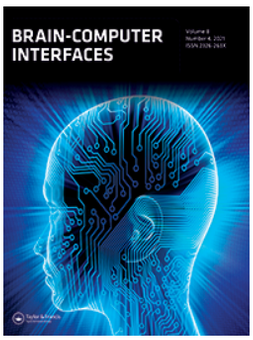

Brain-Computer Interfaces

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbci20

Improving longitudinal P300-BCI performance for people with ALS using a data augmentation and jitter correction approach

Alyssa Hillary Zisk, Seyyed Bahram Borgheai, John McLinden, Roohollah Jafari Deligani & Yalda Shahriari

To cite this article: Alyssa Hillary Zisk, Seyyed Bahram Borgheai, John McLinden, Roohollah Jafari Deligani & Yalda Shahriari (2021): Improving longitudinal P300-BCI performance for people with ALS using a data augmentation and jitter correction approach, Brain-Computer Interfaces, DOI: 10.1080/2326263X.2021.2014678

To link to this article: https://doi.org/10.1080/2326263X.2021.2014678

Published online: 23 Dec 2021.

BRAIN-COMPUTER INTERFACES
https://doi.org/10.1080/2326263X.2021.2014678

Taylor & Francis
Taylor & Francis Group

ORIGINAL RESEARCH

Improving longitudinal P300-BCI performance for people with ALS using a data augmentation and jitter correction approach

Alyssa Hillary Zisk[a], Seyyed Bahram Borgheai[b], John McLinden[b], Roohollah Jafari Deligani[b] and Yalda Shahriari[a,b]

[a]Interdisciplinary Neuroscience Program, University of Rhode Island, Kingston, RI, –USA; [b]Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Fascitelli Center for Advanced Engineering, –USA Kingston, RI, USA

ABSTRACT

The P300 Brain-Computer Interface (BCI) is a well-established communication channel for severely disabled people. However, variation in P300 latency, or latency jitter, is both increased in people with amyotrophic lateral sclerosis (ALS) and negatively associated with BCI performance. In this study, we proposed an augmentation and correction (A/C) characterization scheme with data augmentation and correction for jitter, both relying on time-shifted responses with individualized parameters determined based on latency jitter. We tested this approach offline on longitudinal data collected from six participants with ALS. While our longitudinal analysis showed decreased BCI performance and increased latency jitter over time with both our proposed characterization scheme and conventional methods, our proposed A/C characterization scheme significantly improved character selection accuracy, required for usability, along with recall and F-scores, showing the effectiveness of our proposed approach. These results should inform further work on improving longitudinal BCI performance and reliability for people with ALS.

ARTICLE HISTORY
Received 14 August 2021
Accepted 1 December 2021

KEYWORDS
Brain-computer interfaces; amyotrophic lateral sclerosis; data augmentation; latency jitter; jitter correction

1. Introduction

As people with amyotrophic lateral sclerosis (ALS) develop significant motor disability and lose voluntary motor control, including speech-related motor control [1], they frequently require tools for augmentative and alternative communication. Currently available tools including brain-computer interfaces (BCIs) support communication for people with ALS, but people with ALS have been reported to show reduced BCI performance in comparison to neurotypical users [2–6]. In addition, BCI users can experience substantial variations in BCI performance within and across days [7–9]. Given these concerns, significant attention has been paid to both understanding the correlates of BCI performance and improving selection accuracy [10–15]. Much of this research is dedicated to understanding and improving BCIs based on the visual P300 response, a positive electrical deflection occurring 250–500 ms after an attended rare event [16,17]. In a longitudinal study of P300-based BCI home users, Shahriari and colleagues found that BCI performance was positively correlated with P200 amplitude, parietal alpha-band spectral power, and occipital beta-band spectral power, and that these measures were significantly increased for successful BCI sessions as compared to unsuccessful sessions. They also found BCI performance was negatively correlated with occipital delta-band power [9]. Mak and colleagues found that among participants with ALS, increased event-related potential (ERP) amplitudes and theta-band spectral power were associated with increased P300 BCI performance [18]. Geronimo and colleagues found that higher cognitive scores, including scores measuring attention, were associated with both increased P300 quality and BCI performance [6,19].

Trial to trial variation in P300 latency, known as latency jitter, has been found to be negatively associated with BCI performance in a mixed group of neurotypical participants and potential end-users [20], in neurotypical participants [21,22], and in people with ALS [4]. For example, Zisk and colleagues recently determined that this latency jitter is elevated in people with ALS as compared to neurotypical controls [4], and latency jitter is a factor affecting BCI performance for people with ALS [4,20,23].

As studies have shown latency jitter can predict BCI performance [20,21], Mowla and colleagues used latency estimation and a secondary classifier to improve BCI performance, though they did not report an online implementation of this method [24]. Togashi and

CONTACT Yalda Shahriari yalda_shahriari@uri.edu Interdisciplinary Neuroscience Program, University of Rhode Island, Tyler Hall, 9 Greenhouse Rd, Kingston, RI 02881, USA.

Washizawa similarly utilized Bayesian latency estimation to improve P300 BCI performance [25]. Considering differences in latencies between experimental paradigms which elicited P300 responses rather than variability within participants using a single paradigm, Iturrate and colleagues calculated the latency shift between paradigms and then trained a classifier for one paradigm using data from another, time-shifted to compensate for the latency differences between the experimental conditions [26]. They found that in cases of insufficient training data from any given paradigm, including latency-corrected training data from other paradigms improved performance.

In recent years, data augmentation for BCIs has gained attention as a strategy for improving performance [27–32].The purpose of data augmentation is to increase the size of the training data, and thereby improve the reliability and generalizability of the classification algorithms. As electroencephalography (EEG) data varies significantly between different participants, many EEG classifiers are subject-specific, though pooling data from multiple participants has also been the focus of some research [17,32–34] with a similar goal of improving generalizability and reliability. Iturrate and colleagues collected data from multiple experimental paradigms that produced P300 responses, but with different latencies. Their transfer of data between different experiments that evoke P300 responses similarly works toward the goal of improving generalizability and stability with limited training data [26]. In other studies, the use of time shifted epochs has supported the extraction of multiple segments per stimulus, providing a larger training data set [31,35,36] or helping mitigate class imbalances [37]. For example, Kim and colleagues used a 100 ms shift based on the results of a pre-analysis to find optimal windows for classifying error-related potentials. Their use of two epochs per event with different starting times allowed for variation in when participants detected errors and doubled the size of their data set in a reinforcement learning method [36]. By requiring both the shifted and unshifted epoch to be classified correctly for the classification to be considered correct, they additionally improved the reliability of any positive feedback. In their study, Kim and colleagues noted an improvement with this data augmentation scheme as compared to using a single epoch without augmentation [36]. Sakai and colleagues compared several data augmentation methods based on the time and amplitude fluctuations associated with bio-signals such as EEG, including set time shifts of ±10 ms for all participants, which tripled their training data sets while considering potential latency variability [35]. Their data augmentation protocol improved classification performance, with greater improvements found when the training set was smaller. Krell and colleagues similarly considered several data augmentation methods, including time-shifted data, for augmenting P300 training data [31]. They initially tested single time-shifts as several factors, including both lag in the display and increases in workload, can lead to delayed P300 latencies. The single time-shifts provided improvements for some participants, but no single time-shift was reported to be consistently helpful. They then tested symmetrical time-shifts and reported that ±40 ms shifts increased the data set but did not significantly affect performance [31]. In all three studies, unshifted epochs, beginning at the time of the stimulus, were used alongside overlapping time-shifted epochs extracted from the recorded EEG data. These three studies sought to classify responses which can vary in latency, and their use of time-shifted data both increased the number of epochs available for training and provided epochs with earlier and/or later responses of interest [31,35,36]. As data augmentation with time-shifted data provides time-shifted responses in the training data, this augmentation approach provides additional latency variability that may improve robustness to this same form of variability [27].

In this study we therefore proposed a correction strategy that relied on latency jitter at multiple levels. In particular, we proposed to improve classification performance for P300 data longitudinally recorded from people with ALS using both data augmentation and jitter correction, tested offline. The data augmentation utilized time-shifted responses to both target and non-target trials, with individualized time shifts based on latency variations present in the training set. The jitter correction procedure was also implemented through allowing limited time-shifts of the epochs to be classified. We quantified our performance improvements through the use of a reference classifier using neither data augmentation nor jitter correction. We then investigated longitudinal relationships between clinical measures, latency jitter, and BCI performance in our participants with ALS.

2. Materials and methods

2.1. Participants

Six participants with ALS (age 57 ± 15.7 years, 1 female) were recruited for this study (see Table 1). All participants had at least some post-secondary education. Participants other than ALS-01 had normal or corrected to normal vision, while ALS-01 was in the late stages of

Table 1. Demographic information for participants with amyotrophic lateral sclerosis (ALS).

Subject Number	Age	Sex	Time since diagnosis (years)	Revised ALS Functional Rating Scale (ALSFRS-R) (out of 48)	ALSFRS-R Bulbar Subscore	Artificial Ventilation	Means of Communication
ALS-01	29	M	4	0	0	Yes	No reliable means
ALS-02	55	M	11	4	0	Yes	Eye-tracking
ALS-03	70	M	8	14	5	No	Non-verbal sound
ALS-04	67	M	2	7	5	Yes	Eye-tracking
ALS-05	69	F	11	23	11	No	Verbal
ALS-06	52	M	3	22	12	No	Verbal
Mean ± SD	**57.0 ± 15.7**	-	**6.5 ± 4.0**	**11.6 ± 9.5**	**5.5 ± 5.2**	-	-

locked-in syndrome with significant ocular impairments. Participants were diagnosed with ALS 6.5 ± 4.0 years prior to the start of the study and had an average functional rating scale-revised (ALSFRS-R) score of 11.6 ± 9.5, with a minimum score of 0 indicating no voluntary motor functions and complete dependence on life-sustaining technologies including mechanical ventilation and a maximum score of 48 indicating normal functioning [38]. Three participants had gastrostomies as well as tracheostomies. ALS-01′s sole form of communication was an idiosyncratic and error-prone yes/no pupil dilation his caregiver read subjectively, which deteriorated over the course of the recordings, losing reliability as a means of communication. Two other participants with artificial ventilation (ALS-02 and 04) used eye-tracking devices to communicate. ALS-03 could still move his index finger and make non-verbal sounds to sustain minimal communication. ALS-05 and 06 retained the ability to speak, though ALS-05 had lost non-facial movement, and ALS-06 could barely move a joystick with one hand. Participants were tested in their homes or care centers. The study protocol was approved by the Institutional Review Board (IRB) of the University of Rhode Island (URI), and all participants provided informed consent or assent for the study and received financial compensation.

