usable BCI application. However, to gain the user's trust, minimal levels of prediction accuracy, approximately around 70% should be achieved [2]. These accuracy levels are not consistently reached in offline experiments. On the other hand, the nature of the human brain, capable of learning and adapting its behaviour suggests that closing the BCI loop through the provision of feedback has the potential to support the adjustment of the user's neural network over time, which could lead to improvements in BCI performances [67], [70]. This is why it is essential to conduct experiments where neurofeedback is provided, allowing the user to learn how to optimize their strategy in a mutually beneficial learning process involving both the user and the system. In

the experiment, this process was observed, especially for subjects 000 and 005. Initially, classification accuracy is just about chance level for the first online blocks and then it increases over the time with a different speed, eventually achieving high accuracy levels for the detection problem, surpassing those reached in offline sessions. While the improving trend in accuracy over consecutive online blocks was also noted in [67], it was primarily attributed to the increasing availability of data for training. Instead, in our case, the improvement in performance in the last blocks can be attributed to an enhancement in data quality, since post-hoc accuracy analysis of models with a constant training set size exhibited the same trend.

It's important to note that the mean online accuracy obtained in our study is lower than that of [67] and [68]. However, it should be recognized that in [67] subjects were demanded to continuously imagine repeating the word for 10s, not performing a single repetition trial and that both studies used 64-channel EEG, whereas here only 8 channels were employed. Moreover, the trend observed in subjects 000 and 005 suggests that if more blocks were performed, higher accuracy levels could have been maintained. In fact, the peak values are achieved in the final blocks are very high, reaching 77.8% and 89.9%. This demonstrates the potential of this system and paves the way for further research in implementing an IS BCI system providing online feedback based on Mentalab Explore+, a portable 8-channel EEG setup.

6.4.2 Strategy changes with and without feedback

The visualization of the TF classification maps clearly reveals that subjects employ different strategies when feedback is provided compared to when it is not. In a sense, it would be expected that trials would become more consistent when feedback is offered, given that users receive indications about the ability of the system to comprehend their strategy through neurofeedback. Indeed, the most discriminable areas get shrinked in the TF map for online blocks and focus on more specific area: they are less spread out. Nevertheless, the maximal accuracy levels achieved are lower.

On the other hand, it's essential to highlight that in subjects 000 and 005, the fixed model (trained only on trials without feedback) fails to make predictions beyond chance levels for online blocks: this emphasizes how distinct the way to perform IS is when feedback is provided compared to when it is not. The offline blocks seem to have limited relevance in predicting the blocks with feedback.

Although the two paradigms (online and offline) are fundamentally similar, the timing differs; the rhythm changes because when feedback is given, each trial includes a feedback period, adding 3 extra seconds. The blocks become longer, and also the inter-block pauses are extended due to the model training process, which can take up to 90s, as opposed to the 30s pause between offline blocks. On the other hand, the purpose of providing feedback is to encourage users to refine their strategy, making the task more predictable to the system. Hence, this difference was overall expected and since it is so visible, it suggests the potential use of a different model for future investigations. Indeed, the cumulative model taking into consideration all the offline blocks could be surpassed by the cumulative-only-feedback model. Using the latter in future experiments may allow for more blocks with feedback within the same timeframe (without the need for the lengthy offline training section based on data recorded without feedback, which apparently is less valuable for predicting purposes) and could enhance the number of blocks where feedback functions effectively.

7. Conclusion

This thesis introduces a BCI for detecting IS using a portable 8-channel EEG system, Mentalab Explore+. Notably, such devices have not been employed in previous state-of-the-art studies aimed at developing a single repetition IS BCI, marking a significant contribution to the field. The system was tested in online setting, providing real-time feedback to the user. This demonstrates a substantial advancement, considering that, to date, only a limited number of studies ([67], [68]) have tested their EEG-based IS BCI in real-time.

After an initial phase to familiarize with the tools and the paradigm, experiments were held to create a model able to detect IS in offline analysis from the EEG signal. It performed properly on five out of six healthy subjects. Considering the use of an 8-channel research-grade system, the achieved performances are competitively positioned in a fair comparison with the state-of-the-art, commonly employing EEG systems featuring 64-channel full-scalp setups. The system was then tested online on three healthy subjects, showing its ability to provide trustworthy feedback to the user in real-time while also revealing certain limitations. The results emphasized that an offline training session is less beneficial for the model compared to data acquired with feedback. This suggests that, for future developments of the system, initiating to provide online feedback earlier in the session may be advantageous, as the imagination strategy appears to change in trials where feedback, even if partly incorrect, is given. In the online sessions, the model demonstrated classification performance surpassing the usability threshold of 70% in two subjects out of three; peaks of 89% accuracy were reached in the final blocks. This indicates a bilateral learning process happening, prompting the need for further studies to quantify the contributions of the two adaptive controllers, the user and the BCI system.

Six subjects were tested in offline experiments and only three participated in online sessions. Future studies will be certainly aimed at increasing the pool of participants to increase the significance of these results. Then, real-time feedback about IS detection may be strongly beneficial for increasing subjects' consistency about imagination strategy and expanding the paradigm and the model towards a BCI system able to decode the specific imagined word, dealing with a multi-class classification problem which was not addressed in this thesis.

Acknowledgements

I sincerely thank Professor Van Hulle for giving me the opportunity to develop my thesis project under his supervision in the Laboratory for Neuro- and Psychophysiology of the KU Leuven Medical School. I express my gratitude to Aurélie and Bob for their fundamental daily supervision and guidance throughout the project. A special thanks goes to my advisor, Professor Ambrosini, for her availability in supporting me and providing valuable advice in remote mode during the whole research conducted at the foreign location.

References

[1] J.-D. Bauby, "Le Scaphandre et le Papillon," Paris, France, 1997.

[2] J. R. Wolpaw, N. Birbaumer, D. J. Mcfarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," 2002. [Online]. Available: www.elsevier.com/locate/clinphCLINPH2001764

[3] J. del R. Millán and J. Mouriño, "Asynchronous BCI and local neural classifiers: An overview of the adaptive brain interface project," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 159–161, Jun. 2003, doi: 10.1109/TNSRE.2003.814435.

[4] T. O. Zander and C. Kothe, "Towards passive brain-computer interfaces: Applying brain-computer interface technology to human-machine systems in general," in *Journal of Neural Engineering*, Apr. 2011. doi: 10.1088/1741-2560/8/2/025005.

[5] C. Mühl, D. Heylen, and A. Nijholt, "Affective Brain-Computer Interfaces," in *The Oxford Handbook of Affective Computing*, Oxford University Press, 2015. doi: 10.1093/oxfordhb/9780199942237.013.024.

