Biocybernetics and Biomedical Engineering 45 (2025) 685–696

Contents lists available at ScienceDirect

Biocybernetics and Biomedical Engineering

journal homepage: www.elsevier.com/locate/bbe

Reevaluating performance in c-VEP BCIs: The impact of calibration time

Víctor Martínez-Cagigal [a,b,c,*], Eduardo Santamaría-Vázquez [a,b,c], Sergio Pérez-Velasco [a,b,c], Ana Martín-Fernández [a,b], Roberto Hornero [a,b,c]

[a] Biomedical Engineering Group, University of Valladolid, ETSIT, Paseo de Belen, 15, 47011, Valladolid, Spain
[b] Centro de Investigacion Biomedica en Red de Bioingenieria, Biomateriales y Nanomedicina (CIBER-BBN), Valladolid, Spain
[c] Valladolid Health Research Institute (IBioVALL), C. Rondilla Sta. Teresa, 47010, Valladolid, Spain

ARTICLE INFO

Keywords:
Calibration
Code-modulated visual evoked potentials (c-VEP)
Brain-computer interface (BCI)
Neurotechnology
Electroencephalography (EEG)

ABSTRACT

Code-modulated visual evoked potentials (c-VEP) have demonstrated high performance in non-invasive brain-computer interfaces (BCIs). Recently, research has begun to consider practical aspects such as visual comfort, where non-binary sequences and variations in the spatial frequency of stimuli play significant roles. However, calibration requirements remain underexplored in performance comparisons. This study aims to analyze a multivariable tradeoff crucial to the practical application of c-VEP-based BCIs: decoding accuracy, decoding speed, and calibration time. Visual comfort is retrospectively evaluated using two pre-recorded datasets. Models were trained with increasing calibration cycles and tested across varying decoding times, depicting learning and decoding curves. The datasets comprised 32 healthy subjects, and featured different stimulus paradigms: plain non-binary stimuli and checkerboard-like binary stimuli with spatial frequency variations. Results showed that all conditions achieved over 97 % grand-averaged accuracy with sufficient calibration. However, a clear tradeoff emerged between calibration duration and performance. Achieving 95 % average accuracy within a 2 s decoding window required mean calibration durations of 28.7±19.0 s for binary stimuli, or 148.7±72.3 s for non-binary stimuli. The binary checkerboard-based condition with a spatial frequency of 1.2 c/º (C016) proved to be particularly effective, achieving over 95 % accuracy within 2 s decoding window using only 7.3 s of calibration, and reporting a significant improvement in visual comfort. A minimum calibration time of 1 min was considered essential to adequately estimate the brain response, critical in template-matching paradigms. In conclusion, achieving optimal c-VEP performance requires balancing calibration duration, decoding speed and accuracy, and visual comfort.

1. Introduction

Brain-computer interfaces (BCIs) are defined as technologies that enable users to control applications or devices through their brain signals, typically monitored through electroencephalography (EEG) [1]. Over the past decades, BCIs have advanced considerably, emerging as a technology with potential applications in areas such as neurorehabilitation [2], alternative communication [3], and even as a novel control mechanism in gaming [4]. However, despite extensive research efforts, most of these applications remain confined to laboratories due to performance-related challenges, including intersubject variability, limited accuracy or reliability [1]. This constrained performance is intrinsically linked to the limitation that complex thoughts or actions cannot be directly decoded from EEG signals; instead, BCIs rely on detecting EEG changes produced by control signals, which are elicited either through mental tasks (endogenous signals) or responses to external stimuli (exogenous signals) [1].

Efforts to transition BCIs from an orphan technology to a mature tool have traditionally focused on enhancing decoding accuracy, decoding speed, or composite metrics such as the information transfer rate (ITR). Regarding the latter, it is important to distinguish between the theoretical ITR (tITR), which excludes inter-trial pauses for cueing or rest from its computation, and the practical ITR (pITR), which incorporates these pauses to reflect realistic operational conditions of the system. For instance, traditional BCIs based on event-related potentials (ERP) have repeatedly achieved decoding accuracy of over 90 % with pITRs of 10–25 bpm [5]. However, these systems are gradually being surpassed by more reliable control signals based on visual evoked potentials, such as steady-state visual evoked potentials (SSVEPs) or code-modulated visual evoked potentials (c-VEPs), which usually achieve over 90 % with

* Corresponding author.
E-mail addresses: victor.martinez.cagigal@uva.es (V. Martínez-Cagigal), eduardo.santamaria.vazquez@uva.es (E. Santamaría-Vázquez), sergio.perezv@uva.es (S. Pérez-Velasco), ana.martin.fernandez23@uva.es (A. Martín-Fernández), roberto.hornero@uva.es (R. Hornero).

https://doi.org/10.1016/j.bbe.2025.10.006
Received 10 February 2025; Received in revised form 16 October 2025; Accepted 31 October 2025
Available online 13 November 2025

significantly higher pITRs, usually over 30 bpm [5]. SSVEP-based BCIs display commands that flicker at specific frequencies, achieving real-time gaze decoding by detecting oscillatory responses in the EEG that align with the frequency of the command the user is looking at [1]. In contrast, c-VEP-BCIs based on circular shifting use commands encoded with shifted versions of a pseudorandom flickering code, enabling real-time gaze decoding by detecting the difference in phase in comparison with a previously calibrated template [5]. Although performance is generally comparable between these two control signals, c-VEPs are less prone to interference from narrowband baseline EEG activity (e.g., alpha peak) and less constrained by monitor refresh rates [5,6]. Recent research by Shi *et al.* (2024) [6] even suggests that, from an information-theoretic perspective, the maximum tITR achievable through the visual-evoked pathway with c-VEP-based BCIs significantly surpasses that of SSVEP-based systems.

This scientific focus on improving decoding performance has led to applications achieving remarkably high accuracy and speed. Recently, the hybrid approach proposed by Han *et al.* (2023) [7], which combined SSVEP, motion VEPs, and P300 potentials, reported an accuracy of 85.37 % with a pITR of 302.83 bpm in a 216-target speller. Within the c-VEP field, Sun *et al.* (2023) [8] achieved 95.02 % accuracy with a pITR of 183.09 bpm in a 120-target speller. Similarly, Moreno-Calderón *et al.* (2023) [4] reported a mean accuracy of 93.74 % with a tITR of 30.61 bpm in a c-VEP-based multiplayer competitive video game, while Gembler *et al.* (2020) [9] attained 94 % accuracy with a tITR of 88.9 bpm in a dictionary-assisted 32-target speller. Furthermore, the robustness of c-VEP-based systems has been validated in motor-disabled populations as well. Verbaarschot *et al.* [3] achieved a mean accuracy of 79.3 % with a pITR of 20.3 bpm in a 29-class speller tested on patients with amyotrophic lateral sclerosis.

After achieving reliable accuracy and speed metrics, research efforts transitioned toward enhancing visual comfort, as the flickering of high-contrast stimuli based on binary codes has been reported as uncomfortable for some users [5]. Scientific literature consistently indicates that higher stimulation rates are less fatiguing [10,11], and codes that concentrate most of their power in high-frequency bands further improve visual comfort [12–14]. Another study has shown that modulated codes were associated with a significant decrease in user comfort [15]. Recent studies have also demonstrated that reducing the maximal amplitude depth by lowering the contrast between stimuli, or even using non-binary codes [6,16], can enhance visual comfort without significantly compromising decoding accuracy. Moreover, the design of the stimuli has been shown to play a critical role in user convenience. For instance, Fernández-Rodríguez *et al.* (2023) [17] highlighted that higher spatial frequency variations in checkerboard-like stimuli increase visual comfort, while more recent findings by Dehais *et al.* (2024) [18] revealed that Gabor and Ricker-based textures are not only comfortable but also nearly imperceptible in peripheral vision compared to traditional flickering stimuli when implemented within a nonoverlapping burst-code paradigm.

Despite these advancements, the calibration requirements for c-VEP BCIs have been underexplored [5]. The scientific literature has predominantly focused on the trade-off between decoding accuracy and speed, whereas analyses of calibration duration are still scarce in the field. In practice, achieving high performance frequently requires multiple calibration runs customized for individual users, with recalibration required for each new session, even for the same user [19]. Therefore, it is appropriate to consider that BCI performance is determined by a multifactor tradeoff involving (1) decoding accuracy, (2) decoding speed (i.e., the time required to select a command), (3) calibration time, or even (4) visual comfort, among other potential factors.

Several studies have sought to reduce or eliminate calibration in c-VEP-based BCIs. For example, Miao *et al.* (2024) [20] applied a cross-subject transfer learning approach, successfully reducing calibration time to as little as 1 min while maintaining high decoding performance (99 % accuracy at a pITR of 250 bpm). Zheng *et al.* (2024) [21] developed a calibration-free system using narrow-band random sequences (NBRS), achieving 87.48 % accuracy with a pITR of 57.48 bpm. Stawicki *et al.* (2022) [22] explored the use of multi-session data for calibration, demonstrating that 336 s of data collected over multiple days were necessary to reach 90 % accuracy. Their findings also highlighted the potential of session-to-session transfer learning to significantly reduce calibration time. Similarly, Thielen *et al.* (2021) [19] proposed a stepwise reduction of calibration data, eventually eliminating the need for calibration by incorporating their "reconvolution" encoding model into a semi-supervised calibration algorithm. This system achieved a grand-averaged online accuracy of 99.76 % with a pITR of 66.43 bpm [19]. The authors further validated this zero-training calibration strategy using an unsupervised mean-difference maximization (UMM) method, attaining accuracy levels exceeding 94 % with 10.5 s trials [23]. Additional unsupervised calibration algorithms have also been proposed by Spüler *et al.* (2013) [24] and Turi *et al.* (2020) [25], though these approaches rely on specific c-VEP paradigms. For instance, the former designed a system calibrated with only two targets [24], while the latter required prior knowledge of an n-gram dictionary within a speller framework [25]. As shown, these studies provide valuable insights into strategies for reducing or even eliminating calibration through the use of unsupervised algorithms, thereby enhancing session performance and even enabling the visualization of learning and decoding curves across various approaches [19]. However, all of these studies used a binary-encoded, plain-stimuli c-VEP paradigm, which prevents the evaluation of calibration requirements in alternative paradigms specifically designed to enhance visual comfort. Furthermore, the primary objectives of these studies focused in proposing signal processing algorithms that minimized the amount of required training data, thereby reducing calibration time through advances in decoding. To date, the interdependencies among decoding accuracy, decoding speed, and calibration duration within c-VEP paradigms optimized for visual comfort remain insufficiently explored.

To address this research question, the present study aims to determine the calibration duration necessary to effectively control a c-VEP-based BCI and to investigate the interplay between accuracy, decoding speed, and calibration time. To this end, we assessed the impact of increasing calibration durations on performance using two datasets recorded under multiple conditions specifically designed to enhance user experience. Furthermore, we retrospectively analyze visual comfort based on these pre-recorded datasets, in order to offer additional insights into the usability of the BCI systems. The subjective nature of the visual comfort rating makes it impossible to simulate in an offline analysis; therefore, it can only be considered in relation to the ratings collected during the original experiments. The datasets involved the use of non-binary codes [16] and variations in spatial frequencies of checkerboard patterns [17], both of which are made publicly available for the first time as part of this study.