2.2. Experimental protocol

Each participant took part in 5–12 (9.5 ± 2.6) sessions of recording over 2.5–13.7 (10.9 ± 4.3) months. These sessions took place at least two weeks apart. Including preparation such as the application of gel to electrodes and impedance calibration, each session typically lasted 2–2.5 hours. To familiarize participants with the BCI setup, including the recording protocol and the task, each participant took part in a single familiarization session before the main experimental recordings, in which they completed the same tasks without recording the data and could get clarification about the experimental tasks. Each session contained one run of a standard P300 spelling protocol, in which a 6 × 6 matrix of characters containing letters and numbers was displayed to participants, with each row and column intensified 10 times (i.e. 10 repetitions) per character selection [10,39]. Intensifications lasted 93.75 ms and were separated by an inter-stimulus interval of 62.5 ms. Participants copy-spelled 14 characters with 4 second pauses between characters, counting intensifications of their intended (target) character, without real-time feedback in each session.

2.3. Data acquisition

EEG data were recorded using a g.USBamp amplifier (g.tec Medical Technologies) with a 256 Hz sampling rate. Data were recorded from eight channels commonly used in P300 protocols, Fz*, Cz, P3, Pz, P4, PO7, PO8, and Oz [30]. However, as Fz was occupied by sensors for other studies recorded in the same session as the current experiment, it was replaced by the nearest available channel, FAF2, denoted as Fz*. All experimental protocols, data acquisition, and stimulus presentation were controlled using BCI2000 software [40].

2.4. Signal pre-processing

All analyses were conducted offline in MATLAB R2019a. EEG data were detrended and bandpass filtered at 0.5–30 Hz with a Hamming window-based zero-phase filter. For P300-based BCI applications, a 0–800 ms post-stimulus window is common [5–9,41,42], covering important ERP features including the P200, N200, P300, and late negativity [5,9]. Thus, to allow this typical 800 ms segment epoch to be shifted by up to 100 ms in either direction for use in classifier-based latency estimation (CBLE) [20], the data were segmented into 1 s epochs, from 100 ms pre-stimulus to 900 ms post-stimulus epochs. From each of these epochs, 53 time-shifted 800 ms sub-epochs per stimulus were extracted using an 800 ms moving window with a step size of one sample, ~3.9 ms at 256 Hz. These 800 ms sub-epochs were subject to a moving average procedure,

where each value was replaced by the local mean calculated over a 13-sample moving window, and then downsampled by a factor of 13, following the feature reduction procedure used in previous studies [20]. The downsampled sub-epochs from all channels were concatenated and then treated as potential features for further classification. The true class labels were 1 for the sub-epochs extracted from the 1 s epochs around target stimuli and 0 for the sub-epochs extracted from the 1 s epochs around non-target stimuli.

2.5. Data analysis

Figure 1 provides an overview of our proposed data analysis method, a characterization scheme including data augmentation and jitter correction, hereafter called augmentation/correction (A/C) characterization. In this method, first, latency shifts were calculated on the training set using CBLE [20], providing a series of classifier scores. Then, the training data was augmented with time-shifted sub-epochs. The data augmentation parameters were determined based on the calculated latency shifts from the training set. The allowable range of time shifts to be used in the jitter correction procedure was then determined over the training set. Throughout the A/C process, stepwise linear discriminant analysis (SWLDA) classifiers were used with typical parameters for P300 speller applications: in each step, the most significant feature for predicting if an epoch was a target with $p < 0.1$ was added, and then if applicable the least significant feature with $p > 0.15$ was removed, up to a maximum of 60 included features or until no features satisfied entry/removal criteria [41,42]. SWLDA classifiers were similarly used for all comparison conditions. For all analysis conducted within the training set, the training data were divided into five folds of approximately equal size, by character. That is, classifiers were trained on four folds for application to the fifth, and this procedure was repeated four more times to test each fold once.

Then, on the test set, for each stimulus we extracted all of the 800 ms sub-epochs with time shifts within the range determined on the training set. These 800 ms sub-epochs were fed to the final classifier, and the maximum classifier score over the selected 800 ms sub-epochs corresponding to a stimulus was retained as the score for that stimulus to correct for latency jitter. These steps are explained in more detail in sections 2.5.1–2.5.3.

To ensure that the proposed A/C approach could be implemented in practical environments, for each participant, data from prior sessions were used to predict performance and determine correction parameters for future sessions. Beginning with each participant's third session, session performances were evaluated by taking that participant's two prior sessions as the training set. That is, first, classifiers were trained and parameters were determined using data from each participant's first two sessions and then evaluated the data of their third session as its test set; then classifiers were trained on the second and third sessions to evaluate their fourth session, and so forth. As 5–12 sessions were recorded from each participant, we therefore had 3–10 training sets and corresponding test sessions per participant, for a total of 45 training sets and corresponding test sessions over all 6 participants.

2.5.1. Latency shift and latency jitter calculations

Figure 1 shows a schematic of how classifier score series and latency shifts were calculated over the training set. All calculations of classifier score series, latency shifts,

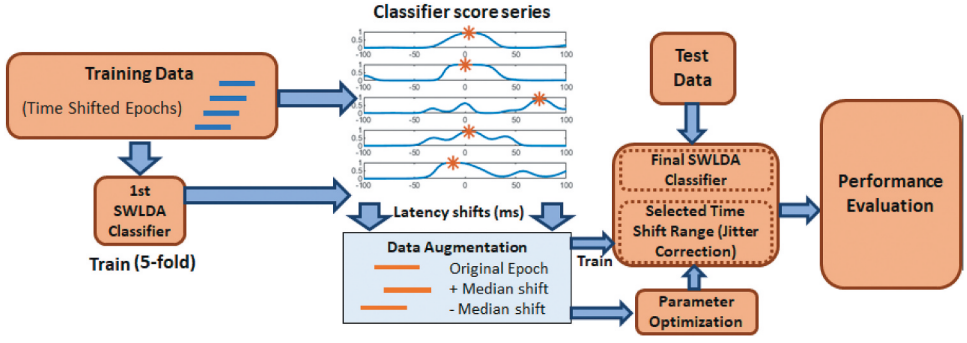

Figure 1. The schematic illustrates the training process for the A/C characterization scheme with data augmentation and jitter correction. **Latency shift and jitter calculation**: A classifier was trained, time-shifted sub-epochs were classified, and the classifier scores for time-shifted sub-epochs were used to extract latency shifts for each stimulus. **Data augmentation and parameter optimization**: The latency shifts calculated were used to determine the constant symmetric shift, *M*, which was used to augment the training data. Parameters were optimized and a final SWLDA classifier was trained on the full augmented training data. **Classification and jitter correction**: The final SWLDA classifier was applied to the test data with jitter correction, and performance was evaluated.

and latency jitter relied on classifier-based latency estimation (CBLE), as proposed by Thompson and colleagues [20] and used in our prior investigation of latency jitter as well [4]. In CBLE, the sensitivity of a classifier to latency variability is used to estimate latency shifts for single trials using time-shifted data [20,21]. As a first step for CBLE, an SWLDA classifier was trained on the unshifted 0–800 ms sub-epochs from four fifths of the training set (5-fold training) using the true class labels of 0 for sub-epochs corresponding to non-target stimuli and 1 for target stimuli. Then, for each stimulus in the fifth fold, the downsampled 800 ms sub-epochs, including all 53 time-shifted sub-epochs, were extracted and fed to the classifier. This resulted in 53 classifier scores, each of which is the post-probability that the shifted sub-epoch corresponded to a target stimulus. The time shift corresponding to the highest classifier score in the series was extracted as the *'latency shift'* for that specific stimulus, recorded in milliseconds. This procedure was repeated five times, such that data from each fold was fed to a classifier trained on the other four folds, providing a latency shift for each stimulus in the training set. Then, *latency jitter* was calculated as the variance of the latency shifts calculated through classifier-based latency estimation (vCBLE) for all target stimuli, similar to the procedure reported by Thompson and colleagues [20].

2.5.2. Data augmentation and parameter optimization

After extracting the latency shifts from all stimuli in the training set, the training data were augmented using symmetrically time-shifted data similar to the protocol in [35], but with an individualized adaptive time shift calculated using the latency shifts in the data (see Figure 1). First, the median of the absolute latency shifts was calculated over all target stimuli in the training set. This median, $\boldsymbol{M}$, was used as the constant symmetric time shift to augment that training set. Specifically, an -$\boldsymbol{M}$ to -$\boldsymbol{M}$ + 800 ms sub-epoch and an $\boldsymbol{M}$ to $\boldsymbol{M}$ + 800 ms sub-epoch were extracted for each stimulus. This procedure tripled the original training data (3x).

In addition, if the latency jitter in the training set was greater than 1000 ms^2, then per-epoch data augmentation was also performed based on individual, per-stimulus latency shifts to compensate for the excess latency variation in the training data. In this case, for all stimuli, both target and non-target, we additionally extracted the sub-epoch corresponding to the latency shift calculated for the stimulus in section 2.5.1. That is, for a stimulus with a calculated latency shift of $\boldsymbol{S}$ ms based on its classifier score series that reached its maximum for the $\boldsymbol{S}$ to $\boldsymbol{S}$ + 800 ms sub-epoch, this S to S + 800 ms sub-epoch was also added to our training set. In cases where this additional sub-epoch was used, the number of sub-epochs extracted from the training set was quadrupled (4x), with the original 0–800 ms sub-epoch, two symmetrically time shifted sub-epochs (-$\boldsymbol{M}$ to -$\boldsymbol{M}$ + 800 ms and $\boldsymbol{M}$ to $\boldsymbol{M}$ + 800 ms), and a jitter corrected sub-epoch ($\boldsymbol{S}$ to $\boldsymbol{S}$ + 800 ms) corresponding to each stimulus.