[6] A. Kübler and K.-R. Müller, "An Introduction to Brain-Computer Interfacing," in *Toward Brain-Computer Interfacing*, The MIT Press, 2007, pp. 1–26. doi: 10.7551/mitpress/7493.003.0003.

[7] D. R. Beukelman, S. Fager, L. Ball, and A. Dietz, "AAC for adults with acquired neurological conditions: A review," *Augmentative and Alternative Communication*, vol. 23, no. 3, pp. 230–242, Jan. 2007, doi: 10.1080/07434610701553668.

[8] F. R. Willett *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, Aug. 2023, doi: 10.1038/s41586-023-06377-x.

[9] D. A. Moses *et al.*, "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217–227, Jul. 2021, doi: 10.1056/nejmoa2027540.

[10] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based Spoken Communication: A Survey."

[11] L. Vygotsky, R. Rieber, and A. Carton, *The Collected Works of L.S. Vygotsky: Volume 1: Problems of General Psychology, Including the Volume Thinking and Speech*. Plenum, 1987.

[12] J. T. Panachakel and A. G. Ramakrishnan, "Decoding Covert Speech From EEG-A Comprehensive Review," *Frontiers in Neuroscience*, vol. 15. Frontiers Media S.A., Apr. 29, 2021. doi: 10.3389/fnins.2021.642251.

[13] S. Saminu *et al.*, "A Recent Investigation on Detection and Classification of Epileptic Seizure Techniques Using EEG Signal," *Brain Sci*, vol. 11, no. 5, p. 668, May 2021, doi: 10.3390/brainsci11050668.

[14] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A State-of-the-Art Review of EEG-Based Imagined Speech Decoding," *Frontiers in Human Neuroscience*, vol. 16. Frontiers Media S.A., Apr. 26, 2022. doi: 10.3389/fnhum.2022.867281.

[15] M. Saeidi *et al.*, "Neural decoding of eeg signals with machine learning: A systematic review," *Brain Sciences*, vol. 11, no. 11. MDPI, Nov. 01, 2021. doi: 10.3390/brainsci11111525.

[16] A. Libert and M. M. Van Hulle, "Predicting premature video skipping and viewer interest from EEG recordings," *Entropy*, vol. 21, no. 10, Oct. 2019, doi: 10.3390/e21101014.

[17] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topogr*, vol. 2, no. 4, pp. 275–284, 1990, doi: 10.1007/BF01129656.

[18] S. H. Lee, M. Lee, and S. W. Lee, "Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2647–2659, Dec. 2020, doi: 10.1109/TNSRE.2020.3040289.

[19] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans Biomed Eng*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008, doi: 10.1109/TBME.2008.921154.

[20] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans Biomed Eng*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005, doi: 10.1109/TBME.2005.851521.

[21] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Front Neurosci*, no. MAR, 2012, doi: 10.3389/fnins.2012.00039.

[22] L. Yang and M. M. Van Hulle, "Real-Time Navigation in Google Street View® Using a Motor Imagery-Based BCI," *Sensors*, vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031704.

[23] F. Yger, M. Berar, and F. Lotte, "Riemannian Approaches in Brain-Computer Interfaces: A Review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10. Institute of Electrical and Electronics Engineers Inc., pp. 1753–1762, Oct. 01, 2017. doi: 10.1109/TNSRE.2016.2627016.

[24] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian Geometry Applied to BCI Classification," 2010, pp. 629–636. doi: 10.1007/978-3-642-15995-4_78.

[25] C. Cooney, A. Korik, F. Raffaella, R. Folli, and D. Coyle, "Classification of imagined spoken word-pairs using convolutional neural networks," vol. 2019, pp. 338–343, 2019, doi: 10.3217/978-3-85125-682-6-62.

[26] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open Access database of EEG signals recorded during imagined speech." [Online]. Available: www.phys.unsw.edu.au/jw/hearing.html

[27] K. Koizumi, K. Ueda, and M. Nakao, "Development of a Cognitive Brain-Machine Interface Based on a Visual Imagery Method," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, Jul. 2018, pp. 1062–1065. doi: 10.1109/EMBC.2018.8512520.

[28] M. R. Asghari Bejestani, G. R. Mohammad Khani, V. R. Nafisi, and F. Darakeh, "EEG-Based Multiword Imagined Speech Classification for Persian Words," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/8333084.

[29] T. Proix *et al.*, "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features," *Nat Commun*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-021-27725-3.

[30] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, *Towards an EEG-based Intuitive BCI Communication System Using Imagined Speech and Visual Imagery; Towards an EEG-based Intuitive BCI Communication System Using Imagined Speech and Visual Imagery*. 2019. doi: 10.0/Linux-x86_64.

[31] L. Wang, X. Zhang, and Y. Zhang, "Extending motor imagery by speech imagery for brain-computer interface," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2013, pp. 7056–7059. doi: 10.1109/EMBC.2013.6611183.

[32] N. Hashim, A. Ali, and W. N. Mohd-Isa, "Word-Based Classification of Imagined Speech Using EEG," in *Lecture Notes in Electrical Engineering*, Springer Verlag, 2018, pp. 195–204. doi: 10.1007/978-981-10-8276-4_19.

[33] N. Nieto, V. Peterson, H. L. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition," *Sci Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01147-2.

[34] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomed Eng Lett*, vol. 10, no. 2, pp. 217–226, May 2020, doi: 10.1007/s13534-020-00152-x.

[35] S. Zhao and F. Rudzicz, "CLASSIFYING PHONOLOGICAL CATEGORIES IN IMAGINED AND ARTICULATED SPEECH."

[36] Y. V. Varshney and A. Khan, "Imagined Speech Classification Using Six Phonetically Distributed Words," *Frontiers in Signal Processing*, vol. 2, Mar. 2022, doi: 10.3389/frsip.2022.760643.

[37] A. A. Torres-García, C. Alberto Reyes-Garcia, L. Villaseñor-Pineda, T.-G. Alejandro Antonio, R.-G. Carlos Alberto, and V.-P. Luis, "Toward a silent speech interface based on unspoken speech Domain adaptation for automatic deceptive text detection View project TOWARD A SILENT SPEECH INTERFACE BASED ON UNSPOKEN SPEECH," 2012. [Online]. Available: https://www.researchgate.net/publication/257984054

[38] M. J. Abdulaal, A. J. Casson, and P. Gaydecki, "Critical Analysis of Cross-Validation Methods and Their Impact on Neural Networks Performance Inflation in Electroencephalography Analysis," *IEEE Canadian Journal of Electrical and Computer Engineering*, vol. 44, no. 1, pp. 75–82, 2021, doi: 10.1109/ICJECE.2020.3024876.