2. Subjects

To enhance the generalizability of the findings across diverse codes and stimulus conditions, the following datasets were used in this study: (1) non-binary m-sequences encoded with shades of gray [16], and (2) checkerboard patterns with varying spatial frequencies [17]. All BCI-related operation was managed using open-source MEDUSA″ v2023 (https://www.medusabci.com) applications [26]. In both of them, the visual stimuli were displayed on a Full HD LED monitor, with a fixed refresh rate at 120 Hz (XGM24F23.8″, Keep Out Gaming, Turkey). EEG signals were captured using a g.USBamp device from 16 g.LADYbird active Ag/AgCl electrodes (g.Tec, Austria) positioned at F3, Fz, F4, C3, Cz, C4, CPz, P3, Pz, P4, PO7, POz, PO8, Oz, I1, and I2, grounded at AFz and referenced to the earlobe, according to the International System 10-10. The selected 16-electrode montage was designed to cover

the parieto-occipital region, with an emphasis on maintaining spatial symmetry. All participants provided informed consent prior to participating in both studies. To promote reproducibility, both datasets have been made publicly available.

2.1. Non-binary m-sequences database

In this dataset, a total of 16 healthy participants (aged 28.80 ± 5.02 years, 11 males, 5 females) engaged in BCI spelling tasks using the open-source "*P*-ary c-VEP Speller" application of MEDUSA″ [26], accesible at medusabci.com/market/pary_cvep. The c-VEP stimuli were encoded using non-binary m-sequences, i.e. pseudorandom codes with more than two distinct events (gray levels). Five p-ary m-sequences were evaluated: binary GF(2^6) with a base of 2, GF(3^5) with a base of 3, GF(5^3) with a base of 5, GF(7^2) with a base of 7, and GF(11^2) with a base of 11. Note that GF(p^r) denotes the Galois Field containing p elements (i.e., distinct levels), where r represents the order of the primitive polynomial used to generate the m-sequence [16]. This procedure will be elaborated upon in Section 3.1. The elements were encoded by varying shades of gray; e.g. GF(2^6) was represented with black and white flashes, while GF(5^3) incorporated three equidistant gray tones in addition to black and white flashes, and so forth [16].

Each participant completed a single session that included a calibration phase consisting of 300 cycles (repetitions of each p-ary m-sequence), followed by an online spelling task of 32 trials (with 10 cycles per trial) for each condition. Fig. 1 illustrates all p-ary m-sequences, as well as the command layout. Online selections were made using a 4×4 command matrix (chance level of 6.25 %), consisting of alphabetic characters from A to P. For further details, refer to Martínez-Cagigal *et al.* (2023) [16]. The database containing all raw EEG data is publicly available at doi.org/10.35376/10324/70945 [27].

2.2. Checkerboard database

This dataset comprises data from 16 healthy participants (aged 29.63 ± 4.06 years; 11 males, 5 females) who undertook BCI spelling tasks using a branch of the open-source "c-VEP Speller" application of MEDUSA″ [26], available at medusabci.com/market/cvep_speller [17]. In this study, c-VEP stimuli followed a binary GF(2^6) m-sequence of 63 bits. It is noteworthy that this binary m-sequence is identical to the one used in the binary condition of the previously mentioned non-binary m-sequences database. The encoded events were represented through black-background checkerboard (BB-CB) patterns, with "1" events depicted by a checkerboard pattern and "0" events by a black flash. The conditions differed in the spatial frequency of the stimuli, i.e. the amount of squares within the checkerboard pattern, measured in cycles (pairs of squares of two alternative colors) per degree of visual angle (c/º). Participants were seated at a viewing distance of 60 cm from the screen. Eight spatial frequency conditions were tested: C001 (0 c/º), C002 (0.15 c/º), C004 (0.3 c/º), C008 (0.6 c/º), C016 (1.2 c/º), C032 (2.4 c/º), C064 (4.79 c/º), and C128 (9.58 c/º) [17]. Each condition label corresponds to the quantity of individual black-and-white squares (e.g., C008 represents an 8 × 8 square matrix).

Each participant completed a single session consisting of a calibration phase with 240 cycles, followed by an online spelling task comprising 18 trials (with 8 cycles per trial) for each spatial frequency condition. The stimuli presentation and command layout is depicted in Fig. 1. Online selections were made using a 3×3 command matrix (chance level of 11.11 %). Further details can be found in Fernández-Rodríguez *et al.* (2023) [17]. The database containing all raw EEG data is publicly available at doi.org/10.35376/10324/70973 [28].

3. Methods

3.1. The circular shifting paradigm

In both datasets, we implemented the circular shifting paradigm, which encodes application commands using time-delayed versions of a pseudorandom code, typically chosen for its low autocorrelation properties [5]. Calibration involves extracting the brain response elicited by the original code, mainly over the primary visual cortex, to generate a main template. Templates for the remaining commands are generated by shifting the main template according to the lag of each command. During online decoding, the response to a given stimulus is compared with these shifted templates, selecting the command associated with the highest correlation [5]. The lag arrangements used for each database are depicted in Fig. 1 [16,17].

As outlined briefly in Sections 2.1 and 2.2, we used maximal length sequences (m-sequences) to encode application commands. These m-sequences are pseudorandom temporal series generated by linear feedback shift registers (LFSRs) and are characterized by near-optimal autocorrelation properties [5]. An m-sequence is defined by three key parameters: (1) the base p, representing the number of distinct levels or events (e.g., for binary m-sequences, $p = 2$); (2) the order r, which corresponds to the number of taps in the LFSR; and (3) the arrangement of these taps, expressed as a polynomial with coefficients constrained to a Galois Field (GF) with p elements. As previously noted, these m-sequences are denoted using the notation GF(p^r) [16]. In addition to satisfying various mathematical constraints, the length of an m-sequence is precisely $N = p^r - 1$ bits, repeating cyclically [5,16]. Additional details regarding their implementation for command encoding are provided in Martínez-Cagigal *et al.* (2023) [16]. Notably, while increasing the length of an m-sequence expands the number of commands that can be encoded with it, it also extends the time required to display a complete cycle.

3.2. Signal processing

The signal processing applied in this study is adapted from the standard "reference processing pipeline" for c-VEPs, widely recognized as the most commonly used approach in circular shifting-based studies and is well-known for its high performance capabilities [5]. We believe that this approach provides a solid foundation, ensuring that the results are as generalizable and representative as possible across diverse c-VEP implementations. This processing pipeline is detailed below.

The EEG signals were pre-processed using a combination of 7-th order infinite impulse response (IIR) Butterworth filters. First, a 50 Hz notch filter was applied to eliminate power line interference. This was followed by a filter bank comprising three bandpass filters between 1–60 Hz, 12–60 Hz, and 30–60 Hz. This filter bank has been previously used in the literature to enhance the separation between spontaneous brain activity (e.g., alpha peaks) and responses induced by stimuli [16,29]. The upper cutoff frequency was set at 60 Hz, corresponding to the highest fundamental frequency elicited when a user fixates on a display operating at a 120 Hz refresh rate (equivalent to an encoding pattern of 101010 …), without accounting for harmonics or nonlinear interactions. Since the m-sequences used in both datasets do not only exhibit transitions between 0 and 1 (or vice versa) within a single bit, 60 Hz is considered as an upper limit in our analysis. Conversely, the lower cutoff of 1 Hz was chosen to consider delta and theta bands, which also contain information related to the stimulation (e.g., the repetitive presentation of these m-sequences generates low-frequency components and their harmonics) [16].

The calibration of this signal processing pipeline involves extracting three spatial filters (i.e. one for each filtered signal). As the calibration stage involves looking to a command encoded with repetitions of the original m-sequence, we obtain a pre-processed signal $\boldsymbol{X} \in \mathbb{R}^{N_f,k,N_s,N_c}$, where $N_f = 3$ is the number of filters in the filter bank, k is the number of cycles (repetitions) of the m-sequence that have been displayed, N_s is the number of samples, and N_c is the number of EEG channels. For each filter, the concatenated response, denoted as $\boldsymbol{A} \in \mathbb{R}^{kN_s,N_c}$, is calculated by concatenating all cycles. Simultaneously, the averaged response across cycles, $\hat{\boldsymbol{X}}_f \in \mathbb{R}^{N_s,N_c}$, is computed and subsequently repeated k times to produce $\boldsymbol{B} \in \mathbb{R}^{kN_s,N_c}$, ensuring its dimensions align with those of $\boldsymbol{A}$. Canonical correlation analysis (CCA) is applied to find a pair of