Then, the range of time shifts used for jitter correction was optimized. For jitter correction, each stimulus was assigned the maximum classifier score corresponding to a range of time-shifted sub-epochs. Out of all the possible ranges, the optimal range of time shifts to use for jitter correction was optimized using 5-fold cross-validation. For this purpose, a new SWLDA classifier, only used in determining the range of time shifts to be used in jitter correction, was trained on four folds of the augmented training data. As each new epoch added through the data augmentation procedure corresponds to one original stimulus, all additional epochs were assigned to the same fold as their corresponding stimulus and original 0–800 ms epoch.

Using this classifier, the optimal range of time shifts to use for jitter correction was determined using 5-fold cross-validation. The optimal range was selected out of a total of 27 possible ranges corresponding to the central 1, 3, 5, . . ., 53 classifier scores distributed symmetrically around the score for the 0–800 ms epoch, ranging from no correction to the use of the entire classifier score series. These ranges provided maximum allowable time shifts of 0 ms, ±3.91 ms, ± 7.81 ms, . . . ±101.56 ms, corresponding to intervals between data points recorded at 256 Hz. Limited time ranges were tested to partially mitigate the increased false positive rate caused by using the maximum classifier score, or maximum post-probability that an epoch corresponded to a target stimulus [20].

To determine the optimal range, each possible range was tested. To test a possible range of time shifts, for each stimulus (in the training set), classifier score series were calculated using classifiers trained on the augmented data. Then, the maximum classifier score corresponding to an epoch within the range being tested was retained as the final classifier score. That is, for each of the possible ranges of allowable time shifts, classifier scores and class labels were assigned to each stimulus in the training set using the classifier score series calculated for that stimulus. The score for the stimulus and range of allowable time shifts was the maximum score for the stimulus within that range, and the label was assigned according to this score. In effect, if any epoch within the allowable range of time shifts

would have been labeled as a target, then the stimulus was also labeled as a target. If not, then the stimulus was labeled as a non-target.

This was repeated for all possible ranges of allowable time shifts, and the range which maximized the average F-score ($\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$) over the five folds was selected as the optimal range for implementation on the test set.

2.5.3. Classification and jitter correction

The final SWLDA classifier trained on all five folds of the augmented training data and the selected jitter correction range, shown in Figure 1, were then used both to classify data from the test session and to calculate latency jitter on the test session.

For jitter correction, the 800 ms sub-epochs within the time shift range determined in section 2.5.2 were extracted for each stimulus. These time-shifted sub-epochs were fed to the classifier, and the maximum score corresponding to each stimulus was retained for that stimulus. In addition, if this score was greater than 0.5, meaning that at least one of the time-shifted sub-epochs corresponding to the stimulus was labeled a target, then the stimulus was accordingly labeled a target. All performance metrics with the A/C characterization procedure were then calculated using these scores and labels.

In addition to being used for jitter correction, classification of time-shifted sub-epochs was also used to calculate jitter using CBLE. For this purpose, the full classifier score series was calculated, using all 53 time-shifted 800 ms sub-epochs. The time shift for the sub-epoch with the maximum classifier score over this entire range was recorded as the latency shift for that stimulus, then used to calculate vCBLE as in section 2.5.1.

2.5.4. Performance evaluation

Binary classification accuracy, precision, recall, F-score, and character selection accuracy were calculated as measures of performance [43,44]. With TP, TN, FP, and FN respectively representing the number of stimuli that were classified as true positives (correct targets), true negatives (correct non-targets), false positives, and false negatives, we computed accuracy, precision, recall, and F-score as below:

$$\text{Binary Classification Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

$$\text{F}-\text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

The character with the highest summed classifier score over all repetitions of its row and its column was selected. Character selections were calculated for all possible numbers of repetitions, from 1 (only the first intensification of each row and column) through 10 (all 10 intensifications of each row and column per character). For each number of repetitions, character selection accuracies were then calculated as the number of characters correctly selected from a test session divided by the 14 characters in each session.

Although row and column intensifications were consistently repeated 10 times per character selection in the recordings, theoretical utility values were also calculated for each number of repetitions as a measure of potential throughput. These theoretical utility values used the formula $\frac{(2p-1)\log_2(N-1)}{c}$, where p is the portion of characters spelled correctly, N is the number of possible selections (36), and c is the time to select a character, whether correctly or incorrectly [45,46]. Notably, this formula requires character accuracy to be at least 50% to achieve positive utility, as correcting errors would otherwise take infinite time. As intensifying each row and column once took 1.875 seconds and there was a 4 second pause between characters, the time to select a character was 4 + 1.875 r seconds, where r was the number of repetitions used. The maximum theoretical utility and the number of repetitions used to reach it was retained for each test session. For test sessions where the utility was uniformly 0 regardless of the number of repetitions used, the number of repetitions required was considered to be 10, the full number recorded.

In addition to calculating these performance metrics for our proposed A/C characterization scheme, we calculated them for two comparison conditions to test the A/C scheme. First, to compare our proposed A/C characterization scheme to conventional procedures, an SWLDA classifier was trained on the two training sessions without any data augmentation or jitter correction. This classifier was used as the reference classifier. Second, we validated our parameter selection methods by comparing the results of our proposed procedure to the results of data augmentation and jitter correction with random parameters explained in section 2.6.

2.6. Statistical analysis

Statistical analyses other than the random parameter testing were conducted in R version 4.0.5 [47]. Differences between the proposed A/C characterization method and the reference classifier were investigated using paired t-tests comparing participant average metrics (n = 6). Per-stimulus performance metrics, specifically binary classification accuracy, precision, recall, and F-score were averaged within participants. Character selection accuracy for each possible number of repetitions, from 1–10, was also averaged within participants, as was maximum theoretical utility. Participant average jitter, per-stimulus performance metrics, character selection accuracy using all 10 repetitions, and maximum theoretical utility were compared between the proposed A/C characterization method and the reference classifier using Wilcoxon signed-rank tests [48].

We then tested for correlations between performance metrics and latency jitter for both classification methods using repeated measures correlations, (r_{rm}), an analysis of covariance-based regression appropriate for measuring common (overall) intra-individual associations between measures with multiple non-independent observations per participant [49].

We further investigated associations between clinical measures and performance improvements from our proposed method. To do so, we tested for spearman correlations between participant-averages in selection accuracy improvements from our proposed A/C characterization procedure relative to the reference classification approach, and time since diagnosis, ALSFRS-R scores, and ALSFRS-R bulbar subscores. We also tested for correlations between selection accuracies using each method and clinical scores.

Latency jitter and performance metrics were also investigated longitudinally. We utilized repeated measures correlations to investigate common trends across participants. To understand possible changes in performance over time, repeated measures correlations combining information from all participants were investigated between the number of days since the first session and latency jitter, as well as the number of days since the first session and all performance metrics. As prior studies have noted variations in long-term trends between participants, we additionally tested for spearman correlations between character selection accuracies and days since their first session within each participant to consider inter-individual differences in trends.

In addition to the statistical analyses of the proposed A/C scheme and reference classifier, we tested the effectiveness of our selected parameters with a Monte Carlo experiment to determine how likely similar improvements would be to occur by chance [50,51]. To do so, random parameters, specifically symmetric time shift, use or nonuse of per-stimulus augmentation, jitter correction range, were selected and performance metrics calculated 1000 times, with individually randomized parameters for each training set and corresponding test session, in alignment with the individually determined parameters for each training set and corresponding test set in the A/C characterization scheme. Specifically, three parameters were randomized in each case: 1) the constant symmetric time shift (M), which could theoretically vary between ±3.91 ms and ±101.56 ms, with 26 possible values (±3.91 ms, ± 7.81 ms, … ±101.56 ms, corresponding to intervals between data points recorded at 256 Hz); 2) a binary variable (yes or no) representing whether or not data augmentation was additionally performed based on per-stimulus latency shifts; and 3) the range of time shifts used for jitter correction, which could vary from 0 ms (no jitter correction) to ±101.56 ms, with 27 possible allowable ranges (originally optimized with 5-fold cross-validation; randomized here). That is, for each training set and corresponding test session, there were a total of 1404 (=26*2*27) possible parameter combinations. As parameters were individually determined for each training set and corresponding test session (45 training sets and corresponding test sessions), there were therefore 1404^{45} possible parameter combinations over the entire study's data.

For each of the 1000 randomly selected parameter sets, we classified each test session's data using an SWLDA classifier trained on the training set augmented with the randomly selected parameters, with jitter correction using the randomly selected time shift range. Performance metrics from the proposed A/C characterization method were then also compared to performance metrics from the Monte Carlo experiment testing jitter augmentation and data correction with random parameters. Within each of the 1000 random parameter sets, these performance metrics were then averaged over all sessions for each participant, and then the participant averages were averaged to yield overall performance metrics. This process yielded 1000 sets of overall performance metrics, one for each random parameter set. The proportion of random parameter sets for which classification with data augmentation and jitter correction outperformed the proposed A/C characterization method with algorithmically determined parameters was then calculated.