[39] M. J. Abdulaal, A. J. Casson, and P. Gaydecki, "Performance of Nested vs. Non-Nested SVM Cross-Validation Methods in Visual BCI: Validation Study," in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, Sep. 2018, pp. 1680–1684. doi: 10.23919/EUSIPCO.2018.8553102.

[40] J. White and S. D. Power, "k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation," *Sensors*, vol. 23, no. 13, Jul. 2023, doi: 10.3390/s23136077.

[41] R. Li *et al.*, "Training on the test set? An analysis of Spampinato et al. [31]," Dec. 2018.

[42] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, M. Shah, and N. Souly, "Deep Learning Human Mind for Automated Visual Classification," Sep. 2016.

[43] MentaLab Explore+ Device. Mentalab 2023. Available online: https://mentalab.com/mobile-eeg/ (accessed on 9th October 2023).

[44] M. C. Corballis, "Left Brain, Right Brain: Facts and Fantasies," *PLoS Biol*, vol. 12, no. 1, p. e1001767, Jan. 2014, doi: 10.1371/journal.pbio.1001767.

[45] Broca P. ”Sur le siege de la faculte du langage articule”. Bull. Soc. Anthropol. Paris. 1865;6:377–393.”

[46] Wernicke C. ”Der aphasische Symptomencomplex. Eine psychologische Studies auf anatomischer Basis”. Breslau Max Cohn Weigert. 1874.

[47] S. K. Riès, N. F. Dronkers, and R. T. Knight, “Choosing words: left hemisphere, right hemisphere, or both? Perspective on the lateralization of word retrieval,” *Ann N Y Acad Sci*, vol. 1369, no. 1, pp. 111–131, Apr. 2016, doi: 10.1111/nyas.12993.

[48] M. Grimaldi, E. Brattico, Y. Shtyrov, and E. Neuromethods, “Language Electrii ed Principles, Methods, and Future Perspectives of Investigation.” [Online]. Available: http://www.springer.com/series/7657

[49] S. Geva, P. S. Jones, J. T. Crinion, C. J. Price, J. C. Baron, and E. A. Warburton, “The neural correlates of inner speech defined by voxel-based lesion-symptom mapping,” *Brain*, vol. 134, no. 10, pp. 3071–3082, 2011, doi: 10.1093/brain/awr232.

[50] J. B. Watson, “Psychology as the behaviorist views it.,” *Psychol Rev*, vol. 20, no. 2, pp. 158–177, Mar. 1913, doi: 10.1037/h0074428.

[51] L. Vygotsky, *Thought and language.* Cambridge: MIT Press, 1962. doi: 10.1037/11193-000.

[52] G. M. Oppenheim and G. S. Dell, “Inner speech slips exhibit lexical bias, but not the phonemic similarity effect,” *Cognition*, vol. 106, no. 1, pp. 528–537, Jan. 2008, doi: 10.1016/j.cognition.2007.02.006.

[53] S. Geva, P. S. Jones, J. T. Crinion, C. J. Price, J.-C. Baron, and E. A. Warburton, “The neural correlates of inner speech defined by voxel-based lesion-symptom mapping,” *Brain*, vol. 134, no. 10, pp. 3071–3082, Oct. 2011, doi: 10.1093/brain/awr232.

[54] L. A. Farwell and E. Donchin, “Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials,” *Electroencephalogr Clin Neurophysiol*, vol. 70, no. 6, pp. 510–523, Dec. 1988, doi: 10.1016/0013-4694(88)90149-6.

[55] X. Wang, H. Wanniarachchi, A. Wu, and H. Liu, “Combination of Group Singular Value Decomposition and eLORETA Identifies Human EEG Networks and Responses to Transcranial Photobiomodulation,” *Front Hum Neurosci*, vol. 16, May 2022, doi: 10.3389/fnhum.2022.853909.

[56] M. X. Cohen, “A better way to define and describe Morlet wavelets for time-frequency analysis,” *Neuroimage*, vol. 199, pp. 81–86, Oct. 2019, doi: 10.1016/j.neuroimage.2019.05.048.

[57] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier, “Oscillatory γ-Band (30–70 Hz) Activity Induced by a Visual Search Task in Humans,” *The Journal of Neuroscience*, vol. 17, no. 2, pp. 722–734, Jan. 1997, doi: 10.1523/JNEUROSCI.17-02-00722.1997.

[58] MNE - TF representation - Morlet Wavelets. Available online: https://mne.tools/stable/generated/mne.time_frequency.tfr_morlet.html (accessed on 9th October 2023).

[59] MNE - Decoding CSP time-frequency. Available online: https://mne.tools/stable/auto_examples/decoding/decoding_csp_timefreq.html#sphx-glr-auto-examples-decoding-decoding-csp-timefreq-py (accessed on 15th October 2023).

[60] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," Jan. 2012.

[61] T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian Framework for Least-Squares Support Vector Machine Classifiers, Gaussian Processes, and Kernel Fisher Discriminant Analysis," *Neural Comput*, vol. 14, no. 5, pp. 1115–1147, May 2002, doi: 10.1162/089976602753633411.

[62] Sklearn - SVC. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC (accessed on 15th October 2023).

[63] G. R. Müller-Putz, R. Scherer, G. Pfurtscheller, C. Brunner, and R. Leeb, "Better than Random? A closer look on BCI results Presenccia View project Mirage 91-The Graz-BCI Racing Team View project Better than random? A closer look on BCI results," 2008. [Online]. Available: https://www.researchgate.net/publication/275341064

[64] F. Stephan, H. Saalbach, and S. Rossi, "The brain differentially prepares inner and overt speech production: Electrophysiological and vascular evidence," *Brain Sci*, vol. 10, no. 3, Mar. 2020, doi: 10.3390/brainsci10030148.

[65] A. D. Manca and M. Grimaldi, "Vowels and Consonants in the Brain: Evidence from Magnetoencephalographic Studies on the N1m in Normal-Hearing Listeners.," *Front Psychol*, vol. 7, p. 1413, 2016, doi: 10.3389/fpsyg.2016.01413.

[66] A. Jahangiri and F. Sepulveda, "The Relative Contribution of High-Gamma Linguistic Processing Stages of Word Production, and Motor Imagery of Articulation in Class Separability of Covert Speech Tasks in EEG Data," *J Med Syst*, vol. 43, no. 2, Feb. 2019, doi: 10.1007/s10916-018-1137-9.

[67] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG Classification of Covert Speech for Brain-Computer Interfacing," *Int J Neural Syst*, vol. 27, no. 8, Dec. 2017, doi: 10.1142/S0129065717500332.