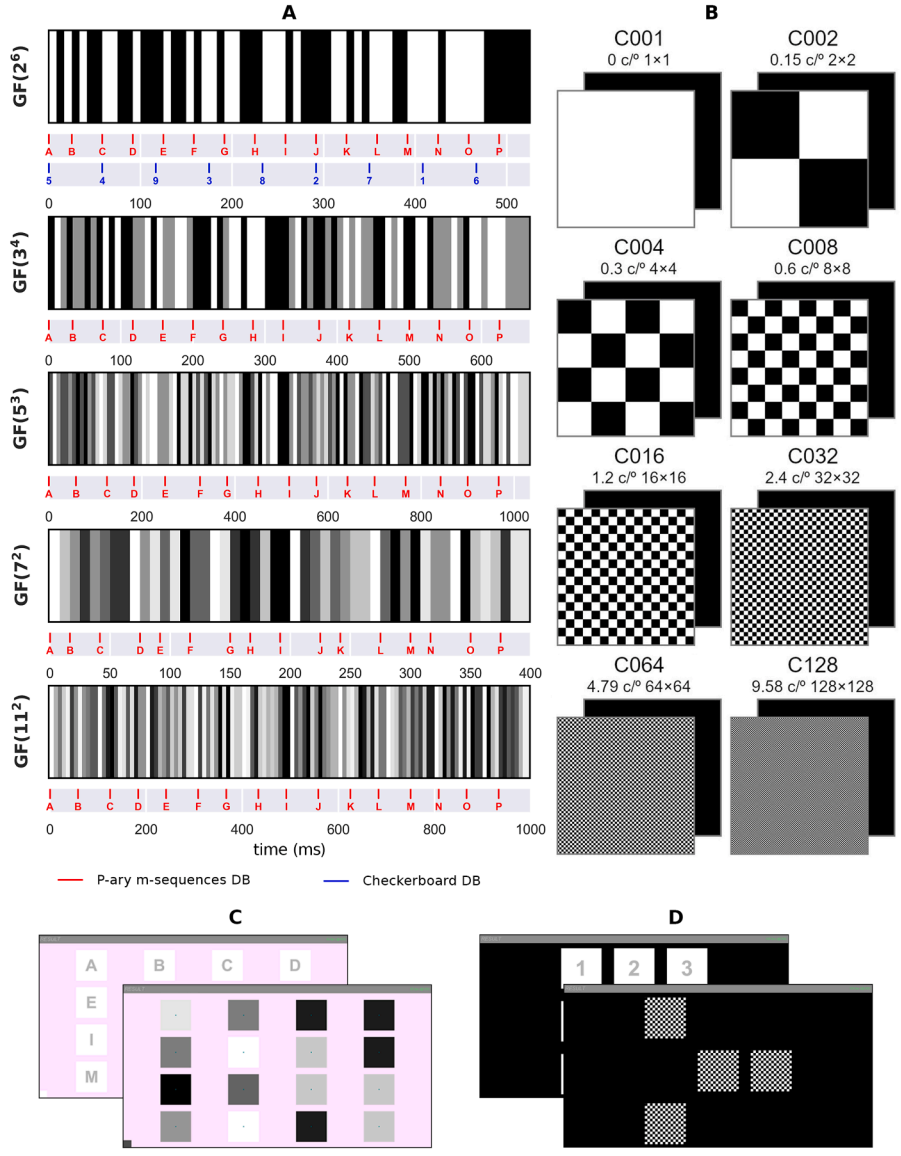


Fig. 1. Stimuli encoding in both databases. (A) Gray encoding of each m-sequence over time, depicting associated temporal shifts for each command in the non-binary m-sequence database (shown in red) and checkerboard database (shown in blue) with respect to the original m-sequence [16]. (B) Binary patterns of the black-background checkerboard (BB-CB) stimulus for the eight distinct spatial frequencies assessed in the checkerboard database. All patterns (event 1) were coupled with a black square (event 0). The temporal flickering of these stimuli was equivalent to the GF(2^6) m-sequence shown in (A) [17]. (C, D) Command layouts and flashing stimuli for the non-binary m-sequence database using condition GF(11^2) (C) [16], and checkerboard database using condition C016 (D) [17].

linear projections $\boldsymbol{w}_a \in \mathbb{R}^{N_c,1}$ and $\boldsymbol{w}_b \in \mathbb{R}^{N_c,1}$ that maximize the correlation between projected versions of these two signals, i.e. $\boldsymbol{A}\boldsymbol{w}_a$ and $\boldsymbol{B}\boldsymbol{w}_b$. This involves optimizing:

$$\arg\max_{\boldsymbol{w}_a,\boldsymbol{w}_b} \frac{\boldsymbol{w}_a^\top \boldsymbol{A}^\top \boldsymbol{B}\boldsymbol{w}_b}{\sqrt{(\boldsymbol{w}_a^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{w}_a)\cdot(\boldsymbol{w}_b^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{w}_b)}}. \quad (1)$$

Once the solution is obtained, only the first canonical component $\boldsymbol{w}_b$ is used afterward to project the averaged response (i.e., $\hat{\boldsymbol{X}}_f \boldsymbol{w}_b$) and get the main template $\tilde{\boldsymbol{x}}_{f0} \in \mathbb{R}^{N_s,1}$. Templates for the rest of the commands $\tilde{\boldsymbol{x}}_{fi}$ are calculated by circularly shifting $\tilde{\boldsymbol{x}}_{f0}$ according to the lag associated to each command θ_i. This procedure is repeated for each filter f, so we end up with $N_f \times m$ templates, where m is the number of commands (i.e. $m = 16$ for the non-binary database, and $m = 9$ for the checkerboard database) [16,17].

In the test stage, the goal is to identify the command to which the user was paying attention, which is unknown. To achieve this, the same pre-processing pipeline is applied to generate $\boldsymbol{Z}_{test} \in \mathbb{R}^{N_f,k_t,N_s,N_c}$, where k_t represents the number of cycles in the test trial. For each filter, the signal is cycle-averaged and projected using the trained spatial filter $\boldsymbol{w}_b$, resulting in the test response $\tilde{\boldsymbol{z}}_f \in \mathbb{R}^{N_s,1}$. This response is then compared with all template signals corresponding to that filter, yielding a vector $\boldsymbol{\rho}_f$ that contains the Pearson correlation coefficients for each command. The correlations are subsequently averaged across the filter bank, and the command corresponding to the highest coefficient is selected, i.e. $y = \arg\max_i \frac{1}{m}\sum_f \boldsymbol{\rho}_f$.

3.3. Evaluation of the calibration duration

To evaluate the impact of calibration duration on the system's overall decoding accuracy and speed, we adopted an incremental methodology. For a given number of calibration cycles k, the following steps were performed: (1) the model was trained with the training sets as described in Section 3.2, (2) the selected command was predicted for each trial for increasing decoding times t_d, and (3) performance metrics (accuracy and

Table 1
Cycle durations of all m-sequences, along with calibration and testing intervals.

			Calibration interval (s)			Testing interval (s)		
M-seq.	Length (bits)	Cycle dur. (s)	Min.	Max. (NB)	Max. (C)	Min.	Max. (NB)	Max. (C)
$GF(2^6)$	63	0.525	0.525	157.500	126.000	0.05	5.250	4.200
$GF(3^4)$	80	0.667	0.667	200.000	160.000	0.05	6.667	5.333
$GF(5^3)$	124	1.033	1.033	310.000	248.000	0.05	10.333	8.267
$GF(7^2)$	48	0.400	0.400	120.000	96.000	0.05	4.000	3.200
$GF(11^2)$	120	1.000	1.000	300.000	240.000	0.05	10.000	8.000

M-seq.: m-sequence, cycle dur.: cycle duration, C: checkerboard database [28], NB: non-binary m-sequences database [27].
The cycle duration was computed based on a display refresh rate of 120 Hz.
The calibration step was set equal to the cycle duration, whereas the testing step was consistently fixed at 50 ms.
Note that an m-sequence $GF(2^6)$ was applied across all conditions within the checkerboard database [17].

tITR) were computed. It is important to note that predictions are made using windows of progressively increasing length, simulating real-time continuous decoding rather than waiting for the completion of each stimulation cycle. This approach enables decision-making even when the full m-sequences have not yet been fully presented. These steps were applied for an increasing number of calibration cycles k. The chronological structure of the datasets is preserved consistently throughout both the calibration and decoding stages. Thus, the value of k ranged from 1 cycle up to the maximum number of cycles (300 for the non-binary m-sequences database and 240 for the checkerboard database) with an increment of 1 cycle. The value of t_d began at 50 ms and increased in 50 ms increments up to the maximum trial duration, corresponding to 10 cycles for the non-binary m-sequences database [27], and 8 cycles for the checkerboard database [28]. As the duration of a cycle is determined by the length of the m-sequence, the maximum durations for both the calibration and decoding windows are dependent on the specific experimental condition [16,17]. Detailed durations are provided in Table 1.

Accuracy is defined as the ratio of correct decisions to the total number of decisions. The ITR, expressed in bits per minute (bpm), was calculated using the formula originally proposed by Wolpaw *et al.* (2002) [30]:

$$\mathrm{ITR} = Q\left[\log_2(m) + P\log_2(P) + (1-P)\log_2\left(\frac{1-P}{m-1}\right)\right], \tag{2}$$

where Q denotes the number of selections per minute, m represents the number of commands, and P is the accuracy. From this point onward, we mainly report tITRs that exclude inter-trial pauses for cueing or rest in the computation of Q, as these intervals may vary in real-world applications. In some figures, the equivalent pITR is also shown, estimated by accounting for an inter-trial interval of 2 s (1 s for presenting the selected command and 1 s pause between trials), consistent with the configuration used during the recording of both datasets [16,17].

To determine when a pair of calibration cycles and decoding time (k, t_d) yields an accuracy statistically equivalent to the maximum achievable accuracy distribution in both datasets, we applied a two one-sided tests (TOST) procedure based on the Wilcoxon signed-rank test [31]. The maximum achievable accuracy distribution comprised the accuracy values for each subject at the minimal calibration value k yielding the highest grand-averaged accuracy within the condition. The equivalence margin was set to $\delta = 4\%$, corresponding to 5 % of the range between chance level and perfect accuracy ($\delta = \lfloor 5 \cdot (1 - \frac{1}{m}) \rfloor$). To account for multiple comparisons, p-values were adjusted using the Benjamini-Hochberg false discovery rate (FDR) procedure [31].

4. Results

An independent analysis using the same experimental setup was conducted to evaluate multiple latency measurements, the results of which are presented in Table 2. A detailed description of the methodology employed for this analysis is provided in A. Table 2 reports the transmission latency between black-to-white stimulus onsets, measured at (1) a phototransistor located at the top-left of the screen $t_{t,tl}$, (2) a phototransistor located at the bottom-right of the screen $t_{t,br}$, and (3) the estimated mid-point of the screen; as well as the actual raster latency t_r. The data indicate an average transmission latency of approx. 40.003 ms, and an actual raster latency of 5.305 ms. While average and standard deviation values are reported for $t_{t,tl}$ and $t_{t,br}$, it is crucial to note that their actual distributions are discrete, reflecting the chunk-based data transmission of the EEG equipment, i.e. $t_{t,tl}$ exhibited values of $\{35.156, 39.062\}$ ms, whereas $t_{t,br}$ exhibited values of $\{39.062, 42.969, 46.875\}$ ms. Notably, the difference between these discrete values is precisely 3.906 ms, corresponding to the duration of a single sample (1/256 Hz).

Table 2
Transmission latency measurement results.

	Mean	STD
Top-left, $t_{t,tl}$	37.351 ms	1.938 ms
Bottom-right, $t_{t,br}$	42.655 ms	1.301 ms
Estimated mid-point	40.003 ms	1.359 ms
Actual raster latency, t_r	5.305 ms	1.873 ms

Transmission times were computed as the difference between the flash onset timestamps and the peak of the phototransistors t_{tl} and t_{br}. Mid-point was estimated as $(t_{tl} + t_{br})/2 - t_o$. STD: standard deviation.

Grand-averaged performance results (accuracy, tITRs) as a function of (1) calibration cycles (and calibration time) and (2) decoding time for the non-binary m-sequences and the checkerboard datasets are shown in Figs. 2 and 3, respectively. The contours define regions in which accuracy distribution remains statistically equivalent to that of the distribution producing the maximum grand-averaged accuracy for each condition, while simultaneously minimizing the required calibration time. Detailed ranges of accuracy, tITR, and pITR associated with the respective contours are presented in the bottom-right table. Individual performance results for each participant and condition are shown in the supplementary material. The average estimated transmission latency was subtracted from the decoding time axis of all plots to enable fair comparison across different EEG equipment. For instance, a point labeled as 10 ms in the figures actually corresponds to EEG activity received at 50 ms post-stimulus onset, due to the approximate 40 ms transmission delay. Importantly, this correction does not imply that decoding was performed using a 10 ms EEG window, but rather that the time axis was shifted to account for the latency observed.