3. Results

Average target and non-target responses at channel Cz from a representative session for each participant are shown in Figure 2. As visible in the figure, the extent to

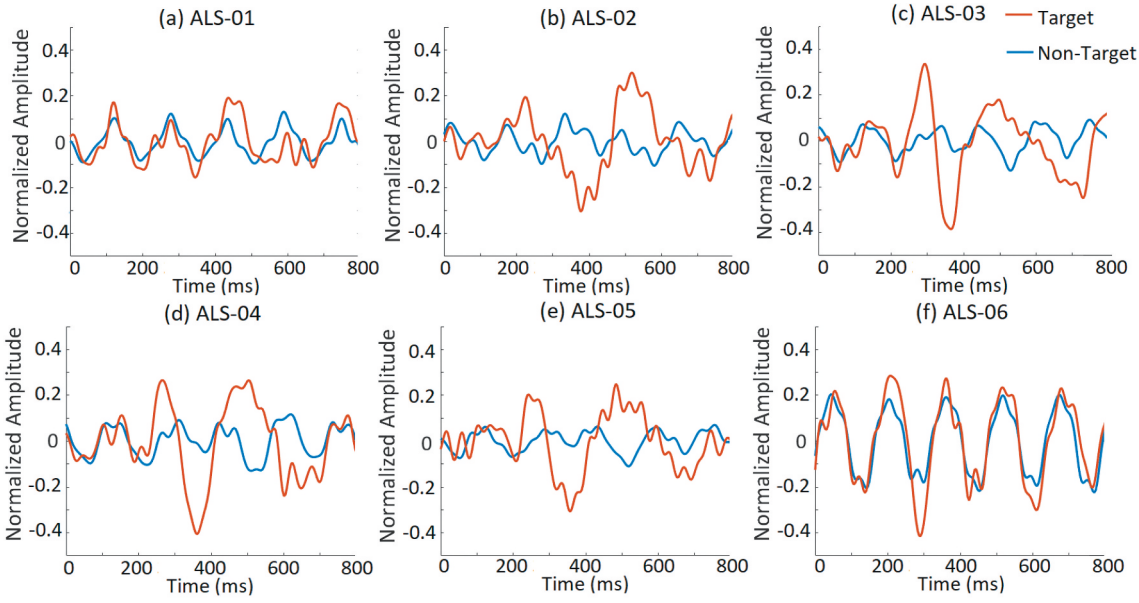

Figure 2. Average target and non-target responses at channel Cz from a representative session for participants ALS-01 (a) through ALS-06 (f).

which ERP features such as the P200, N200, and P300 are visibly present in the average target response varied between participants.

3.1. Augmentation and correction quantification

The symmetric shifts used for data augmentation and ranges of allowable time shifts for jitter correction varied between participants and sessions, as did latency jitter. For each participant, the minima, maxima, means, and standard deviations of their determined A/C parameters and calculated latency jitters over all their recorded sessions are reported in Table 2. The symmetric shifts used for data augmentation varied between ±11.72 ms and ±58.59 ms, though shifts greater than ±30 ms were only selected for ALS-01, who was both the participant in the locked-in state and the participant with the highest average latency jitter. The selected ranges of allowable time shifts used for jitter correction ranged from 0 (no allowable time shift) to ±101.56 ms, though ranges greater than ±55 ms were also only selected for the participant in the locked-in state. The selected parameters for each combination of training and testing session numbers, specifically the symmetric shifts used in data augmentation, the relative size of the augmented training set compared to the original training data (4x if per-epoch augmentation was used versus 3x if not), and the time shift range, are available in Table A.1 for each participant and session.

3.2. Evaluation of the augmentation and correction scheme

Table 3 tabulates the individual results on all participant average measures evaluating the augmentation and correction scheme. Character accuracy when using all 10 repetitions was significantly (p = 0.031) higher with the A/C classifier at 75.69 ± 32.40% than with the reference classifier at 71.55 ± 32.33%. In particular these changes were 2.9%, 5.7%, 2.4%, 7.1%, 2.6%, and 4.1% for participants ALS-01 through ALS-06, respectively. The difference between maximum theoretical utility with the A/C characterization scheme (16.49 ± 9.65 bit/min) and the reference classifier (13.54 ± 8.63 bit/min) was not

Table 2. Per-participant summary statistics for augmentation and correction (A/C) parameters and latency jitter.

	Augmentation Shift (ms)			*Correction Window (ms)*			*Jitter (ms^2)*		
Participant	Min	Max	Mean ± SD	Min	Max	Mean ± SD	Min	Max	Mean ± SD
ALS-01	19.53	58.59	41.02 ± 11.83	0	101.56	40.23 ± 43.19	2915	4817	3843 ± 548
ALS-02	11.72	23.44	16.80 ± 3.71	7.81	54.68	23.44 ± 14.96	509	1807	999 ± 435
ALS-03	15.63	19.53	16.93 ± 2.26	7.81	19.53	13.02 ± 5.97	591	1345	939 ± 380
ALS-04	15.63	27.34	20.09 ± 4.18	27.34	46.88	36.38 ± 6.32	1035	3591	2037 ± 1144
ALS-05	11.72	19.53	15.14 ± 2.50	7.81	27.34	12.70 ± 7.16	557	2985	1381 ± 947
ALS-06	15.63	27.34	19.53 ± 4.51	7.81	50.78	27.34 ± 17.47	383	1524	1018 ± 378

Table 3. Average accuracy metrics for both reference and augmentation and correction (A/C) classification schemes for each participant. *significant at $p < 0.05$, Wilcoxon signed-rank test. Means and standard deviations (SD) are provided for each classification method.

	*Character Accuracy (%)**		*Utility (bit/min)*		*Binary Accuracy (%)**	
Participant	Reference	A/C	Reference	A/C	Reference	A/C
ALS01	10.71	13.57	0	0	81.73	80.95
ALS02	87.14	92.86	17.46	22.76	85.14	83.77
ALS03	97.62	100.00	20.32	22.19	88.51	88.37
ALS04	60.20	67.35	5.39	9.23	79.78	76.94
ALS05	84.82	87.50	18.50	22.17	85.44	84.56
ALS06	88.78	92.86	19.57	22.59	86.68	85.20
Mean ± SD	**71.55 ± 32.32**	**75.69 ± 32.40**	13.54 **± 8.63**	16.49 **± 9.65**	**84.55 ± 3.23**	**83.29 ± 3.93**
	Precision (%)		*Recall (%)**		*F-score**	
Participant	Reference	A/C	Reference	A/C	Reference	A/C
ALS01	30.09	36.31	7.04	7.50	0.11	0.11
ALS02	55.42	52.96	58.93	72.86	0.56	0.60
ALS03	66.40	63.49	66.40	72.74	0.63	0.67
ALS04	46.37	46.39	30.00	38.27	0.33	0.36
ALS05	64.83	60.14	49.24	60.22	0.52	0.57
ALS06	66.40	60.82	52.55	62.35	0.57	0.59
Mean ± SD	**54.92 ± 14.50**	**53.35 ± 10.43**	**43.53 ± 21.27**	**52.32 ± 25.33**	**0.45 ± 0.20**	**0.48 ± 0.21**

significant ($p = 0.059$). The reduction in repetitions required to maximize utility with the A/C scheme (5.67 ± 2.51) as compared to the reference classifier (6.67 ± 1.99) was also not significant ($p = 0.058$). As ALS-01 never reached the 50% character selection accuracy required for utility to be positive [35], his utility was uniformly 0 and remained unchanged by our procedure. However, ALS-02 through ALS-06 had average improvements in theoretical utility of 5.3, 1.9, 3.8, 3.7, and 3.0 bits/min, respectively. Binary classification accuracy, however, was significantly reduced ($p = 0.031$) with the A/C classifier at 83.29 ± 3.93% compared to the reference classifier at 84.55 ± 3.23% despite no significant change in precision ($p = 0.562$) and significant improvements in both recall and F-score. Specifically, the A/C classifier had a significantly ($p = 0.031$) higher recall of 52.32 ± 25.33% than the reference classifier at 43.53 ± 21.27%. The A/C classifier also provided a significantly ($p = 0.031$) higher F-score of 0.48 ± 0.21 than the reference classifier, at 0.45 ± 0.20.

When fewer repetitions were used per character, the proposed A/C characterization scheme was still observed to provide improvements in character selection accuracy as compared to the reference classifier, as shown in Figure 3. Character selection accuracy was improved by an average of 5.63% using the proposed A/C classifier as compared to the reference classifier over all numbers of repetitions and participants. Both initial selection accuracy and the extent of the improvement varied between participants. In particular, for ALS-01, character selection accuracy was improved by 3.0% on average over all possible numbers of repetitions, though this improvement did not allow for successful BCI control due to poor initial performance.

For ALS-02, character selection accuracy was improved by an average of 8.6% over all possible numbers of repetitions used. Character selection accuracy first reached an acceptable level (≥70% [40]), for ALS-02 using at least five repetitions using the reference classifier at 72.9%, as compared to three repetitions using the proposed A/C classifier, at 77.1%. Averaged over all sessions, ALS-02 achieved a maximum utility of 22.7 bit/min with the proposed A/C characterization scheme using 1–7 repetitions (mean 3.8 ± 1.8), as compared to 17.5 bit/min with the reference classifier using 3–10 repetitions (mean 5.3 ± 2.5).

For ALS-03, character selection accuracy was improved by an average of 3.1% over all numbers of repetitions, requiring at least four repetitions to reach acceptable accuracy with the reference classifier (81.0%) as opposed to three with the proposed A/C classifier (76.2%). Averaged over all sessions, ALS-03 achieved his maximum utility of 22.1 bit/min with 4–6 repetitions used (mean 4.7 ± 1.2) as compared to 20.3 bit/min using 5–6 repetitions (mean 5.7 ± 0.6).

ALS-04 never reached acceptable character selection accuracy, but the proposed classifier improved selection accuracy by an average of 9.8% over all possible numbers of repetitions. However, he achieved positive utility in several sessions, averaging 9.2 bit/min with 4–10 repetitions (mean 7.4 ± 2.6) with the A/C characterization scheme as compared to 5.4 bit/min with 6–10 repetitions (mean 8.2 ± 2.2).

For ALS-05, the average improvement in character selection accuracy over all numbers of repetitions was 3.8%, first achieving an acceptable accuracy using 3 repetitions at 75.0% with the proposed A/C characterization scheme as opposed to 4 repetitions at 74.1% character selection accuracy with the reference classifier.