[68] J. Moon and T. Chau, "Online Ternary Classification of Covert Speech by Leveraging the Passive Perception of Speech," *Int J Neural Syst*, vol. 33, no. 9, Sep. 2023, doi: 10.1142/S012906572350048X.

[69] S. D. Power, A. Kushki, and T. Chau, "Intersession Consistency of Single-Trial Classification of the Prefrontal Response to Mental Arithmetic and the No-Control State by NIRS," *PLoS One*, vol. 7, no. 7, p. e37791, Jul. 2012, doi: 10.1371/journal.pone.0037791.

[70] G. Pfurtscheller, "The hybrid BCI," *Front Neurosci*, 2010, doi: 10.3389/fnpro.2010.00003.

Appendix A

Table A1: EEG-based IS BCI state-of-the-art

Ref.	Set-up	Protocol	Signal processing	Performance
[28]	Clinical quality equipement. 21 electrodes, 2 ECG channels, 1 EMG channel. Sampling rate: 500Hz	Silent talk of 6 Persian words (“Up”, “Down”, “Left”, “Right”, “Yes”, “No”) and silence class. Imagination is performed with closed eyes. The word (or the instruction for silence) is talked to subjects on earphones. As the audio finishes the subjects imagine saying the word for three times, within a 2.5s time window. 5 subjects. 5 sessions are performed by each subject along 5 weeks. In total 200 trials (composed of 3 repetitions) per word and 600 trials for silence class are recorded.	The complete 2.5s trial is used as time window. Trials are band-pass (0-32Hz) filtered. Trials with high signal energy are rejected, as probably are corrupted by artifact. For each channel, amplitude FFT is computed in the interval 1-32Hz with frequency bins of 1Hz, resulting in 32 amplitude values. These values are fed into an SVM.	Reported accuracy are obtained through a Monte-Carlo cross-validation procedure. Hyperparameter tuning process of the SVM is not clearly explained. Classification accuracy is also reported considering the samples for different classes coming from different sessions, or from different parts of the same session reaching detection accuracy (chance level: 50%) of 97 and 96%. But considering the samples for both classes coming from the same parts of the sessions, detection accuracy reduces to 58%. This shows how important it is to consider the non-stationarity of EEG signals in the time and even more among different sessions.
[33]	Clinical quality equipment. 128 EEG channels. Sample rate: 1024Hz. Also EMG and EOG channels recorded.	Inner speech of 4 Spanish words (“Up”, “Down”, “Right”, “Left”): each participant is indicated to imagine his/her own voice as if he/she is giving a direct order to the computer, repeating the corresponding word. Each trial starts with 0.5s of concentration interval showing a white fixation dot. Then a white triangle oriented in one of the 4 directions is shown superposed to the with dot for 0.5s. Once the triangle disappears, the subject has 2.5s to repeat imagining the	The signal is band-pass (0.5-100Hz) and notch filtered; re-referenced at external channels to minimize common mode voltage due to line noise and body drift. ICA is performed to identify artifacts.	No classification is performed. This work is aimed at introducing a new public dataset with already pre-processed signal.

		word instructed by the triangle. Then, the fixation dot turns blue (1s): the trial is over, subjects stop the imagination phase. For 1.5-2s (randomized time interval) nothing is shown and the subject can rest. 10 subjects. 60 trials per word (240 trials in total).		
[36]	Clinical quality signal. 64 channels. Sample rate: 2048Hz,	The subjects are asked to perform overt speech and then imagined speech of six English words ("could," "yard," "give," "him," "there," and "toe"). Each trial starts with the written instruction of the word. It is then followed by a fixation cross remaining on the screen for 1s. As it disappears the subjects pronounce overtly the word in a time window of 2s (blank screen). Then fixation cross appears again for 1s, and as it disappears the subjects perform imagined speech of the same word (blank screen for 2s). The subjects themselves did a self-assessment of performing correct/incorrect trials after performing the imagined speech task. Then a 1s pause follows and the next trial begins. 15 subjects. 50 trials per word (300 per subject)	CAR re-reference is applied. The signal is band-pass (0.01-250Hz) and notch filtered. ICA is applied to remove eye blinking artifacts. Only the first second of each trial was used as time window. A wavelet-based (db4) filter bank is applied to obtain 8 frequency bands. From the 6 frequency bands up to 64Hz, 3 features are extracted: RMS, SD and relative wavelet energy. Then PCA is applied to reduce feature number (features explaining 95% of variance) and the resulting combination are fed into a RF or a SVM classifier to compare their performances. SVM resulted the best. No hyperparameter tuning is performed, but fixed, reasonable values are used.	Reported classification performances are the average accuracy obtained in a 5-fold CV process on each subject. In the 6-class classification problem (chance level: 16.6%) the achieved accuracy (average along subjects) is 28.6%. Considering the pair-wise binary classification problem, as an average of all the possible pair combination (chance level: 50%) the accuracy obtained is 74%.
[34]	Clinical quality device. 64 electrodes. Sampling rate: 1000Hz.	Imagined speech of four words: "Left", "Right", "Up", "Down". On the screen a question about which direction is pointing the arrow is shown together with an arrow pointing in one of the 4 directions. The subjects perform	Signals are band pass (0.5-128Hz) and notch filtered; artifacts are managed through ICA-based ADJUST algorithm. Time window to be used as one epoch is the whole 10s mental repetition trial. Extracted features are energy and entropy, computed on the original EEG signals and on DB4 wavelet coefficients. The set of	It is reported the best average cross-validated accuracy for each algorithm, without performing a nested cross-validation. In the multi-calss classification problem (chance level: 25%), the average accuracy is 48% (best subject 50%).