Fig. 4 presents the grand-averaged activation patterns across filter banks and subjects for both datasets, computed using the maximum number of available calibration cycles. The temporal evolution of these patterns throughout the calibration process, as well as the corresponding activation patterns for each participant is provided in the supplementary material. The activation patterns were computed following the method proposed by Haufe *et al.* (2014) [32], as given by:

$$\boldsymbol{a}_b \propto \Sigma_{\boldsymbol{B}} \boldsymbol{w}_b, \tag{3}$$

where $\Sigma_{\boldsymbol{B}} \in \mathbb{R}^{N_c, N_c}$ denotes the covariance matrix of the averaged response, and $\boldsymbol{a}_b \in \mathbb{R}^{N_c, 1}$ corresponds to the activation values for each EEG channel [32]. Given that the sign in the context of CCA is ambiguous (i.e., the correlation between $\boldsymbol{A}\boldsymbol{w}_a$ and $\boldsymbol{B}\boldsymbol{w}_b$ is equivalent to that between $-\boldsymbol{A}\boldsymbol{w}_a$ and $-\boldsymbol{B}\boldsymbol{w}_b$), we applied a sign-adjustment procedure to the weights of each spatial filter prior to averaging across the filter bank. That is, we enforced that the maximum absolute weight within each spatial filter was positive before computing the corresponding activation pattern.

Fig. 5 illustrates the grand-averaged c-VEPs recorded at Oz for both datasets. The onset timestamp was delayed by 40.003 ms to account for the average estimated transmission latency in the calculation of

the c-VEPs. The temporal evolution of these c-VEPs with increasing calibration duration is also depicted, along with the correlation between each c-VEP and the final c-VEP. This correlation serves as a visual metric to assess the calibration duration required to reach a steady state where no significant changes in the brain response are observed. To mitigate large amplitude outliers, a standard deviation-based artifact rejection method was applied. First, the standard deviation for each channel, $\sigma \in \mathbb{R}^{1,N_c}$, was computed across the entire calibration signal (i.e., considering all trials). An individual cycle was excluded if the standard deviation across all channels within that cycle exceeded 3σ [16].

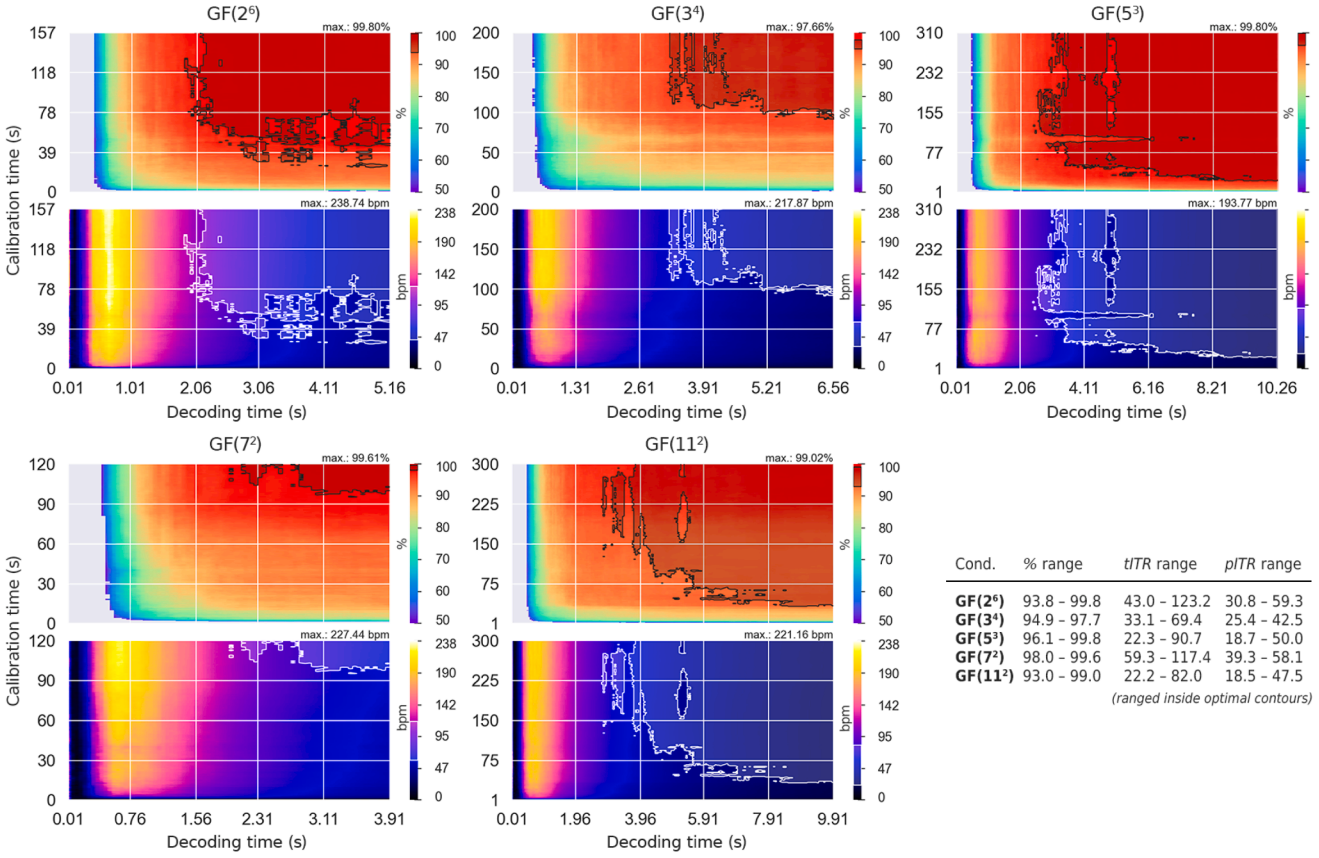
5. Discussion

5.1. Calibration and performance

The results clearly demonstrate that performance in terms of accuracy and decoding speed is influenced by the duration of the calibration phase. Specifically, Figs. 2 and 3 underscore the ability of all conditions (binary and non-binary m-sequences, plain, and checkerboard patterns) to achieve over 97 % accuracy with adequate calibration. However, these performance metrics undergo substantial changes when the decoding and calibration durations vary. Optimizing and adapting these variations to individual subjects may be crucial for signal processing algorithms (e.g., classifiers, early stopping, and asynchronous approaches) applied in practical BCI applications in communication and control.

The figures also show that accuracies statistically equivalent to the maximum grand-averaged accuracy for each condition can be achieved with relatively limited calibration. Notably, some conditions reach this plateau more quickly than others. For example, GF(5^3) attains optimal accuracy at $t_d \simeq 3$ s after 60 s of calibration, whereas GF(7^2) requires 100 s of calibration to reach the same performance at $t_d \simeq 3$ s. Results for the binary m-sequence with classical plain stimuli were consistent across both datasets; however, the spatial frequency of the binary stimuli also played a decisive role, with the best outcomes obtained for C016 and C032. Although the curves differ in shape, C016 achieved optimal accuracy with 60 s of calibration at $t_d \simeq 1$ s, while C032 reached the same performance with only 30 s of calibration but required a longer decoding time of $t_d \simeq 1.6$ s. Generally, it is also shown that the checkerboard dataset achieves higher performances compared to the non-binary m-sequences dataset. This is attributable to two primary factors: (1) decoding is performed with a smaller number of classes (9 commands versus 16); and (2) binary codes required less calibration to achieve comparable results.

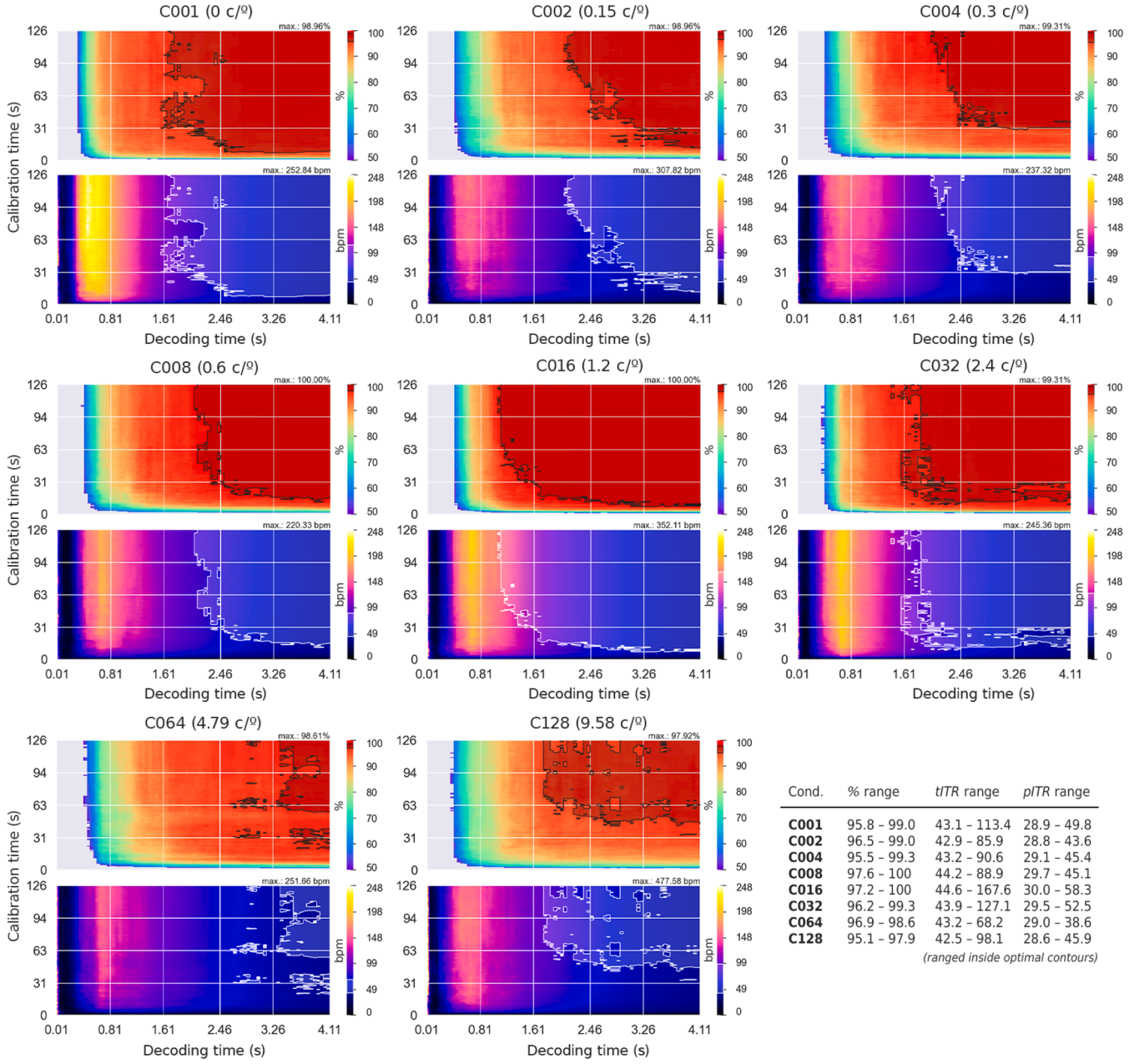
Interpreting results with a high number of variables can be challenging at first glance. To assist readers, Table 3 summarizes the calibration durations required to achieve 80 %, 90 %, and 95 % accuracy for decoding durations (t_d) of 1, 2, and 3 s. We selected these values heuristically to facilitate a fair and rapid comparison between conditions; however, readers are encouraged to focus on the statistical test results reported above for a rigorous interpretation. As shown, all conditions surpass 80 % accuracy with just 1 s of decoding. Notably, binary codes require calibration durations ranging from 3.7 to 20.5 s, with C001 standing out by achieving 80 % accuracy using just 3.7 s. Expectedly, the required calibration time decreases drastically for de-