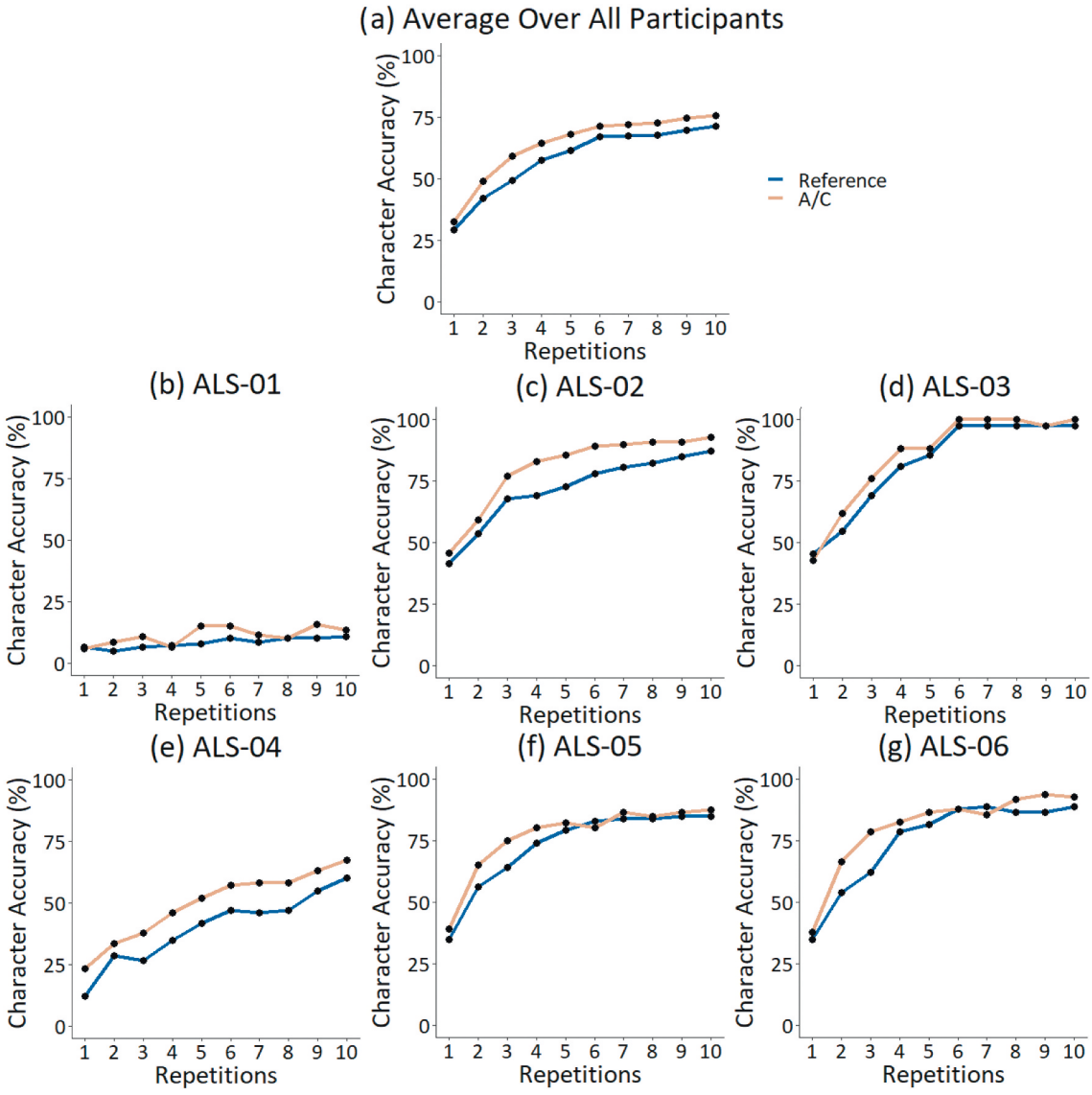

Figure 3. Average character selection accuracies at each number of repetitions used, over all participants (a) and for each participant (b-g) using both the reference (blue) and augmentation and correction (A/C; tan) classification schemes.

Averaged over all sessions, ALS-05's achieved her maximum utility with the A/C characterization scheme at 22.2 bit/min with 2–7 repetitions (mean 4.0 ± 2.0), as compared to 18.5 bit/min with 2–10 repetitions (mean 5.5 ± 3.0).

Over all possible numbers of repetitions, character selection accuracy was improved for ALS-06 by 5.4% using the A/C characterization scheme as compared to the reference classifier, first achieving an acceptable accuracy using 3 repetitions at 78.6% with the proposed A/C characterization scheme as opposed to 4 repetitions at 78.6% character selection accuracy with the reference classifier. Averaged over all sessions, ALS-06's utility was maximized at 22.6 bit/min with the A/C characterization scheme, using 3–9 repetitions (mean 4.1 ± 2.2), as compared to 19.6 bit/min using 3–10 repetitions (mean 5.3 ± 2.4) with the reference classifier.

Regardless of classification method, latency jitter was negatively associated with BCI performance. Using the proposed A/C characterization method, there were significant correlations between latency jitter and five performance metrics, specifically character accuracy (r_{rm} = −0.87, p < 0.001), utility (r_{rm} = −0.73, p < 0.001), binary classification accuracy (r_{rm} = −0.72, p < 0.001), precision (r_{rm} = −0.58, p < 0.001), and F-score (r_{rm} = −0.62, p < 0.001) indicating that as the latency jitter increased, that the proposed A/C method improved performance overall

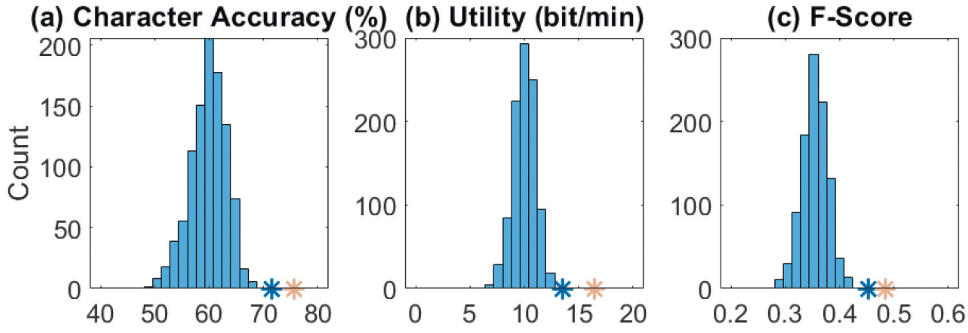

Figure 4. Random parameter sets histograms, for character accuracy (a), utility (b), and F-score (c), as well as average performance metrics over all participants using the proposed A/C characterization scheme (tan asterisks) and the reference classifier (blue asterisks).

but did not mitigate the negative relationship between jitter and performance. However, the correlation between latency jitter and recall using the proposed A/C characterization method was not significant (r_{rm} = −0.16, p = 0.320). Using the reference classifier, latency jitter correlated significantly with character selection accuracy (r_{rm} = −0.80, p < 0.001), utility (r_{rm} = −0.73. p < 0.001), binary classification accuracy (r_{rm} = −0.73, p < 0.001), precision (r_{rm} = −0.72, p < 0.001), and F-score (r_{rm} = −0.63, p < 0.001) but the negative trend was not significant for recall (r_{rm} = −0.30, p = 0.057), for significant correlations with the same five performance metrics.

The histograms of the random parameter testing for character accuracy, utility, and F-score are illustrated in Figure 4. All of randomly selected parameter sets provided all three measures lower than those achieved with our proposed A/C characterization scheme. In addition,

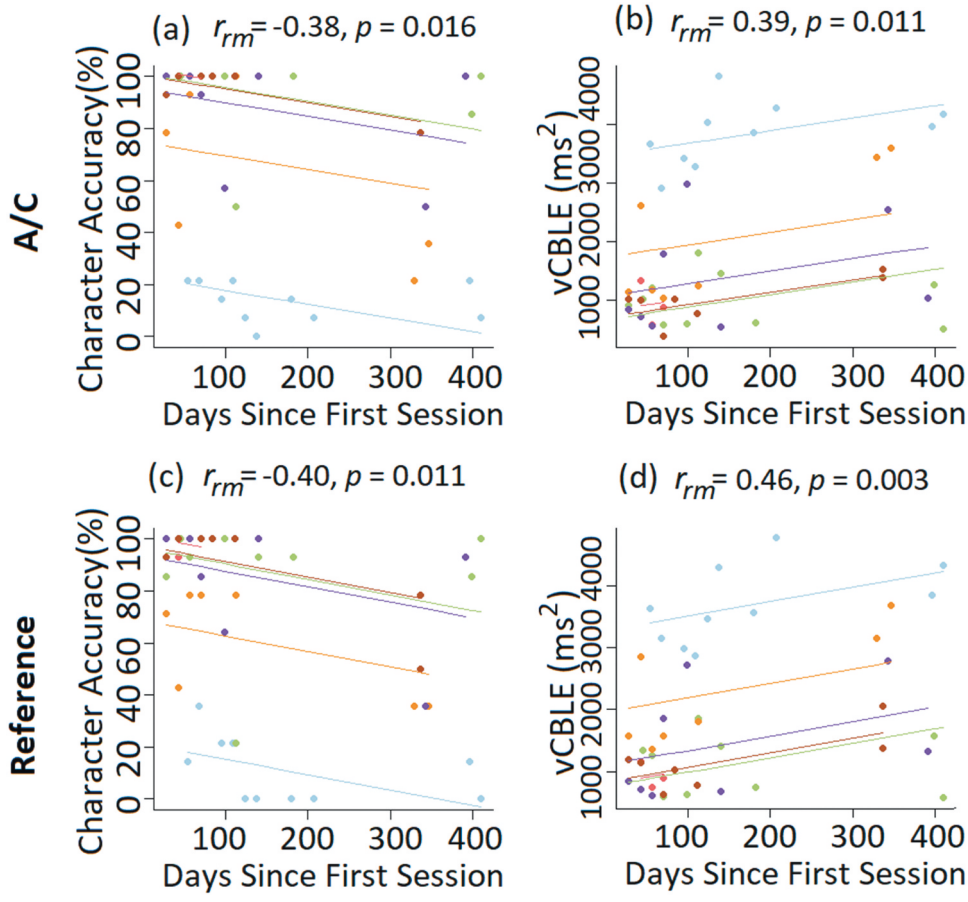

Figure 5. Longitudinal repeated measures correlation plots for character accuracy (left, a&c), and latency jitter (vCBLE, right, b&d). Note: each color indicates one participant.

all the random parameter sets provided average binary classification accuracy, precision, or recall lower than those achieved with our proposed characterization scheme. Similarly, all performance metrics except for recall, were lower than for the reference classifier for all 1000 random parameter sets. However, recall with the reference classifier was only higher than recall with 781 of the 1000 random parameter sets. These results indicate that our parameter selection algorithm provides higher performance than the use of random parameters for data augmentation and jitter correction.