		imagined speech of that word for 10s, by imagining to continuously repeat it. The cue to start the 10s of mental repetition is given through a rhythm completion strategy: two equidistant "beep" are played and the imagination period starts when the third one should be to complete the rhythm. 6 subjects; 50 trials per word (200 trials in total).	extracted features is then fed into various ML algorithms. The best performing results to be Gaussian-ELM. The tuning parameters were chosen using a 10-fold cross-validation procedure.	For binary classification problem, considering pair-wise classification accuracies (chance level 50%), an average 85% is achieved. Subset of electrodes were also tested and non-significantly different performance were achieved by using only the electrodes covering specific speech brain areas.
[18]	Same as [30]	Same as [30]. 22 subjects.	Same as [30] but the final classification is performed by using a SVM classifier instead of an RLDA classifier.	As in [30], reported performances are the average accuracy obtained from a 10-fold cross-validation procedure. Hyperparameter tuning process of the SVM is not clearly explained. In IS detection (binary classification performed 12 times, one per each word vs "Rest" condition, chance level: 50%) the reported average accuracy along all the 12 words is 80.7%. In multi-class classification problem, considering all the 13 classes (chance level: 7.7%), they achieve an accuracy of 40%. Different subsets of channels and different frequency ranges are also investigated. In particular, it is shown that non-significative difference can be noted in classification accuracies when using only specific speech related brain areas with respect to using all the 64 channels. They also show gamma band (30-120Hz) to be the most informative.
[25]	Using the public database by [26]	Using the public database by [26]	Only 0.5s of the 4s available are used. Different strategies are tried, in particular a Convolutional Neural Network (CNN) taking as input raw EEG data achieves best classification performances.	A nested-cross-validation procedure is used to tune hyperparameters of the architecture and then report performances in a robust way. The achieved

				accuracy in binary classification between each word pair (50% chance level) is 66%.
[30]	Clinical data quality. 64 channels. Sample rate: 1000Hz.	The subjects are instructed to perform inner speech by imagining the given word as if they were performing real speech, without moving any articulators nor making the sound. 12 English words are used and the 13th class is a "Rest" class where the same stimulus is shown as for when a words is instructed but the subjects are demanded not to do anything particular. One block is composed of one trial for each of the 12 words and for the "Rest" class in a random order. Each trial has 2s where the audio instruction about the word is played. Then a white fixation cross appears on the screen for 0.8-1.2s (randomized interval). As the cross disappears the subject should imagine saying once the word that was played. After 2s the white cross appears back for another attention period of 0.8-1.2s followed again by a 2s imagination time. 4 repetitions per each instruction are performed. Then a 3s relax time before the subsequent instruction happens. 7 subjects. 22 blocks in total, with instructions (88 repetitions) per word.	The signal is band-pass (0.5-40Hz) filtered and segmented into 2s epochs from the beginning of each trial. Binary classification of "Rest" epochs vs "Imagine" epochs (taking the first trial of each block to keep the dataset balanced) is performed using spatio-frequency features based on CSP: logarithmic variances of the first and last three CSP components are used as input features for the classifier. Classification is performed by shrinkage regularized linear discriminant analysis (RLDA). RLDA is a classification method adding a regularization term to a covariance matrix using optimal shrinkage parameter, which doesn't need any hyperparameter tuning.	Reported performances are the average accuracy obtained from a 10-fold cross-validation procedure which randomly divides all the trials of each subject into ten equally sized subsets and classifies ten times using the nine subsets as a training set and one subset as a test set. It is not performed a CV procedure where all the repetitions following the same instruction are kept in the same set. For imagined speech detection ("Rest" vs "Imagined Speech", 50% chance level), they report an average accuracy along the subjects of 80%. The same classification is also performed by using subsets of channels obtaining average accuracies of 80% and 77% when electrodes covering only the left speech brain areas and occipital areas are used. For multi class classification, by using all the 12 words (chance level: 8.3%) an accuracy of 14.7% is achieved.
[12]	Using KaraOne public database [35].	Using KaraOne public database [35]. Imagined speech of 7 phonemic/syllabic prompts and of 4 English words ("pat", "pot", "knew", "gnaw").	Only the first 3s from each imagined speech trial are used. To increase data to be used, each channel is considered as an independent data vector.	The reported accuracy is the best score obtained in the cross-validation process used to optimize DNN architecture. A nested cross-validation is not implemented. In the multi-class classification problem (chance level

	Clinical data quality. 64 channels. Sample rate: 1000Hz. Only 11 channels (covering speech brain areas) are used.	Each trial consists of 4 steps: 5s rest; then the prompt is shown written on the screen while an auditory utterance is played – this is followed by 2s when the subject moves the articulators into position to begin pronouncing the prompt; 5s imagined speech state, when the subject continuosly mentally repeats the word; a speaking state when the subject speaks the prompt aloud. 12 subjects. 12 trials per word.	Statistical features are extracted from each EEG epoch (RMS, variance, kurtosis, skewness, 3rd order momentum); the 3s EEG signals are decomposed into 7 levels by using db4 wavelets and the same statistical features are also extracted from the last approximation coefficient and from the last 3 detailed coefficients. These features are fed into a DNN with 2 hidden layers. The DNN architecture is optimized based on a 5-fold cross-validation: the performance of several possible architectures are compared and the best is chosen. Since each channel was processed by itself, during the cross-validation process, it was paid attention to keep all the signals coming from the same trial in the same split. This is important, since the presence of a couple of channels from the test trials in the training set can lead to high spurious accuracy due to data leakage.	9%) it is claimed an average accuracy along 8 subjects of 57%.
[32]	Emotiv Epoc device, 14 EEG channels with dry electrodes. Only 6 left sided channels are used. Sampling rate: 128Hz.	Imagined speech of "Yes" or "No". Each session is composed of 10 repetitions per each word. As the word is displayed the subjects imagine to say that word. It remains displayed for 2s. After each repetition, 2s pause is there. 10 consecutive "Yes" are followed by 10 consecutive "No". 4 subjects. 5 sessions, 50 trials per word.	2s time windows are used. They are band-pass (0.16-43Hz) and notch filtered. Mel Frequency Cepstral Coefficients (MFCC) are extracted and fed into a KNN classifier.	The average accuracy reached for this binary classification (chance level 50%) is 58%, reported as average of 5 train-test splits (using a ratio of 60% train, 40% test). A nested CV is not implemented to tune hyperparameters. The work claims a good performance in comparison with other works using dry electrodes, but it should be considered the absence of a randomized order.