Cond.	% range	*tITR* range	*pITR* range
GF(2⁶)	93.8 - 99.8	43.0 - 123.2	30.8 - 59.3
GF(3⁴)	94.9 - 97.7	33.1 - 69.4	25.4 - 42.5
GF(5³)	96.1 - 99.8	22.3 - 90.7	18.7 - 50.0
GF(7²)	98.0 - 99.6	59.3 - 117.4	39.3 - 58.1
GF(11²)	93.0 - 99.0	22.2 - 82.0	18.5 - 47.5

(ranged inside optimal contours)

Fig. 2. Grand-averaged accuracy (top plots) and tITR (bottom plots) as a function of the calibration time (y-axis), and decoding time (x-axis) for the non-binary m-sequences database. Each plot represents a different stimulation m-sequence: binary GF(2^6), GF(3^4), GF(5^3), GF(7^2), and GF(11^2). The average estimated transmission latency was corrected. Only accuracies exceeding 50 % are shown. The maximum accuracy and maximum tITR for each condition are reported in the upper-right corner of the respective plots. Black (accuracy) and white (tITR) filled contours overlaid on the plots and colorbars indicate regions where the accuracy was statistically equivalent to the maximum grand-averaged accuracy (p-values < 0.05, TOST equivalence testing with Wilcoxon signed-rank tests, FDR corrected). The bottom-right table presents the ranges of accuracy, tITR, and estimated pITR within the specified regions.



Cond.	% range	*tITR* range	*pITR* range
C001	95.8 – 99.0	43.1 – 113.4	28.9 – 49.8
C002	96.5 – 99.0	42.9 – 85.9	28.8 – 43.6
C004	95.5 – 99.3	43.2 – 90.6	29.1 – 45.4
C008	97.6 – 100	44.2 – 88.9	29.7 – 45.1
C016	97.2 – 100	44.6 – 167.6	30.0 – 58.3
C032	96.2 – 99.3	43.9 – 127.1	29.5 – 52.5
C064	96.9 – 98.6	43.2 – 68.2	29.0 – 38.6
C128	95.1 – 97.9	42.5 – 98.1	28.6 – 45.9

(ranged inside optimal contours)

Fig. 3. Grand-averaged accuracy (top plots) and tITR (bottom plots) as a function of the calibration time (y-axis), and decoding time (x-axis) for the checkerboard database. Each plot represents a different spatial frequency variation: C001 (0 c/º), C002 (0.15 c/º), C004 (0.3 c/º), C008 (0.6 c/º), C016 (1.2 c/º), C032 (2.4 c/º), C064 (4.79 c/º), C128 (9.58 c/º). The average estimated transmission latency was corrected. Only accuracies exceeding 50 % are shown. The maximum accuracy and maximum tITR for each condition are reported in the upper-right corner of the respective plots. Black (accuracy) and white (tITR) filled contours overlaid on the plots and colorbars indicate regions where the accuracy was statistically equivalent to the maximum grand-averaged accuracy (p-values < 0.05, TOST equivalence testing with Wilcoxon signed-rank tests, FDR corrected). The bottom-right table presents the ranges of accuracy, tITR, and estimated pITR within the specified regions.

coding times of 2 or 3 s, with C001 and C032 leading by requiring only 3.1 s of calibration. In contrast, the worst performing m-sequence was $GF(3^4)$, requiring calibration times of 82.7 s, 23.3 s, and 12.7 s for t_d values of 1, 2, and 3 s, respectively. Regarding a threshold of 90 % accuracy, only a subset of binary conditions reached this level with 1 s of decoding. Remarkably, C001 maintained 90 % accuracy with only 7.9 s of calibration, followed closely by C032 (12.1 s) and C016 (16.3 s). These three conditions also performed best at t_d values of 2 and 3 s, requiring calibration durations of just 4.2–5.2 s. Finally, it is noteworthy that only condition C016 achieved more than 95 % accuracy with 1 s of decoding, requiring a total calibration time of 101.8 s. For decoding times of 2 and 3 s, C016 and C032 emerged as the best performing conditions, achieving more than 95 % accuracy with calibration times between 5.8 and 7.3 s. In contrast, $GF(11^2)$ exhibited the poorest performance, requiring calibration times of 199–221 s to exceed 95 % accuracy.

Referring back to Figs. 2 and 3, it is important to note that increasing the cumulative number of calibration cycles (or calibration time) does not always lead to improved accuracy or speed. This could be because no artifact rejection was applied during the performance evaluation to ensure an equal number of cycles were used to train each condition. However, as observed (e.g., around a calibration time of 50 s for C064), the

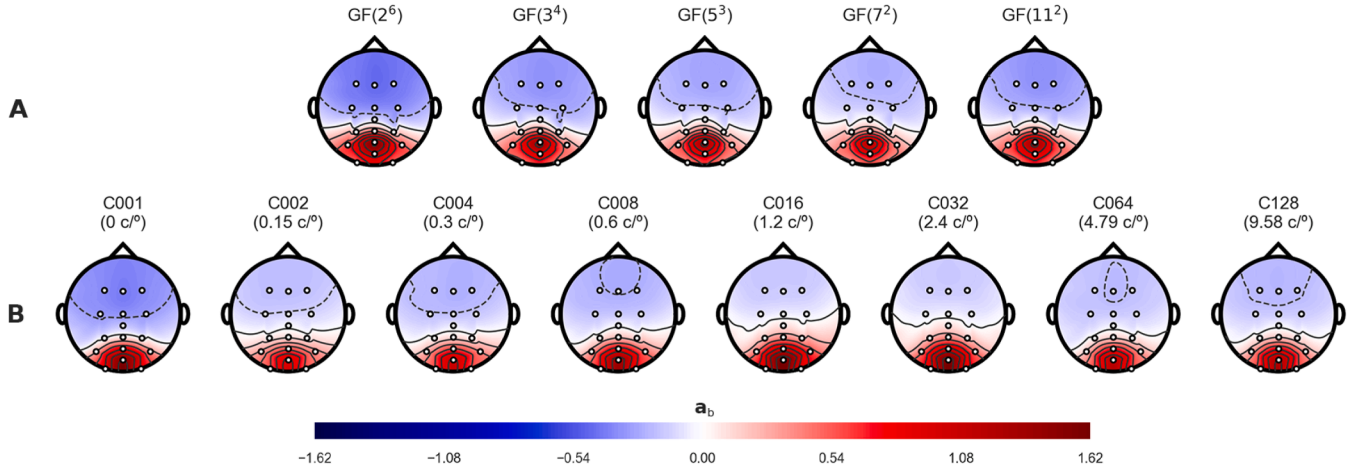


Fig. 4. Grand-averaged activation patterns across subjects and filter banks for (A) the non-binary m-sequences database and (B) the checkerboard database. The activation patterns have been calculated using the maximum number of calibration cycles available in each database.

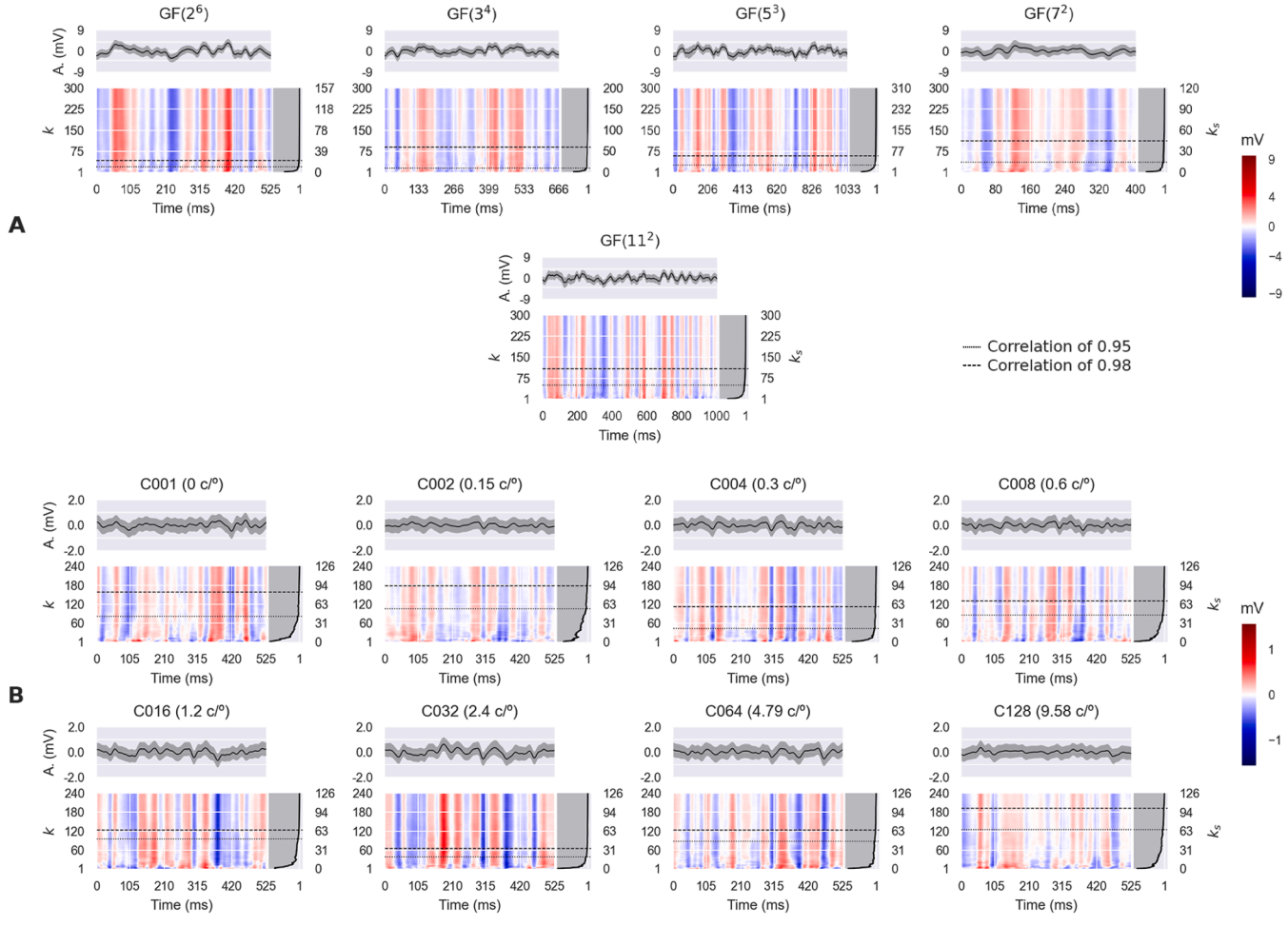


Fig. 5. Grand-averaged code-modulated visual evoked potentials (c-VEPs) at Oz across subjects for the non-binary m-sequences (A) and checkerboard (B) databases are shown as a function of the number of calibration cycles (k) or calibration duration (k_s) in seconds. All c-VEPs were adjusted to account for the average estimated transmission latency. The top panels present the grand-averaged c-VEPs computed using all available cycles, with shaded areas representing the standard deviation. A standard deviation-based artifact rejection method was applied to exclude large amplitude outliers. The right panels illustrate the correlation between each c-VEP and the final c-VEP, computed using all available calibration cycles. Thresholds where the correlation exceeds 0.95 and 0.98 are indicated by dotted and dashed black lines, respectively.