3.3. Clinical and longitudinal trends

Spearman correlations between participant average character selection accuracies and clinical features, specifically age, time since diagnosis, ALSFRS-R scores, and ALSFRS-R bulbar sub-scores, were not significant for either classification method ($p > 0.05$). Spearman correlations between average performance improvements from the proposed A/C characterization scheme and clinical scores were also not significant.

Repeated measures correlation plots for the longitudinal analysis of character selection accuracy, and latency jitter over time are shown in Figure 5. Character accuracy decreased significantly over time with both the proposed A/C characterization scheme ($r_{rm} = -0.38$, $p = 0.016$) and the reference classifier ($r_{rm} = -0.40$, $p = 0.011$), indicating a decrease in performance over time using both approaches. Latency jitter increased over time with both the A/C characterization scheme ($r_{rm} = 0.39$, $p = 0.011$) and the reference classifier ($r_{rm} = 0.46$, $p = 0.003$), which aligns with both the decrease in performance over time and negative associations between jitter and performance.

Single-participant longitudinal trends in character selection accuracy are shown in Figure 6. Spearman correlations between selection accuracy and the numbers of days since the first session were significant and negative in ALS-01 for the reference classifier ($\rho = -0.65$, $p = 0.041$) but not the proposed A/C characterization scheme ($\rho = -0.46$, $p = 0.177$) indicating some longitudinal improvement in performance using our proposed A/C characterization scheme. There was no significant trend in performance over time with either the A/C scheme ($\rho = -0.04$, $p = 0.917$) or the reference classifier ($\rho = -0.04$, $p = 0.919$) for ALS-02. There was similarly no significant trend with the proposed (ρ and p both undefined) or reference ($\rho = 0.87$, $p = 0.333$) classification schemes for ALS-03, for whom

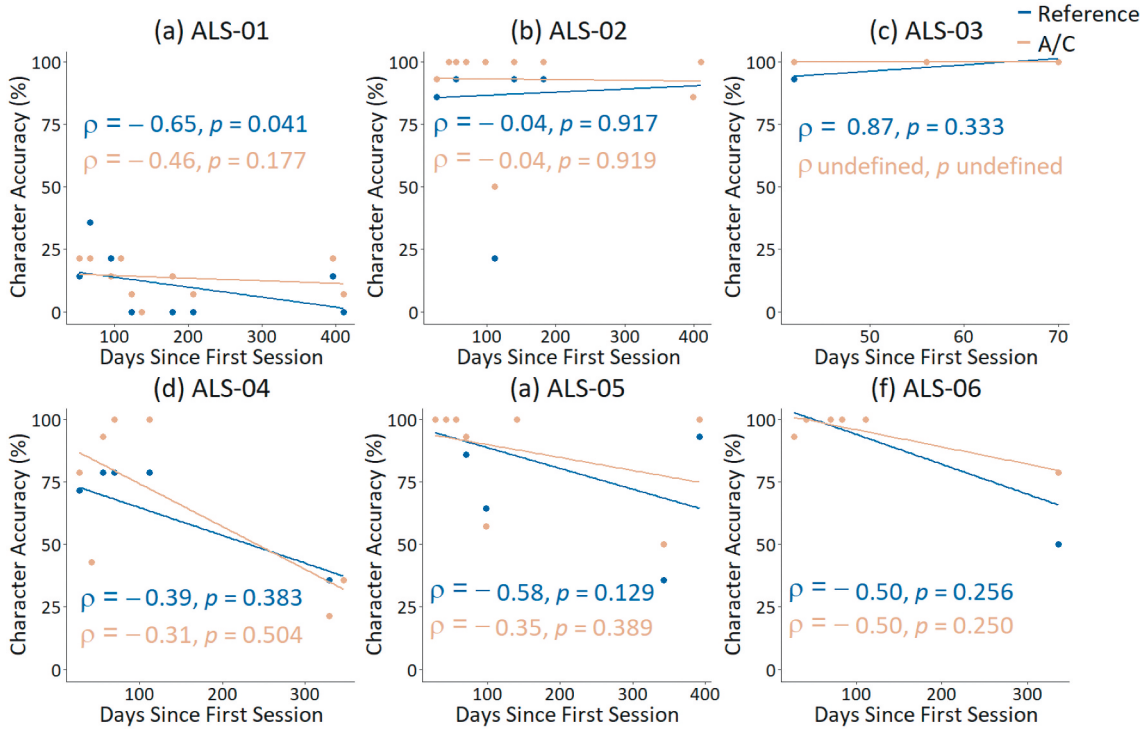

Figure 6. Longitudinal plots of single-session character selection accuracies over time for ALS-01 (a) through ALS-06 (f) using both the proposed augmentation and correction (A/C) classification (tan dots and lines) and the reference classification (blue dots and lines). Each dot represents the result from a single session. For dates where only one dot is visible, the character selection accuracies were the same with both methods.

performance metrics were only extracted from three sessions. For ALS-04, neither the correlation between accuracy with the A/C scheme ($\rho = -0.31$, $p = 0.504$) nor with the reference classifier ($\rho = -0.39$ $p = 0.383$) and time since the first session was significant. Neither negative correlation was significant for ALS-05 (A/C $\rho = -0.35$, $p = 0.389$; reference $\rho = -0.58$, $p = 0.129$) or ALS-06 (A/C $\rho = -0.50$, $p = 0.250$; reference $\rho = -0.50$, $p = 0.257$) indicating that longitudinal trends were not typically significant on the individual level, though the non-significant trends were generally toward decreasing performance for both approaches.

4. Discussion

In this study, we proposed an augmented/corrected (A/C) classification scheme that relies on latency variations at two levels, using both data augmentation and jitter correction procedures to improve P300-based BCI classification performance in people with ALS. Our proposed approach demonstrated significantly improved character selection accuracy and detection of target stimuli relative to classical reference SWLDA classifiers. Classification performance improvements with EEG data augmentation were reported to vary based on both tasks and augmentation methods in a recent review paper, though none of the papers covered by that review specifically addressed P300 tasks [27]. However, prior P300 studies have found some success with data augmentation. For example, Krell and colleagues considered multiple augmentation methods and found improvements similar to ours using a rotational data augmentation scheme with P300 data. However, their use of one consistent symmetric time-shift to augment P300 data across all neurotypical participants, did not find significant improvement, whereas we showed performance improvements with the individualized time-shifts used in our study [31]. Our proposed method, comparatively, found larger and more consistent improvements in selection accuracy than some prior augmentation approaches with P300 data, and similar improvements to one. Augmentation procedures similar to ours have been implemented in prior studies with neurotypical participants, increasing the amount of training data [31,35,36] and thereby improving performance in the two non-P300 studies [35,36]. These prior studies examined augmentation using constant time-shifts across participants, while the current study determined individual time shifts for each subject separately. Augmentation with symmetric time-shifts has also been reported to improve performance in Sakai and colleagues' study using data recorded during an intrinsic motivation task with neurotypical participants [35]. A constant but non-symmetric shift was also used to improve feedback in the detection of error-related potentials, again with neurotypical participants [36]. However, Krell and colleagues found no significant effect on performance after augmenting P300 data with symmetric time-shifts similar in size to the larger selected shifts from the current study [31]. In this study, we used individualized parameters rather than constant parameters across all sessions or all participants, as participants with ALS generally experience more latency jitter than neurotypical controls (i.e. increased within-subject variability in ALS), and as latency jitter can significantly vary between participants with ALS (i.e. between-subject variability in ALS) [4]. By individualizing the time-shifts used based on latency variations in the data, utilizing the median absolute latency shift in the training data for augmentation, we were able to both increase the amount of training data and improve performance. We also investigated changes in performance over time to evaluate how our proposed method can facilitate robust long-term use of the P300-based BCI system. While our proposed classification procedure improved performance overall, it could not completely eliminate the decline in performance over time, likely due to the inherent disease progression.

Our jitter correction procedure relying on the maximum classifier score within a given allowable range of time shifts to correct for latency variations similarly improved selection accuracy. Considering this latency variation has also shown improvement in P300 classification metrics in some prior studies [24,25], and denoising using a matrix representation of single-trials supports both effective single-trial latency detection and improved classification performance [52]. Prior investigations involving classifier-based latency estimation noted qualitatively that taking the maximum classifier score within a given range of time-shifts as our study did, increased the risk of false-positives, or detecting a P300 response for non-target stimuli, but did not quantitatively specify the size of this increase [20,24]. Rather than using this maximum score, Mowla and colleagues used a secondary classifier relying on a wavelet transform of the classifier score series to improve performance [24]. Here, by utilizing individualized parameters in the current study, we successfully improved character selection accuracy utilizing these classifier score series without a secondary classifier despite some decrease in single-trial binary classification accuracy from the aforementioned increase in false-positives [20,24], which occurs as taking the maximum score over a range increases the final score for all stimuli, including non-targets.