[27]	Clinical quality experiment. 65 electrodes. Sampling rate: 1000 Hz.	Inner-speech task of six Japanese words: "Up", "Down", "Left", "Right", "Forward", "Backward". The word is shown on the screen for 4s, then it is substituted by a fixation point for a second. As it disappears the subjects imagine saying the word, repeating it for 4s. 16 subjects, 10 trials per word (40s recording per word).	Signals are band-pass (1-120Hz) and notch filtered. ICA is used to remove eye blinks and CCA to reduce EMG sources of noise. The time window to be used for creating epochs is the whole 4s repeating period. Power spectral density is computed (using Welch method) per each channel and the powers in 12 frequency bands (10Hz wide, non-overlapping) are the extracted. The 12x65 powers are then fed into a SVM to classify the epochs to the specific word.	Reported performance of the model are obtained using a 20-fold cross-validation scheme. Hyperparameter tuning process of the SVM is not clearly explained. The obtained accuracy for the multi-class classification problem (chance level: 16.6%) is very high: 83% using all electrodes available, 81% and 75% using only specific subsets of electrodes corresponding to pre-frontal cortex and frontal pole. The relevance of frontal electrodes and high gamma frequency bands in the decoder, make the authors speculate that "the high accuracies in the tasks were caused not only by EEG components but also by EMG artifacts".
[26]	Ag-AgCl cup electrodes. 6 EEG channels. Sampling rate: 1024Hz.	Imagined speech of 6 Spanish words ("up", "Down", "Left", "Right", "Forward", "Backward") and vowels. The word to be imagined is played as an audio and shown on the screen. Then the subjects imagine that word in correspondence of an audio "beep" reproduced 3 times within a window of 4s. 4s rest is then allowed. 15 subjects. 40 imagined speech trials per word.	The signal is band-pass (2-40Hz) filtered. The used time window is the whole 4s period with the three imagined repetitions. The EEG is processed with discrete wavelet transform and the relative wavelet energies are used as features to be the fed into a SVM or a random forest classifier.	The reported accuracy is the best score obtained in the cross-validation process used to optimize the hyperparameters of the two proposed classifiers. A nested CV is not implemented. Reported accuracy is 20% with a chance level of 16.6%. The focus of this work is to present a public dataset and show that it is possible to use it for multi-class classification problems. Indeed, an accuracy better than chance level is achieved.
[31]	15 electrodes covering speech brain areas (Broca's and Wenicke's areas).	Imagined speech, also intended as mental reading, of the Chinese words for "Left" and "One". The trials start with a preparation period of 1s. Then the Chinese character for the word to be imagined is shown for 1s. When it disappears (blank screen) the subjects	Signals are band-pass (0.1-100Hz) and notch filtered. For epochs belonging to the class with the imagination of a word, only the last 2s of each trial are used as time window. As "silence" epochs, 2s windows are segmented from the resting periods following the imagination.	Accuracy is reported as the average of scores obtained through a 10-fold cross-validation applied on each subject data. Hyperparameter tuning process of the SVM is not clearly explained. Averaging along all the subjects, the reported accuracy is 85% in imagined speech detection (binary

	Sampling rate: 250Hz. EOG recorded to remove ocular artifacts.	imagine repeating the word for 4s. Finally, a rest period of variable duration is present. 6 subjects; 75 trials per character (150 trials in total).	CSP filtering is applied and four statistical features are extracted. Also, synchronization phenomenon between brain areas is exploited by extracting cross-correlation (τ=0) values between pairs of channels. They are then fed into a SVM classifier.	classification of "Left" vs "Rest" and "One" vs "Rest" – chance level 50%). Instead, accuracy in classifying which word is imagined ("Left" vs "One" – chance level 50%) is lower: 66%.

Appendix B

In this appendix the ERP (Figures B1 and B2) and TF representations (Figures B3 and B4) of section 4.3.1 are complemented. The same figures comprehending all the electrodes are shown.

Then for the TF representation (Figures B3 and B4), also the epochs related to the "NONE" class are represented, to offer a fair comparison between the two different mental conditions and show that the ERD identified in "IMAGINE" epochs (with different timing for the two paradigms) is not present in the "NONE" epochs.

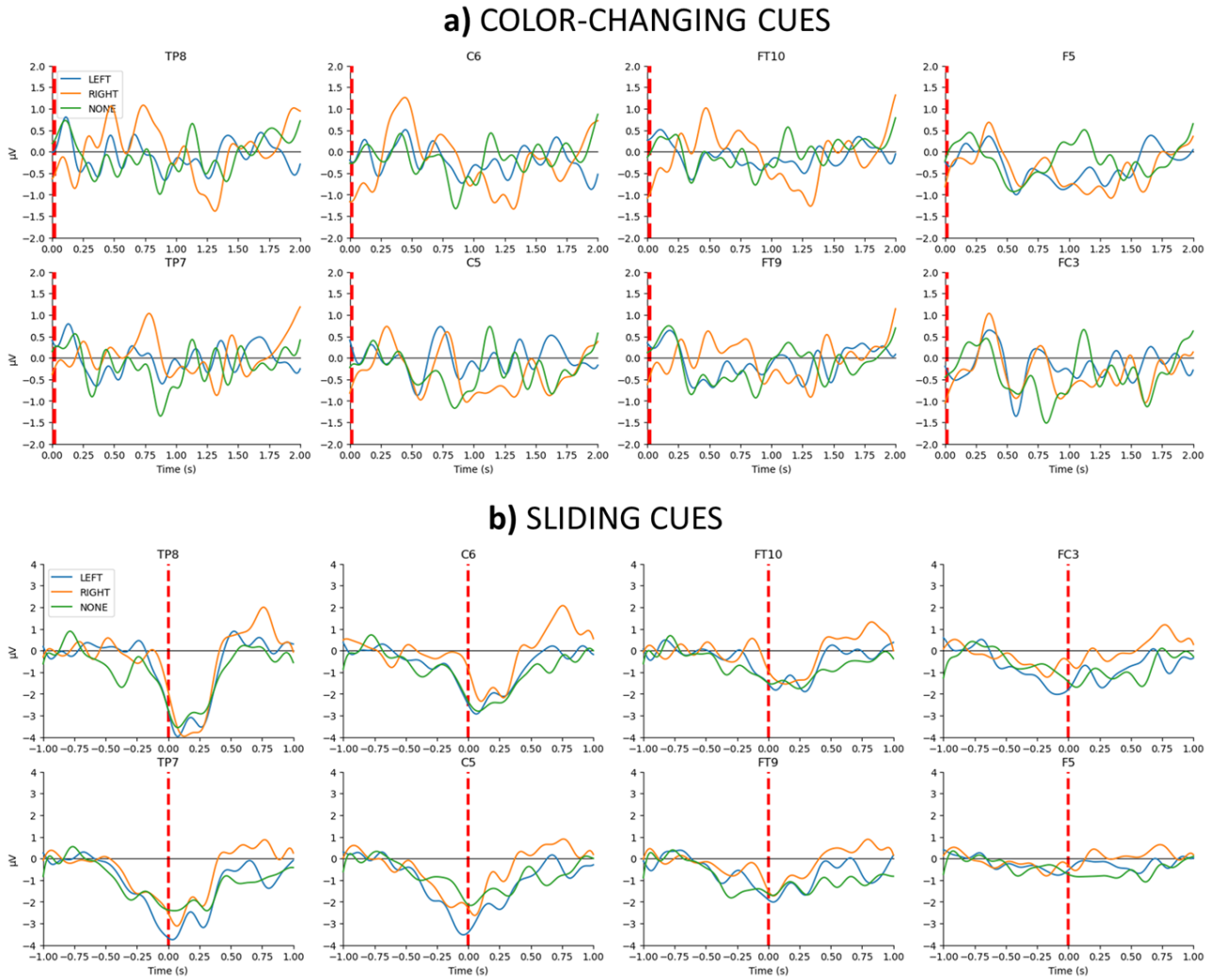

Figure B1: Time-locked average ERP. Complementing Figure 9 (all the channels are represented).