Table 3
Calibration time needed to achieve a specific average accuracy for decoding durations of 1, 2, and 3 seconds.

Condition \ t_d	Accuracy > 80 %			Accuracy > 90 %			Accuracy > 95 %			Comfort*
	1 s	2 s	3 s	1 s	2 s	3 s	1 s	2 s	3 s	
GF(2^6)	12.1 s	4.7 s	3.7 s	150.2 s	19.4 s	8.4 s	n.a.	41.5 s	34.6 s	6.7 ± 1.7
GF(3^4)	82.7 s	23.3 s	12.7 s	n.a.	84.0 s	53.3 s	n.a.	n.a.	160.7 s	7.2 ± 1.7
GF(5^3)	36.2 s	9.3 s	7.2 s	n.a.	22.7 s	14.5 s	n.a.	n.a.	23.8 s	7.4 ± 1.8‡
GF(7^2)	21.6 s	5.2 s	4.0 s	n.a.	28.4 s	17.6 s	n.a.	76.4 s	70.4 s	8.6 ± 2.1†,‡
GF(11^2)	40.0 s	12.0 s	8.0 s	n.a.	40.0 s	33.0 s	n.a.	221.0 s	199.0 s	8.6 ± 1.8†,‡
C001 (0 c/º)	3.7 s	3.1 s	3.1 s	7.9 s	5.2 s	4.2 s	n.a.	11.0 s	7.3 s	6.7 ± 2.6
C002 (0.15 c/º)	15.8 s	7.3 s	5.8 s	n.a.	13.7 s	8.9 s	n.a.	66.7 s	13.7 s	6.9 ± 1.8
C004 (0.3 c/º)	16.8 s	7.3 s	6.8 s	115.0 s	13.1 s	8.4 s	n.a.	33.6 s	15.2 s	7.4 ± 1.5
C008 (0.6 c/º)	12.1 s	5.8 s	5.8 s	70.9 s	10.0 s	8.4 s	n.a.	15.2 s	10.5 s	7.8 ± 1.5
C016 (1.2 c/º)	6.8 s	4.2 s	3.7 s	16.3 s	4.7 s	4.7 s	101.8 s	7.3 s	5.8 s	8.0 ± 1.3†
C032 (2.4 c/º)	5.2 s	3.1 s	3.1 s	12.1 s	5.2 s	4.2 s	n.a.	7.9 s	6.3 s	7.6 ± 2.0
C064 (4.79 c/º)	19.9 s	6.8 s	6.3 s	n.a.	10.5 s	8.4 s	n.a.	32.0 s	10.0 s	7.6 ± 1.9
C128 (9.58 c/º)	20.5 s	6.8 s	6.8 s	n.a.	17.9 s	13.7 s	n.a.	43.0 s	32.0 s	7.8 ± 1.9

All calibration times are reported in seconds; 'n.a.' indicates that the specified accuracy was not achieved for the given decoding time, and t_d represents the decoding time in seconds. * Grand-averaged visual comfort and standard deviation, derived from the values obtained through qualitative questionnaires from the original datasets. The results from both databases [16,17] were normalized against the common plain binary stimuli (i.e., GF(2^6) and C001) to ensure a fair comparison. Results for the non-binary m-sequences database were extracted from the online 120 Hz condition. Statistically significant improvements in visual comfort between each condition and the common plain binary stimuli are denoted by † at 60 Hz and by ‡ at 120 Hz. These results were derived from the original publications [16,17], in which Wilcoxon signed-rank tests were applied and the resulting p-values were subsequently corrected for false discovery rate using the Benjamini-Hochberg procedure.

inclusion of noisy cycles in the calibration stage could impede achieving the optimal performance plateau. This phenomenon is more evident in the individualized plots provided in the supplementary material. Simple artifact rejection algorithms, such as those based on standard deviation [16], could effectively mitigate this issue.

The ITR serves as a comprehensive metric that combines decoding accuracy, decoding speed, and the number of commands into a single value. While this unified measure facilitates the comparison of different BCI systems, Figs. 2 and 3 demonstrate that higher tITRs do not always correspond to acceptable levels of accuracy. Traditionally, the scientific literature has concurred that a minimum accuracy of 70 % is necessary for a BCI system to be "controllable" [33–35]. Although this threshold might be considered arbitrary, it underscores the necessity for practical BCI applications to meet or exceed this level of accuracy. In the present case, due to the rapid decoding pace of c-VEP, tITRs exceeding 200 bpm are observed even at accuracy levels below 70 % (and over 400 bpm for decoding times of 10 ms). This highlights the importance of interpreting ITR results cautiously, especially in SSVEP and c-VEP-based BCIs, as high ITR values can sometimes mask suboptimal accuracy, which is critical for practical usability. Nevertheless, optimal accuracy was achieved within more moderate tITR and pITR ranges. For binary m-sequence conditions, tITR values ranged from approximately 40 to 170 bpm and pITR values from 28 to 59 bpm, whereas non-binary conditions were slower, with tITR values ranging from 22 to 117 bpm and pITR values from 18 to 58 bpm.

5.2. Calibration and visual comfort

The previously discussed results have demonstrated a clear tradeoff between decoding accuracy, decoding speed, and calibration time. Additionally, the use of non-binary m-sequences and checkerboard databases allows for the consideration of an additional factor: visual comfort. Table 3 also presents the grand-averaged visual comfort values derived from qualitative questionnaires, where a score of 0 indicates no comfort and a score of 10 indicates very high visual comfort. Note that visual comfort was rated independently from the perceived duration of each trial, for every condition and participant. This separation was intended to isolate potential biases in comfort ratings that might arise due to variations in trial duration associated with different non-binary m-sequences. Indeed, a group-level t-test on the Spearman correlation coefficients between perceived comfort and perceived speed ratings revealed no significant effect ($p = .646$). The complete original questionnaires used in both studies are included in the corresponding datasets. These encompass not only the ratings for visual comfort, but also assessments of decoding speed and system usability scale (SUS) ratings [27,28]. For a detailed exposition of the methodological validity underlying these findings, readers are referred to the original publications [16,17]. The ratings reported in the original manuscripts describing both databases were mean-normalized against the common plain binary stimuli condition (i.e., GF(2^6) and C001) to facilitate a fair comparison [16,17]. However, it is important to interpret these inter-database comparisons with caution, as users only evaluated the conditions included in the specific study they participated in, rather than all the conditions presented here. It is also worth noting that, since the spatial frequency conditions were encoded using binary m-sequences, the interaction between spatial frequency and non-binary coding remains unexplored.

In Martínez-Cagigal *et al.* (2023) [16], we provided significant evidence that higher m-sequence bases are associated with reduced eyestrain perceived by users at both 60 Hz and 120 Hz presentation rates. Specifically, GF(7^2) and GF(11^2) were reported to be significantly less annoying at 120 Hz, while GF(5^3) also showed the same effect at 60 Hz. The effect of increasing perceived visual comfort as the base increases is clearly reflected in the results reported in Table 3, where $GF(7^2)$ and $GF(11^2)$ obtained a mean visual comfort value of 8.63, followed by $GF(5^3)$ with 7.38, $GF(3^4)$ with 7.19, and $GF(2^6)$ with 6.69. Although no linear relationship exists between the base of the m-sequence and the calibration time required to achieve high accuracy, Fig. 2 and Table 3 indicate that non-binary m-sequences generally require more calibration to reach comparable performance levels to binary codes. Notably, GF(7^2) and GF(5^3) achieved results that approach or even overcome those of GF(2^6), e.g. at $t_d = 3$s and 95 % accuracy, GF(5^3) requires only 23.8 s of calibration, whereas GF(2^6) needs 34.6 s. For example, at $t_d = 3$ s and 80 % accuracy, GF(2^6) requires 3.7 s of calibration, followed by GF(7^2) with 4.0 s and GF(5^3) with 7.2 s. However, the sta-

tistical analysis of Fig. 2 indicates that GF(5^3) and GF(11^2) are highly promising candidates, as they simultaneously achieve near-optimal accuracy and a substantial reduction in eyestrain [16] when the decoding time is slightly extended to $t_d \simeq 4$ s. Notably, the former requires only about one minute of calibration.

With regard to the checkerboard database, previous research has shown that increasing spatial frequency is associated with significant improvements in visual comfort [17]. C016 condition was recommended, as it was significantly more comfortable than the plain condition while achieving over 95 % accuracy after 2.1 s of decoding time. Notably, the model was trained using all available calibration cycles (240 cycles, 126 s). While our findings align with these results (95 % accuracy for $t_d = 1$ s and 101.8 s of calibration), condition C032 also achieved comparable performance with reduced calibration times. For example, at $t_d = 2$ s, C032 required only 3.1, 5.2, and 7.9 s of calibration to achieve 80 %, 90 %, and 95 % accuracy, respectively, while C016 required 4.2, 4.7, and 7.3 s. According to the results of the original study and Table 3, C016 provided significantly higher comfort (score of 8.00) compared to plain stimuli (score of 6.69) and was preferred by most participants [17]. The statistical analysis further reinforces these findings, suggesting that both C016 and C032 are excellent candidates for achieving optimal accuracy with only moderate calibration requirements. Given its significant advantages regarding visual comfort and the fact that it requires nearly the same or less calibration time than the plain stimuli condition, C016 remains one of the top-performing options when considering the multi-variable tradeoff: decoding accuracy, decoding speed, and calibration time.

5.3. Calibration and c-VEP stability

In template matching-based decoding algorithms, VEPs elicited by the code stimulation must be accurately estimated to train effective spatial filters. Consequently, the duration of calibration is expected to influence the estimation of these c-VEPs, which in turn impacts spatial filter training and decoding performance. The activation patterns shown in Fig. 4 highlight the strong relevance of the parieto-occipital cortex, particularly over Oz and POz. These values reflect the extent to which each EEG channel contributes to the latent component, i.e. the c-VEP template. Given the visual nature of the paradigm, greater weighting over the primary visual cortex, located in the occipital region, is expected [5]. This emphasis is especially prominent over Oz in the checkerboard database, and over POz in the *p*-ary m-sequences database. Although this spatial shift observed between the databases may seem unexpected, the results should be interpreted from a global perspective. Minor variations in spatial activation patterns may stem from non-neurological factors, such as differences in EEG cap placement across subjects and sessions.