Our longitudinal analyses found that latency jitter increased over time, and performance accordingly decreased over time, using both the reference and proposed A/C characterization methods, though performance was improved overall with our proposed method. While participants with ALS in the completely-locked in state have not often been shown to successfully use visual BCIs [2,3,53], prior longitudinal studies which did not involve the completely locked-in state have not typically found BCI performance to decrease over time [7,9,54–56]. Several studies have, however, shown significant day-to-day variation in performance [9,54,56], which could affect investigations of long-term performance changes depending on the analysis methods used. One prior study found no change over time when comparing copy-spelling accuracies between the first and last several sessions [55]. Sellers 2010 BCI for home use study and Holz's 2015 brain painting study both used single-participant designs [7,54], while another found long-term trends to vary between participants [9]. Of our six participants, only one had a significant decline in performance over the course of the study when considered individually, two participants had consistently high performance throughout the study, and three participants appeared to have some decline in performance which did not reach significance when considered individually. It is only by considering common trends across participants with repeated measures correlations that the significant negative trend was uncovered despite both day-to-day and between-participant performance variabilities. BCIs can successfully be used for a significant period of time [7,9], but the consistent failure of current visual P300 BCIs in the completely locked-in state [3,53] indicates that performance must eventually decline, as we found to occur in our present study.

Finally, while our tests of correlations between latency jitter and performance metrics were not a key feature of the study, they confirmed prior results both in our lab [4] and in others work [20–22], namely that increased latency jitter is associated with decreased BCI performance. A classification method that can reduce or eliminate this association, if possible, would likely make BCI performance more robust. However, our proposed method retained this association while improving performance overall.

4.1. Limitations and future work

One limitation of this study, common to many BCI studies of people with ALS, is the relatively low number of participants, due in part both to the rareness of the disease and the difficulties of recording from this population. We therefore did not analyze differences due to gender, though we did consider clinical features in some analyses. The longitudinal recordings we obtained from each participant, however, provide additional data points, mitigating some limitations related to small sample sizes. While we report the average results from a small number of participants, the proposed A/C characterization method was tested on several sessions of longitudinal recordings for each participant, making these participant averages more robust. We additionally used nonparametric statistical methods appropriate to small sample sizes and data which are not normally distributed. For the longitudinal investigation, our use of repeated measures correlations, rather than separately investigating long-term trends for each participant, increased power while maintaining statistical rigor [49]. Future work could also include additional participants and recording sessions.

One potential confound when investigating longitudinal trends is the variation in how much data is used to train the classifier in our proposed A/C scheme, either three or four times the original un-augmented training set. However, the same trends were present with our proposed A/C scheme and the reference classifier with neither data augmentation nor jitter correction. We therefore conclude the longitudinal trends we found are not due to this variation, which was not a factor with the reference classifier.

Another limitation to the current study is inherent to CBLE, which defines a single latency shift for the entire spatiotemporal ERP complex for each stimulus [20,21]. While Thompson's tests with simulated data show the efficacy of CBLE in reflecting P300 latency jitter [57], future work could investigate latency variations between different ERP components. Another facet of CBLE and its use which could be investigated in the future is step size. In our current study, as in Thompson's work [20], epochs were shifted in steps of one sample, allowing the detection of very small latency shifts. Increasing the step size, thereby reducing computational requirements, may be possible without sacrificing the performance improvements yielded by our A/C scheme. Future work could investigate the effect of step size on performance improvements and/or optimize step size for individual BCI users.

Our analyses, while conducted offline, were designed to be appropriate for real-life settings, with all training and parameter selection procedures relying only on data from prior sessions. This would be especially important as practical environments would likely utilize information from prior sessions and/or a short amount of data from the same session to successfully implement in any upcoming BCI experiment. The current study considers

jitter in a simple way relying on individualized parameters to ensure efficacy, and so future work could include the real-time implementation of our proposed A/C method.

5. Conclusion

In this work, we proposed an augmented/corrected (A/C) classification procedure using both data augmentation and jitter correction schemes to improve P300-based BCI classification performance in people with ALS. The proposed method demonstrated an improvement in selection accuracy overall which did not show any relationships with clinical features. Considering common trends across participants, the current work showed decreased BCI performance over time, which was suggested by BCI inefficiency in the completely locked-in state but not consistently demonstrated in the past. When participants were considered individually, however, longitudinal performance trends varied and did not consistently show decreases, which fits with prior studies. Despite improving selection accuracy, our proposed method did not fully eliminate the common downward trend in performance over time.

Acknowledgments

This study was supported by the National Science Foundation (NSF-1913492) and the Institutional Development Award (IDeA) Network for Biomedical Research Excellence (P20GM103430). The authors would like to thank the participants who took part in this study, without whom this study would not have been possible. We would also like to thank the ALS Association Rhode Island Chapter and the National Center for Adaptive Neurotechnologies for their continuous support.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [NSF-1913492]; Institutional Development Award (IDeA) Network for Biomedical Research Excellence [P20GM103430].

ORCID

Alyssa Hillary Zisk http://orcid.org/0000-0003-2266-4855
Seyyed Bahram Borgheai http://orcid.org/0000-0002-3164-6440
John McLinden http://orcid.org/0000-0002-1475-575X
Roohollah Jafari Deligani http://orcid.org/0000-0002-0716-4388
Yalda Shahriari http://orcid.org/0000-0003-1640-0749

References

[1] Gómez-Vilda P, Londral ARM, Rodellar-Biarge V, *et al.* Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production. Neurocomputing. 2015;151:130–138.
[2] Birbaumer N, Piccione F, Silvoni S, et al. Ideomotor silence: the case of complete paralysis and brain–computer interfaces (BCI). Psychol Res. 2012;76 (2):183–191.
[3] Kübler A, Birbaumer N. Brain–computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? Clin Neurophysiol. 2008;119(11):2658–2666.
[4] Zisk AH, Borgheai SB, McLinden J, et al. P300 Latency jitter and its correlates in people with amyotrophic lateral sclerosis. Clin Neurophysiol. 2021;132 (2):632–642.
[5] McCane LM, Sellers EW, Mcfarland DJ, et al. Brain-computer interface (BCI) evaluation in people with amyotrophic lateral sclerosis. Amyotrophic Lat Scler Frontotemporal Degener. 2014;15(3–4):207–215.
[6] Geronimo AM, Simmons Z. The P300 'face' speller is resistant to cognitive decline in ALS. Brain-Comput Interfaces. 2017;4(4):225–235.
[7] Sellers EW, Vaughan TM, Wolpaw JR. A brain-computer interface for long-term independent home use. Amyotrophic Lat Scler. 2010;11 (5):449–455.
[8] Shahriari Y, Sellers EW, McCane LM, et al., "Directional brain functional interaction analysis in patients with amyotrophic lateral sclerosis," in *2015 7th International IEEE/EMBS Conference on Neural Eng. (NER)* Montpellier, France, 2015, pp. 972–975. 10.1109/NER.2015.7146788.
[9] Shahriari Y, Vaughan TM, McCane L, et al. An exploration of BCI performance variations in people with amyotrophic lateral sclerosis using longitudinal EEG data. J Neural Eng. 2019;16(5):056031.
[10] Shahriari Y, Erfanian A. Improving the performance of P300-based brain–computer interface through subspace-based filtering. Neurocomputing. 2013;121:434–441.
[11] Wang J, Gu Z, Yu Z, *et al.* An online semi-supervised P300 speller based on extreme learning machine. Neurocomputing. 2017;269:148–151.
[12] Alvarado-Gonzalez M, Fuentes-Pineda G, Cervantes-Ojeda J. A few filters are enough: convolutional neural network for P300 detection. Neurocomputing. 2021;425:37–52.
[13] Bayliss JD, Ballard DH. Single trial P3 epoch recognition in a virtual environment. Neurocomputing. 2000;32:637–642.
[14] Allison BZ, Kübler A, Jin J. 30+ years of P300 brain–computer interfaces. Psychophysiology. 2020;57(7): e13569.