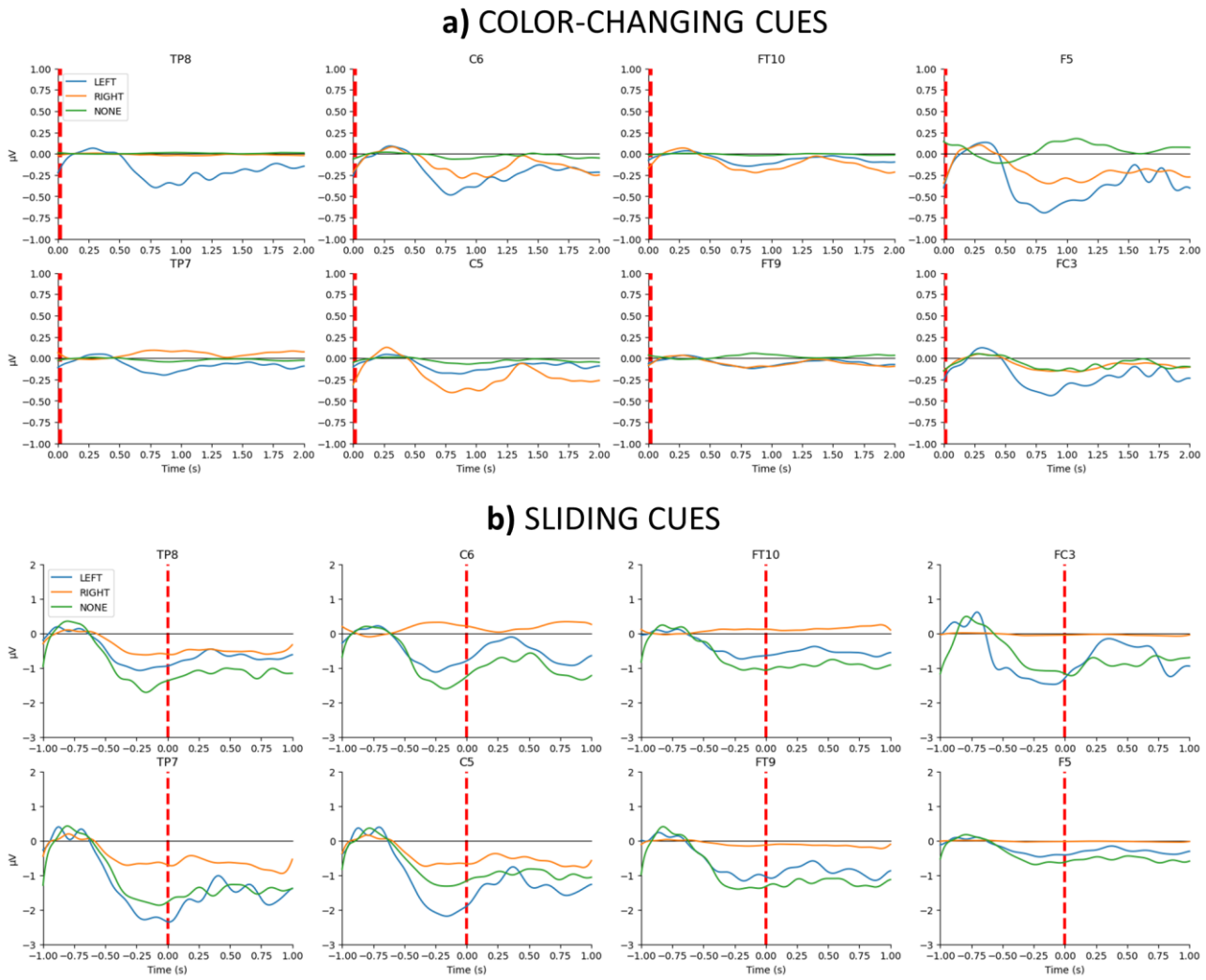

Figure B2: SVD first component ERP. Complementing Figure 10.