As the primary activation occurs in the occipital cortex, Fig. 5 reveals that the stability of c-VEP estimation at Oz is effectively influenced by the number of observations (i.e., the number of calibration cycles). Assuming that the final averaged VEP, computed using the maximum calibration duration for each condition, serves as a reliable estimator of the brain response, the correlation analysis demonstrates notable variability in the stability of the VEPs among the different conditions. Correlation values of 0.95 for the non-binary and checkerboard databases are achieved with calibration durations of 22.35 ± 15.11 s (ranged from 10.00 to 50.00 s) and 43.44 ± 14.34 s (ranged from 20.48 to 65.63 s), respectively. The durations required to reach a correlation of 0.98 increase to 58.96 ± 28.18 s (ranged from 22.05 to 108.00 s) and 71.07 ± 19.76 s (ranged from 34.13 to 100.80 s), respectively. In general, the calibration duration required to achieve a stable VEP estimation is longer for the checkerboard database compared to the non-binary database. This difference may be attributed to variations in the stimulation paradigm (plain versus BB-CB stimuli) or to quality-related factors in the recordings, such as electrode impedance, users' concentration levels, or physical movements. However, the current analysis alone is insufficient to determine whether the observed difference is attributable to the stimulation paradigm. A dedicated study would be required, involving both paradigms tested with the same participants, within the same session, and with comprehensive measurement of these possible confounding factors. Nevertheless, this variability enhances the heterogeneity of the observed results, supporting their generalizability to other recordings and databases. Based on these findings, we can conclude that approximately a minimum of one minute of calibration is necessary to ensure a robust VEP estimation for template-matching algorithms.

5.4. Strengths and limitations

The results have demonstrated that calibration duration is a critical variable that plays a key role in determining performance in BCIs, yet it is often underexplored in the literature, which primarily emphasizes accuracy and speed, or ITR. We have provided a thorough analysis of the interplay between calibration duration, decoding accuracy, and decoding speed in c-VEP-based BCIs. The inclusion of two publicly accessible datasets (focusing on non-binary m-sequences and spatial frequency variations in c-VEP stimuli) enabled us to incorporate an additional crucial factor: visual comfort. It is important to note, however, that visual comfort was evaluated retrospectively using the original eyestrain ratings from the two pre-recorded datasets, rather than being directly measured as a function of calibration time. Importantly, the analysis was not limited to performance metrics alone; we also examined the stability of c-VEPs estimation as a function of increasing calibration times. Furthermore, we accounted for transmission latency by estimating the delay between screen stimulus onsets and signal reception in the computer, where the EEG samples are timestamped. This analysis, often neglected in existing research and influenced by various factors such as the EEG equipment used, is crucial for obtaining latency-independent results that can be generalized across different systems and hardware configurations. The findings revealed that these four variables form a clear tradeoff that must be carefully considered when designing practical and effective BCIs.

Despite the comprehensive nature of our analysis, several limitations must be acknowledged. First, we adhered to a specific signal processing pipeline; i.e., template matching based on a filter bank in combination with CCA and correlation analysis [5]. Investigating the relationship between calibration duration and system performance under alternative signal processing approaches, such as response modeling methods (e.g., reconvolution [19]), bit-wise decoding techniques (e.g., EEG-Inception [36]), or different spatial projection methods (e.g., task-discriminant component analysis [6] or Riemannian geometry [18]), may yield additional insights. In this vein, exploring novel techniques to implement dynamic stopping procedures for both the calibration [37] and decoding [38,39] stages represents a promising next step toward fully adapting data requirements to individual participants. It would be also worthwhile to investigate whether the required calibration duration could be significantly reduced in the presence of noise or artifacts by applying artifact rejection algorithms. Although we evaluated two different stimulus presentations (plain stimuli and BB-CB patterns), it would be valuable to extend this analysis to other stimulus modalities, such as Gabor and Ricker textures, which have recently attracted attention within the c-VEP BCI community [18]. The interaction between these stimulus modalities designed to enhance visual comfort (spatial frequency and non-binary coding) was not investigated, as the BB-CB patterns were encoded using binary stimuli. Developing a novel experimental paradigm that integrates non-binary encoding with spatial frequency variations would be another promising direction for future research. Additionally, all participants in the datasets were healthy individuals. It remains to be tested whether the observed results generalize to motor-disabled populations, who typically exhibit lower BCI performance [3]. Importantly, the performance analysis was conducted using a chronological and sequentially increasing approach, simulating a realistic calibration process. While currently impractical due to computational cost, applying a cross-validation strategy to account for all possible cycle com-

binations during calibration could potentially enhance the robustness of the estimations. Furthermore, although we explored a multi-variable tradeoff, differences between datasets may be influenced by the number of classes to be decoded (16 commands in the non-binary dataset versus 9 commands in the checkerboard dataset). A more detailed analysis of this variable would be valuable to quantify the extent to which the relationship between the number of BCI commands and the required calibration duration influences the results.

6. Conclusions

In this study, we comprehensively analyzed the c-VEP performance considering a multi-variable tradeoff involving (1) decoding accuracy, (2) decoding speed, and (3) calibration time. To achieve this, we applied an incremental methodology to evaluate the impact of calibration duration across two distinct datasets, each incorporating different stimulation paradigms: plain non-binary stimuli and checkerboard-like (i.e., BB-CB) stimuli. These variations in amplitude depth and spatial frequency of checkerboard patterns have proved to be an effective tool to control visual fatigue compared to classical binary approaches. Accordingly, visual comfort was retrospectively evaluated as an additional fourth variable using these datasets.

Our findings demonstrate that all conditions are capable of achieving over 97 % accuracy with sufficient calibration, underscoring the robustness of the c-VEP circular shifting paradigm. However, a clear tradeoff between calibration duration and performance was identified. Binary-coded stimuli required less calibration time compared to non-binary stimuli to achieve similar levels of accuracy. Specifically, achieving an accuracy of over 90 % within a 2 s decoding window required 4–20 s of calibration for binary conditions and 22–84 s for non-binary conditions. To attain higher performance standards, such as exceeding 95 % accuracy within the same decoding time, calibration times increased to 7–67 s for binary stimuli and 76–221 s for non-binary stimuli. Considering visual comfort, the C016 condition emerged as particularly effective, requiring only 7.3 s of calibration to achieve these standards.

To ensure fair comparisons across different equipment setups, transmission latency was estimated and removed from the analysis. While high theoretical ITRs (200–400 bpm) are attainable, they often mask suboptimal accuracy and are thus unsuitable for practical applications. A statistical equivalence test demonstrated that optimal accuracy was attained within a theoretical ITR range of 22–170 bpm, corresponding to an estimated practical ITR of 22–59 bpm. Furthermore, our analysis highlights the importance of a minimum 1 min calibration duration to adequately estimate the c-VEP response, a critical factor in template-matching algorithms.

In conclusion, achieving optimal performance in c-VEP paradigms necessitates a careful balance between calibration duration, real-time performance, and visual comfort. Our results indicate that while shorter calibration durations are feasible for binary-coded stimuli, non-binary conditions demand longer calibration times. This finding underscores the importance of tailoring calibration protocols to specific experimental and user needs.

CRediT authorship contribution statement

Víctor Martínez-Cagigal: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Eduardo Santamaría-Vázquez:** Writing – review & editing, Software, Funding acquisition, Conceptualization; **Sergio Pérez-Velasco:** Writing – review & editing, Resources; **Ana Martín-Fernández:** Writing – review & editing; **Roberto Hornero:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This research was supported by grants TED2021-129915B-I00 and PID2020-115468RB-I00, funded by MCIN/AEI/10.13039/501100011033 and European Regional Development Fund (ERDF) “A way of making Europe”; under the R + D + i project VA140P24 funded by FEDER and Junta de Castilla y León; and under the R + D + i project “EUROAGE + : Red Internacional de Investigación, Innovación y Transferencia de Tecnologías para la Promoción del Envejecimiento Activo” (“Co-operation Programme Interreg Spain-Portugal POCTEP 2021–2027”) funded by “European Commission” and ERDF; and by “Centro de Investigación Biomédica en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)” through “Instituto de Salud Carlos III” co-funded with ERDF funds. A. Martín-Fernández was in receipt of a PIF grant by the ‘University of Valladolid’.

Appendix A. EEG transmission latency measurement

EEG equipment typically transmits EEG samples without associated temporal timestamps; these timestamps are instead assigned by the receiving computer. Consequently, a transmission latency is expected between the precise moment an EEG sample is acquired and the moment it is time-stamped. This latency is influenced by various factors, including the EEG equipment used, the transmission medium (e.g., wireless or wired), the chunk size, the sampling rate, the number of EEG channels, the computational workload of the receiving computer, and the synchronization accuracy of the local clock, among other potential contributors.

In this study, which focuses on precise calibration and decoding times, it is crucial to account for transmission latency, as variations in the previously discussed factors across different setups can significantly impact the results. Consequently, we accurately measured the EEG transmission latency within the specific setup used for both databases. To achieve this, two HW5P-1 phototransistors (Shenzhen Haiwang Sensor Co., Ltd.) [40] were positioned: one in the top-left corner of the screen, and the other one in the bottom-right corner. This arrangement was designed to capture raster latencies, as pixel lines on a screen are refreshed sequentially from top to bottom. As a result, commands rendered at the bottom of the screen experience additional latency compared to those at the top [41].

The signals generated by the pair of phototransistors were fed into the EEG amplifier. On the target computer, they were transmitted via lab streaming layer (LSL) and acquired and processed using MEDUSA″ [26]. A simple Unity-based paradigm was developed, involving one-frame flashes (the background alternated from black to white) triggered randomly at intervals of 1 to 3 seconds. For each stimulus onset, the peaks detected by the phototransistors, denoted as t_{tl} (top-left) and t_{br} (bottom-right), enabled the measurement of two key metrics: (1) the transmission time between the screen onset t_o and the corresponding EEG timestamps, calculated as $t_{t,i} = t_i - t_o$, where $t_i \in \{t_{tl}, t_{br}\}$; and (2) the actual raster latency between EEG samples, calculated as $t_r = t_{br} - t_{tl}$. It is important to note that t_r quantifies screen delay in the domain of EEG sampling, leading to discrete measurements due to the chunk-based nature of EEG data acquisition. This approach contrasts with a theoretical raster latency measurement, which assesses delay between screen onsets by directly comparing phototransistor delays without processing the signals through the EEG amplifier. While the theoretical raster latency measurement provides valuable insight, only t_r holds actual significance in EEG-based BCIs, as it evaluates whether the latency between top-left and bottom-right measurements is sufficient to impact the temporal accuracy of onset time-tagging within the signal sampling domain.

A total of 162 stimulus onsets were recorded during a separate experiment, conducted independently of the participants’ evaluation. However, the same experimental setup was used; i.e., the same equipment,

number and order of EEG channels, monitor, and computer. The developed application, “PhotoMeasure”, has been made freely accessible through the MEDUSA″ App Market to enable these measurements across diverse experimental setups: medusabci.com/market/photomeasure/.