[15] Speier W, Chandravadia N, Roberts D, et al. Online BCI typing using language model classifiers by ALS patients in their homes. Brain-Comput Interfaces. 2017;4(1–2):114–121.
[16] Polich J. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol. 2007;118(10):2128–2148.
[17] Ramadan RA, Vasilakos AV. Brain computer interface: control signals review. Neurocomputing. 2017;223:26–44.
[18] Mak JN, McFarland DJ, Vaughan TM, et al. EEG correlates of P300-based brain–computer interface (BCI) performance in people with amyotrophic lateral sclerosis. J Neural Eng. 2012;9(2):026014.
[19] Geronimo A, Simmons Z, Schiff SJ. Performance predictors of brain–computer interfaces in patients with amyotrophic lateral sclerosis. J Neural Eng. 2016;13(2):026002.
[20] Thompson DE, Warschausky S, Huggins JE. Classifier-based latency estimation: a novel way to estimate and predict BCI accuracy. J Neural Eng. 2012;10(1):016006.
[21] Mowla MR, Gonzalez-Morales JD, Rico-Martinez J, et al. A comparison of classification techniques to predict brain-computer interfaces accuracy using classifier-based latency estimation. Brain Sci. 2020;10(10):734.
[22] Aricò P, Aloise F, Schettini F, et al. Influence of P300 latency jitter on event related potential-based brain–computer interface performance. J Neural Eng. 2014;11(3):035008.
[23] Vucic S. P300 jitter latency, brain-computer interface and amyotrophic lateral sclerosis. Clin Neurophysiol. 2021;132(2):614–615.
[24] Mowla MR, Huggins JE, Thompson DE. Enhancing P300-BCI performance using latency estimation. Brain-Comput Interfaces. 2017;4(3):137–145.
[25] Togashi R, Washizawa Y. Feature extraction of P300 signal using Bayesian delay time estimation. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Kaohsiung, Taiwan, 2013, pp. 1–5. 10.1109/APSIPA.2013.6694191.
[26] Iturrate I, Chavarriaga R, Montesano L, et al. Latency correction of event-related potentials between different experimental protocols. J Neural Eng. 2014;11(3):036005.
[27] Lashgari E, Liang D, Maoz U. Data augmentation for deep-learning-based electroencephalography. J Neurosci Methods. 2020;346:108885.
[28] Abdelfattah SM, Abdelrahman GM, Wang M. Augmenting the size of EEG datasets using generative adversarial networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1–6. 10.1109/IJCNN.2018.8489727.
[29] Riyad M, Khalil M, Abdellah A. MI-EEGNET: a novel Convolutional Neural Network for motor imagery classification. J Neurosci Methods. 2021;353:109037.
[30] Krell MM, Kim SK. Rotational data augmentation for electroencephalographic data. In *2017* 39th Annual International Conference of the IEEE Engineering in Medicine *and Biology* Society *(EMBC)* Jeju, Korea (South), 2017, pp. 471–474. 10.1109/EMBC.2017.8036864.
[31] Krell MM, Seeland A, Kim SK Data augmentation for brain-computer interfaces: analysis on event-related potentials data . Jan . 2018 https://arxiv.org/abs/1801.02730. *arXiv Preprint arXiv:1801.02730*.
[32] Yin Z, Zhang J. Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights. Neurocomputing. 2017;260:349–366.
[33] Wan Z, Yang R, Huang M, *et al*. A review on transfer learning in EEG signal analysis. Neurocomputing. 2021;421:1–14.
[34] Lotte F, Bougrain L, Cichocki A, et al. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J Neural Eng. 2018;15(3):031005.
[35] Sakai A, Minoda Y, Morikawa K. Data augmentation methods for machine-learning-based classification of bio-signals. In 2017 *10th Biomedical Engineering International Conference (BMEiCON*) Hokkaido, Japan, pp. 1–4. 10.1109/BMEiCON.2017.8229109.
[36] Kim SK, Kirchner EA, Stefes A, et al. Intrinsic interactive reinforcement learning–Using error-related potentials for real world human-robot interaction. Sci Rep. 2017;7(1):1–16.
[37] Wang X, Wang X, Liu W, *et al*. One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial EEG. Neurocomputing. 2021;459:212–222.
[38] Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS study group (Phase III). J Neurol Sci. 1999;169(1–2):13–21.
[39] Farwell LA, Donchin E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr Clin Neurophysiol. 1988;70(6):510–523.
[40] Schalk G, Mellinger J. A practical guide to brain–computer interfacing with bci2000: general-purpose software for brain-computer interface research, data acquisition, stimulus presentation, and brain monitoring. New York: Springer Science & Business Media; 2010.
[41] Krusienski DJ, Sellers EW, Cabestaing F, et al. A comparison of classification techniques for the P300 Speller. J Neural Eng. 2006;3(4):299.
[42] Krusienski DJ, Sellers EW, McFarland DJ, et al. Toward enhanced P300 speller performance. J Neurosci Methods. 2008;167(1):15–21.
[43] Nasiri S, Zahedi G, Kuntz S, *et al*. Knowledge representation and management based on an ontological CBR system for dementia caregiving. Neurocomputing. 2019;350:181–194.
[44] Michiels M, Larrañaga P, Bielza C. BayeSuites: an open web framework for massive Bayesian networks focused on neuroscience. Neurocomputing. 2021;428:166–181.
[45] Dal Seno B, Matteucci M, Mainardi LT. The utility metric: a novel method to assess the overall performance of discrete brain–computer interfaces. IEEE Transactions on Neural Systems Rehabilitation Engineering. 2010;18(1):20–28.

[46] Mowla MR, Huggins JE, Thompson DE. Evaluation and performance assessment of the brain–computer interface system Nam, CS, Nijholt, A, and Lotte, F. In: Brain–Computer Interfaces Handbook. New York: CRC Press; 2018. 635–650. https://www.taylorfrancis.com/chapters/edit/10.1201/9781351231954-33/

[47] Team RC. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2021;4.0.5.

[48] Wilcoxon F. Individual comparisons by ranking methods.Biom Bull. 1945;1(6):80–83.

[49] Bakdash JZ, Marusich LR. Repeated measures correlation. Front Psychol. 2017;8:456.

[50] Metropolis N, Ulam S. The monte carlo method. J Am Stat Assoc. 1949;44(247):335–341.

[51] Kroese DP, Chan JC. Monte carlo sampling. Statistical Modeling Computation 2014, Springer.;195–226. DOI:10.1007/978-1-4614-8775-3_7.

[52] Huang Y, Chen X, Zhang J, *et al*. Single-trial ERPs denoising via collaborative filtering on ERPs images. Neurocomputing. 2015;149:914–923.

[53] Murguialday AR, Hill J, Bensch M, et al. Transition from the locked in to the completely locked-in state: a physiological analysis. Clin Neurophysiol. 2011;122(5):925–933.

[54] Holz EM, Botrel L, Kaufmann T, et al. Long-term independent brain-computer interface home use improves quality of life of a patient in the locked-in state: a case study. Arch Phys Med Rehabil. 2015;96(3):S16–S26.

[55] Wolpaw JR, Bedlack RS, Reda DJ, et al. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. Neurology. 2018;91(3):e258–e267.

[56] Borgheai SB, McLinden J, Zisk AH, et al. Enhancing communication for people in late-stage ALS using an fNIRS-based BCI system. IEEE Trans Neural Syst Rehabil Eng. 2020;28(5):1198–1207 doi:10.1109/TNSRE.2020.2980772 .

[57] Thompson DE, Mowla MR, Huggins JE. Evidence of latency variation in the P3 speller brain computer interface. *In Soc. Neurosci*., October 23, 2019. https://www.abstractsonline.com/pp8/#!/7883/presentation/69619

Appendix A: Selected Parameters

Table A.1 shows the selected parameters for each participant and test session, with the two associated training sessions noted for reference. The parameters for both data augmentation and jitter correction are those determined as in section 2.4.3. Symmetric time-shifts for data augmentation were between 7 and 55 ms for participants other than ALS-01, with a maximum of 58.59 ms selected over one pair of training sessions for ALS-01. Correction windows were between 7 and 55 ms for participants other than ALS-01, with a maximum of 101.56 ms selected over two pairs of training sessions for ALS-01. Per-epoch augmentation was used for at least one pair of training sessions for each participant, and for all pairs of training sessions for ALS-01 and ALS-04. The individualized parameters vary both between participants and between training sets.

Table A.1. Parameter selections for each test session, organized by participant.

Session		Participant								
Training Sessions	Test Session	ALS01			ALS02			ALS03		
		Augmentation Shift	Per-epoch Augmentation	Correction Range	Augmentation Shift	Per-epoch Augmentation	Correction Range	Augmentation Shift	Per-epoch Augmentation	Correction Range
1–2	3	±19.53 ms	Yes	±23.44 ms	±19.53 ms	Yes	±31.25 ms	±15.63 ms	No	±11.72 ms
2–3	4	±31.25 ms	Yes	±46.88 ms	±15.63 ms	Yes	±23.44 ms	±19.53 ms	Yes	±19.53 ms
3–4	5	±39.06 ms	Yes	±101.56 ms	±15.63 ms	No	±11.72 ms	±15.63 ms	No	±7.81 ms
4–5	6	±35.16 ms	Yes	±89.84 ms	±15.63 ms	No	±7.81 ms	-	-	-
5–6	7	±35.16 ms	Yes	±101.56 ms	±11.72 ms	No	±11.72 ms	-	-	-
6–7	8	±39.06 ms	Yes	±39.06 ms	±11.72 ms	No	±15.63 ms	-	-	-
7–8	9	±50.78 ms	Yes	±0 ms	±23.44 ms	Yes	±39.06 ms	-	-	-
8–9	10	±58.59 ms	Yes	±0 ms	±19.53 ms	Yes	±54.68 ms	-	-	-
9–10	11	±54.69 ms	Yes	±0 ms	±19.53 ms	Yes	±27.34 ms	-	-	-
10–11	12	±46.88 ms	Yes	±0 ms	±15.63 ms	No	±11.72 ms	-	-	-
Mean ± STD		**41.02 ± 11.83 ms**	-	**40.23 ± 43.19 ms**	**16.80 ± 3.71 ms**	-	**23.44 ± 14.96 ms**	**16.93 ± 2.26 ms**	-	**13.02 ± 5.97 ms**

Session		Participant								
Training Sessions	Test Session	ALS04			ALS05			ALS06		
		Augmentation Shift	Per-epoch Augmentation	Correction Range	Augmentation Shift	Per-epoch Augmentation	Correction Range	Augmentation Shift	Per-epoch Augmentation	Correction Range
1–2	3	±15.63 ms	Yes	±27.34 ms	±11.72 ms	No	±11.72 ms	±27.34 ms	Yes	±42.97 ms
2–3	4	±19.53 ms	Yes	±31.25 ms	±11.72 ms	No	±19.53 ms	±23.44 ms	Yes	±39.06 ms
3–4	5	±19.53 ms	Yes	±46.88 ms	±15.63 ms	No	±7.81 ms	±19.53 ms	Yes	±27.34 ms
4–5	6	±23.44 ms	Yes	±39.06 ms	±15.63 ms	No	±7.81 ms	±15.63 ms	No	±15.63 ms
5–6	7	±15.63 ms	Yes	±39.06 ms	±15.63 ms	No	±7.81 ms	±15.63 ms	No	±7.81 ms
6–7	8	±19.53 ms	Yes	±35.16 ms	±15.63 ms	No	±11.72 ms	±19.53 ms	Yes	±50.78 ms
7–8	9	±27.34 ms	Yes	±39.06 ms	±15.63 ms	No	±7.81 ms	±15.63 ms	No	±7.81 ms
8–9	10	-	-	-	±19.53 ms	Yes	±27.34 ms	-	-	-
Mean ± STD		**20.09 ± 4.18 ms**	-	**36.83 ± 6.32 ms**	**15.14 ± 2.50 ms**	-	**12.70 ± 7.16 ms**	**19.53 ± 4.51 ms**	-	**27.34 ± 17.47 ms**