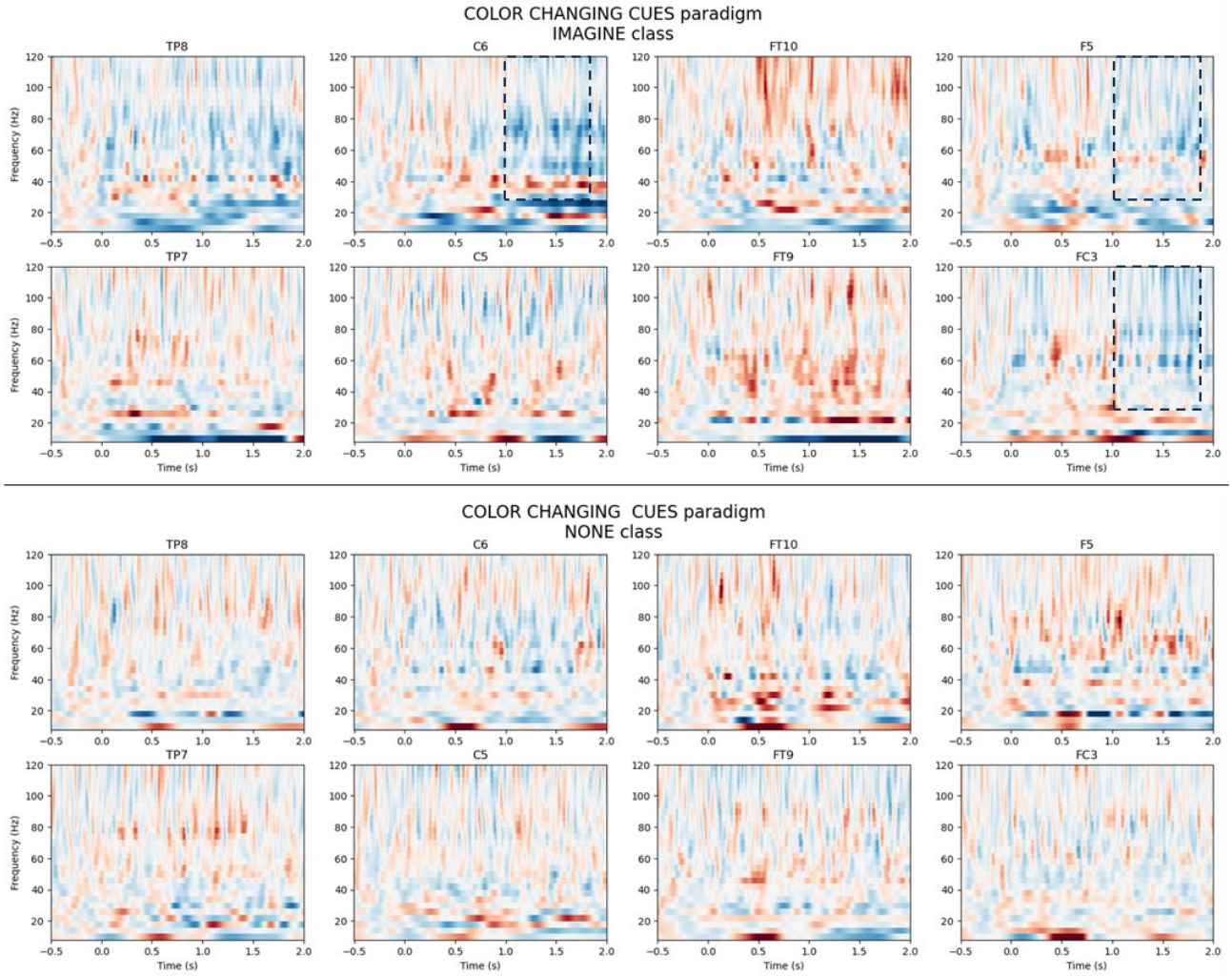

Figure B3: TF representation of *color-changing cues* paradigm per channel. Up the "IMAGINE" class is represented, down the "NONE" class. The power values (always positive) obtained by the Wavelet transform, are normalized with respect to a baseline (-0.5s-0s) period using a z-score scaler. Blue intensity encodes for negative values (hence power reduction with respect to the baseline average), red intensity for positive values (hence power increase). Dashed boxes encircle the identified ERD (reduction of the signal power) happening in γ band at time 1s-2s in channels F5, FC3 and C6.

Differently than Figure 11 where only "IMAGINE" classes are represented, here it is notable that ERD is only happening when the subject performs IS and not in "NONE" epochs. This demonstrates that the identified ERD is related to IS.

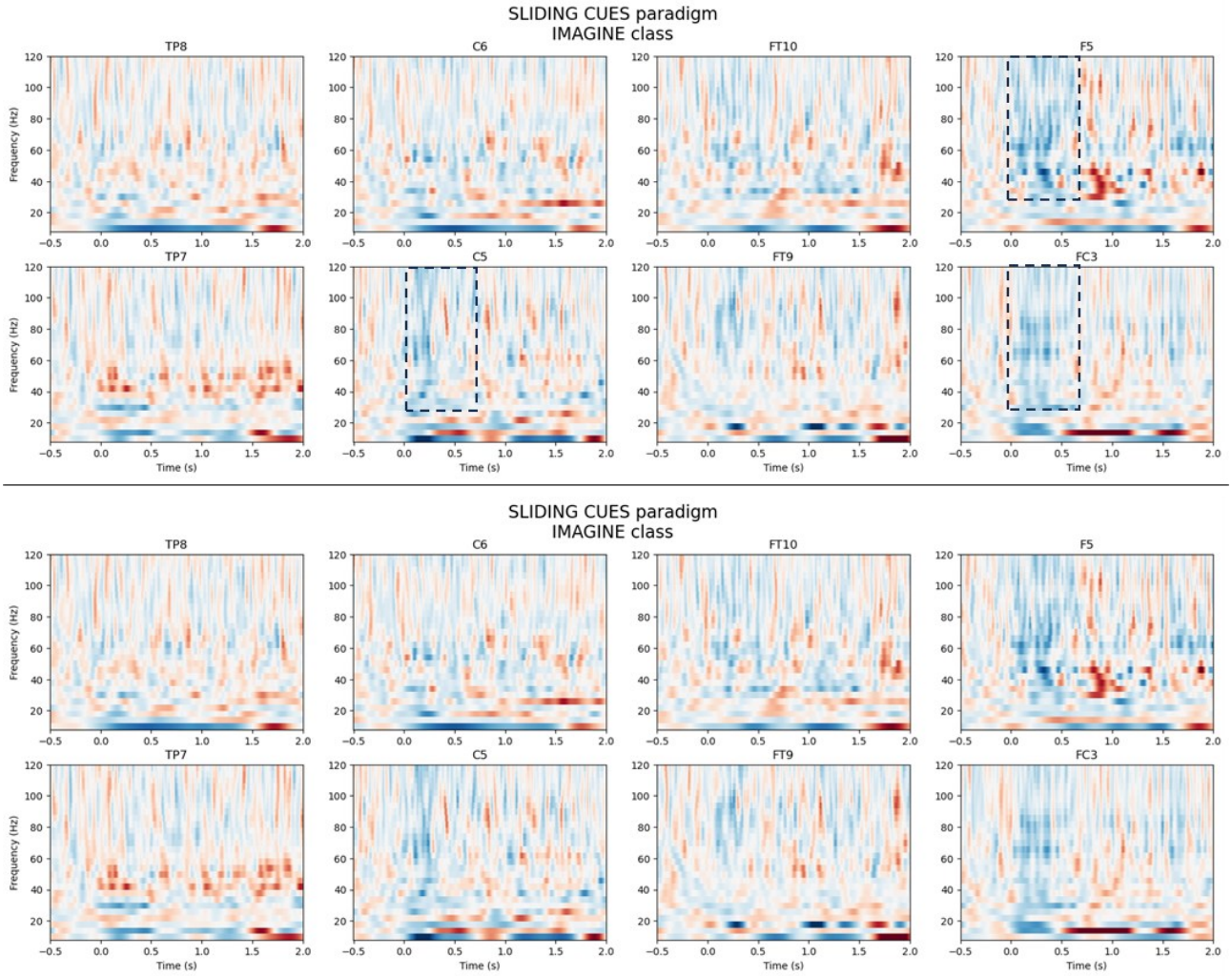

Figure B4: TF representation of *sliding cues* paradigm per channel. Up the "IMAGINE" class is represented, down the "NONE" class. The representation is analog to Figure B3. The ERD (visualized with the dashed boxes) is anticipated by 1s with respect to *color-changing* paradigm. Also here it is only present when the subject performs IS but is only found in channels F5 and FC3.

Differently than Figure 11 where only "IMAGINE" classes are represented, here it is notable that ERD is only happening when the subject performs IS and not in "NONE" epochs. This demonstrates that the identified ERD is related to IS.