Appendix B. Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.bbe.2025.10.006

References

[1] Wolpaw J, Wolpaw EW. Brain-computer interfaces: principles and practice. OUP USA; 2012. ISBN 0195388852. https://doi.org/10.1093/acprof:oso/9780195388855.001.0001

[2] Cervera MA, Soekadar SR, Ushiba J, Millán José del R, Liu M, Birbaumer N, et al., et al. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. Ann Clin Transl Neurol 2018;5(5):651–63. . https://doi.org/10.1002/acn3.544

[3] Verbaarschot C, Tump D, Lutu A, Borhanazad M, Thielen J, van den Broek P, et al. A visual brain-computer interface as communication aid for patients with amyotrophic lateral sclerosis. Clinical Neurophysiology 2021;132(10):2404–15. . https://doi.org/10.1016/j.clinph.2021.07.012

[4] Moreno-Calderón S, Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Marcos-Martínez D, Hornero R. Combining brain-computer interfaces and multiplayer video games: an application based on c-VEPs. Front Hum Neurosci 2023;17. . https://doi.org/10.3389/fnhum.2023.1227727

[5] Martínez-Cagigal V, Thielen J, Santamaría-Vázquez E, Pérez-Velasco S, Desain P, Hornero R, et al. Brain-computer interfaces based on code-modulated visual evoked potentials (c-VEP): a literature review. J Neural Eng 2021;18(6):061002. . https://doi.org/10.1088/1741-2552/ac38cf

[6] Shi N, Miao Y, Huang C, Li X, Song Y, Chen X, et al. Estimating and approaching the maximum information rate of noninvasive visual brain-computer interface. Neuroimage 2024;289(120548). . 2308.13232. https://doi.org/10.1016/j.neuroimage.2024.120548

[7] Han J, Xu M, Xiao X, Yi W, Jung T-P, Ming D. A high-speed hybrid brain-computer interface with more than 200 targets. J Neural Eng 2023;20(1):016025. https://doi.org/10.1088/1741-2552/acb105

[8] Sun Q, Zheng L, Pei W, Gao X, Wang Y. A 120-target brain-computer interface based on code-modulated visual evoked potentials. J Neurosci Methods 2022;375(109597). . https://doi.org/10.1016/j.jneumeth.2022.109597

[9] Gembler FW, Rezeika A, Benda M, Volosyak I, et al. Five shades of grey: exploring quintary m -Sequences for more user-Friendly c-VEP-Based BCIs. Comput Intell Neurosci 2020;2020. . https://doi.org/10.1155/2020/7985010

[10] Başaklar T, Tuncel Y, Ider YZ. Effects of high stimulus presentation rate on EEG template characteristics and performance of c-VEP based BCIs. Biomed Phys Eng Express 2019;5(3). . https://doi.org/10.1088/2057-1976/ab0cee

[11] Gembler F, Stawicki P, Rezeika A, Saboor A, Benda M, Volosyak I, et al. Effects of monitor refresh rates on c-VEP BCIs. In: International workshop on symbiotic interaction. Springer International Publishing. ISBN 978-3-319-91592-0; 2018, p. 53–62. . https://doi.org/10.1007/978-3-319-91593-7

[12] Yasinzai MN, Ider YZ. New approach for designing cVEP BCI stimuli based on superposition of edge responses. Biomed Phys Eng Express 2020;6(4). . https://doi.org/10.1088/2057-1976/ab98e7

[13] Behboodi M, Mahnam A, Marateb H, Rabbani H. Optimization of visual stimulus sequence in a brain-Computer interface based on code modulated visual evoked potentials. IEEE Trans Neural Syst Rehabil Eng 2020;2(c). . https://doi.org/10.1109/TNSRE.2020.3044947

[14] Shirzhiyan Z, Keihani A, Farahi M, Shamsi E, GolMohammadi M, Mahnam A, et al., et al. Introducing chaotic codes for the modulation of code modulated visual evoked potentials (c-VEP) in normal adults for visual fatigue reduction. PLoS ONE 2019;14(3):1–29. . https://doi.org/10.1371/journal.pone.0213197

[15] Thielen J. Addressing BCI inefficiency in c-VEP-based BCIs: a comprehensive study of neurophysiological predictors, binary stimulus sequences, and user comfort. Biomedical Physics & Engineering Express 2025;https://doi.org/10.1088/2057-1976/ade316

[16] Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Marcos-Martínez D, Moreno-Calderón S, Hornero R, et al. Non-binary m-sequences for more comfortable brain-computer interfaces based on c-VEPs. Expert Syst Appl 2023;232(120815). . https://doi.org/10.1016/j.eswa.2023.120815

[17] Fernández-Rodríguez Á, Martínez-Cagigal V, Santamaría-Vázquez E, Ron-Angevin R, Hornero R, et al. Influence of spatial frequency in visual stimuli for cVEP-based BCIs: evaluation of performance and user experience. Front Hum Neurosci 2023;17. . https://doi.org/10.3389/fnhum.2023.1288438

[18] Dehais F, Cabrera Castillos K, Ladouce S, Clisson P. Leveraging textured flickers: a leap toward practical, visually comfortable, and high-performance dry EEG code-VEP BCI. J Neural Eng 2024;21(6). . https://doi.org/10.1088/1741-2552/ad8ef7

[19] Thielen J, Marsman P, Farquhar J, Desain P, et al. From full calibration to zero training for a code-modulated visual evoked potentials brain computer interface. J Neural Eng 2021;18(5):56007. . https://doi.org/10.1088/1741-2552/abecef

[20] Miao Y, Shi N, Huang C, Song Y, Chen X, Wang Y, et al. High-performance c-VEP-BCI under minimal calibration. Expert Syst Appl 2024;249(PB):123679. . https://doi.org/10.1016/j.eswa.2024.123679

[21] Zheng L, Dong Y, Tian S, Pei W, Gao X, Wang Y. A calibration-free c-VEP based BCI employing narrow-band random sequences. J Neural Eng 2024;21(2). . https://doi.org/10.1088/1741-2552/ad3679

[22] Stawicki P, Volosyak I. CVEP training data validation–towards optimal training set composition from multi-Day data. Brain Sci 2022;12(2). . https://doi.org/10.3390/brainsci12020234

[23] Thielen J, Sosulski J, Tangermann M. Exploring new territory: Calibration-free decoding for c-VEP BCI. arXiv preprint arXiv:240315521 2024; https://doi.org/10.48550/arXiv.2403.15521

[24] Spüler M, Rosenstiel W, Bogdan M. Unsupervised online calibration of a c-vep brain-computer interface (bci). In: Artificial neural networks and machine learning–ICANN 2013: 23rd international conference on artificial neural networks sofia, bulgaria, september 10-13, 2013. proceedings 23. Springer; 2013, p. 224–31. https://link.springer.com/chapter/10.1007/978-3-642-40728-4_28.

[25] Turi F, Gayraud N TH, Clerc M. Auto-calibration of c-VEP BCI by word prediction 2020;https://hal.science/hal-02844024v1.

[26] Santamaría-Vázquez E, Martínez-Cagigal V, Marcos-Martínez D, Rodríguez-González V, Pérez-Velasco S, Moreno-Calderón S, et al., et al. MEDUSA©: A novel python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. Comput Methods Programs Biomed 2023;230(107357) . https://doi.org/10.1016/j.cmpb.2023.107357

[27] Martínez-Cagigal V. Dataset: Non-binary m-sequences for more comfortable brain-computer interfaces based on c-VEPs. 2025. https://doi.org/10.35376/10324/70945

[28] Martínez-Cagigal V. Dataset: Influence of spatial frequency in visual stimuli for cVEP-based BCIs: evaluation of performance and user experience. 2025. https://doi.org/10.35376/10324/70973

[29] Gembler FW, Benda M, Rezeika A, Stawicki PR, Volosyak I, et al. Asynchronous c-VEP communication tools–efficiency comparison of low-target, multi-target and dictionary-assisted BCI spellers. Sci Rep 2020;10(1):1–13. . https://doi.org/10.1038/s41598-020-74143-4

[30] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM, et al. Brain-computer interfaces for communication and control. Clinical Neurophysiol 2002;113(6):767–91. . https://doi.org/10.1016/S1388-2457(02)00057-3

[31] Nguyen M. A Guide on Data Analysis. Bookdown; 2020. https://bookdown.org/mike/data_analysis/.

[32] Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 2014;87:96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

[33] Kübler A, Kotchoubey B, Kaiser J, Birbaumer N, Wolpaw JR. Brain-computer communication: unlocking the locked in. Psychol Bull 2001;127(3):358–75. 0033-2909.127.3.358.

[34] Martínez-Cagigal V, Santamaría Vázquez E, Gomez-pilar J, Hornero R, et al. Towards an accessible use of smartphone-Based social networks through brain-computer interfaces. Expert Syst Appl 2019;120. https://doi.org/10.1016/j.eswa.2018.11.026

[35] Pérez-Velasco S, Santamaría-Vázquez E, Martínez-Cagigal V, Marcos-Martínez D, Hornero R. EEGSym: Overcoming inter-subject variability in motor imagery based BCIs with deep learning. IEEE Trans Neural Syst Rehabil Eng 2022;30:1766–74. . https://doi.org/10.1109/TNSRE.2022.3186442

[36] Santamaría-Vázquez E, Martínez-Cagigal V, Hornero R. Bit-wise reconstruction of non-binary visual stimulation patterns from EEG using deep learning: a promising alternative for user-friendly high-speed c-VEP-based BCIs. In: International work-Conference on artificial neural networks. Springer; 2023, p. 603–14. https://doi.org/10.1007/978-3-031-43078-7_49

[37] Sato J, Washizawa Y. Reliability-based automatic repeat request for short code modulation visual evoked potentials in brain computer interfaces. In: 2015 37Th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2015, p. 562–5. https://doi.org/10.1109/EMBC.2015.7318424

[38] Martínez-Cagigal V, Santamaría-Vázquez E, Pérez-Velasco S, Marcos-Martínez D, Moreno-Calderón S, Hornero R. Nonparametric early stopping detection for c-VEP-based brain-computer interfaces: a pilot study. In: 2023 45Th annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE; 2023, p. 1–4. https://doi.org/10.1109/EMBC40787.2023.10341024

[39] Ahmadi S, Desain P, Thielen J. A bayesian dynamic stopping method for evoked response brain-computer interfacing. Front Hum Neurosci 2024;18:1437965. https://doi.org/10.3389/fnhum.2024.1437965

[40] Ltd S. HW5P-1 Specs: Ambient light detector Photosensitive sensor. Tech. Rep.; 2015. https://cdn-shop.adafruit.com/product-files/2831/HW5P-1_2015_1_.pdf.

[41] Nagel S, Dreher W, Rosenstiel W, Spüler M, et al. The effect of monitor raster latency on VEPs, ERPs and brain-computer interface performance. J Neurosci Methods 2018;295:45–50. . https://doi.org/10.1016/j.jneumeth.2017.11.018