

Brain-computer interface performance development over 30 hours of patient training with an auditory event-related potential protocol

Michael Tangermann¹, Alexander de Ranitz^{1*}, Soz Raheem^{1*}, Mariacristina Musso²

¹Donders Institute, Radboud University, Nijmegen, the Netherlands

²German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

*These authors contributed equally

E-mail: michael.tangermann@donders.ru.nl

ABSTRACT: Neural activity measured with electroencephalography can be decoded into control commands using a brain-computer interface (BCI). Longitudinal studies on BCI training are scarce. While some studies examined long-term training on motor imagery (MI) BCIs, multi-session user training using event-related potential (ERP) protocols remains largely unexplored, despite its relevance for novel cognitive rehabilitation protocols. *Methods:* In a post-hoc analysis, we investigated the ERP decoding performance in ten patients with stroke-induced aphasia, who completed a 30-hour multi-session feedback-based language training. Performance was assessed using an adaptive block-Toeplitz LDA classifier and quantified by the area under the ROC curve. *Results:* Decoding performance varied across patients and evolved over time. Several patients improved across sessions, indicating learning, while others showed stable performance. Increases in task difficulty caused sudden performance drops, often followed by recovery and gradual improvement. *Conclusion:* These findings provide insights into the longitudinal dynamics of ERP-based BCI training and highlight its potential for BCI-supported rehabilitation beyond MI paradigms.

INTRODUCTION

Brain signals such as the electroencephalogram (EEG) can be analyzed with machine learning methods and translated into control commands by a brain-computer interface (BCI) system. Such systems enable applications including device control, communication interfaces, mental state monitoring, or rehabilitation after stroke. Decoding information from brain signals is a central component of a BCI system. Decoding performance can be improved, not only with the availability of more data, but also with increased signal quality as a result of *user learning*. The latter means that the user learns to modulate their brain activity, such that more discriminative features can be extracted by the BCI, enabling improved decoding accuracy.

However, decoding performance can also deteriorate due to, e.g., disease progression or impedance problems in long-term implantations, both of which are observed in patients, for whom BCI systems are primarily designed.

Several studies report on the long-term *usage* of a BCI application for communication and control [1–4] or for artistic expression of patients [5]. As the goal of these is to provide a usable application, a recalibration of the BCI is typically avoided once the decoding performance reaches reliable control of the application. Furthermore, an explicit high-density, repeated, supervised *training* of the user, with the goal to further increase the decodability of the brain signals, is not performed, as the time investment for this *training* would leave the patient with less time for *using* the actual application. Nevertheless, user learning may be a side-effect of the continuous usage of the BCI application.

Contrary to these studies with a focus on long-term stability, longitudinal *changes* of brain signals are an explicit goal of (typically non-invasive) studies that employ BCIs to support neuro-rehabilitation [6]. They aim to support the patient to use and strengthen the remaining nodes and connections of, e.g., a stroke-lesioned brain network to overcome functional deficits. In motor rehabilitation protocols, this seems feasible by closing the loop between the (attempted) motor execution and sensory feedback, where the BCI has the role to detect the execution of the motor task. Therefore, it is beneficial if the patient learns to generate increasingly informative oscillatory band power changes over time, such that these can be decoded by the BCI as clearly as possible. As a result, the BCI can provide more informative feedback. For event-related potential (ERP) protocols, however, such longitudinal BCI learning scenarios are scarce. In [7], ten healthy participants used a P300 speller over 11 sessions. Half of the participants received only auditory stimuli and the other half a mixture of auditory and visual stimuli. The classifier was retrained before each session using data from previous sessions. For the auditory stimuli, accuracy improved notably over time, possibly indicating user learning. However, this improvement could in part be due to the classifier being trained on an increasing amount of data over time.

In [8], 16 healthy participants used an auditory P300 speller application over five sessions. The participants could see the letter selected by the BCI application, providing a form of feedback. The classifier was retrained

every session using calibration data collected at the start of the session. Their results showed a significant increase of the information transfer rate (ITR) over time. The authors stabilized the decoding accuracy over sessions by reducing the number of repetitions in a trial for participants with a high decoding accuracy, ensuring the task remained challenging. The increased ITR thus can be explained by the reduced time required per trial and indirectly reflects an improvement of the decoding performance.

While these studies indicate that longitudinal learning may also occur in ERP protocols, several questions remain open. Primarily, since these studies consist only of healthy participants, it is unclear whether learning also occurs in patients, the primary target group for most BCI systems. In the case of [7], it is difficult to discern whether learning has taken place as their experimental procedure and analysis were not focused on *user learning*. Additionally, no individual adaptation of the task difficulty was performed, possibly reducing learning pressure and resulting in some participants reaching a ceiling performance.

In this work, we address this gap by analyzing long-term training effects in an extended auditory ERP-BCI training procedure for patients with stroke-induced aphasia. For this purpose, we re-analyze a longitudinal ERP dataset collected in a study on BCI-assisted language rehabilitation [9], where the task difficulty was continuously adapted to keep the patients challenged.

MATERIALS AND METHODS

Dataset: The EEG data analyzed for this study has been recorded using a language rehabilitation protocol [9]. Each of ten patients (aged 38-76 years) with different levels of stroke-induced aphasia participated in approximately 30 hours of supervised auditory ERP-BCI feedback training. The intensive training used word stimuli to elicit ERP responses, spanned over multiple sessions, and took approx. four weeks per patient.

Sessions: Each patient started with two sessions to collect offline calibration data. In the original rehabilitation protocol [9], this data was used to calibrate the initial classifier. The final session was also an offline session and had served to study any pre-/post training differences. The intermediate sessions were online sessions, in which a trained classifier delivered feedback to the patient after each trial. In our analysis, only the data of the online sessions is considered, resulting in a total of 149 analysed sessions. The amount of sessions differed between patients, ranging from 14 to 28 sessions. However, the total amount of training sums up to 30 hours per patient, with the exception of P2, who underwent 22 hours of training only. Therefore, a patient with a large number of sessions on average had a shorter session duration than a patient with fewer sessions. As a consequence, the number of word presentations ranged between 2940 and 15144 per session. This variability can be clearly seen in Figure 1,

where the amount of recorded epochs, i.e. words presented, is displayed per session.

Experimental task: The online training consisted of a modified AMUSE protocol [10], using six word stimuli to repeatedly probe the patient's ability to detect repeated occurrences of a target word in a sequence of non-target words. No specific instructions on how to attend to targets or ignore non-targets were given. In most of the *trials*, the patient heard a pseudo-random sequence of 90 word stimuli in total, consisting of 75 non-target words and 15 target words. The EEG signal describing the response evoked by a single word stimulus is referred to as an *epoch*. In the original experiment, an adaptive classifier had been trained to estimate the class label (target or non-target) of each stimulus word based on the ERP information contained in the corresponding epoch.

Feedback: Patients received frequent feedback about the classification during the online training sessions. At the end of each trial, graded audiovisual feedback informed the patient whether the target word was decoded correctly. Especially positive feedback was provided in case dynamic stopping occurred, see [9]. It was hypothesized that providing feedback may help patients to find an individual strategy for word processing, which in turn may increase the positive, reinforcing feedback.

Experimental conditions: Multiple experimental conditions were used in this language training protocol. The conditions varied w.r.t. spatial presentation and stimulus onset asynchrony (SOA). The spatial condition determined whether words were presented to the patient through six loudspeakers surrounding the patient's head (6D), or through a mono stream delivered via headphones (HP). The latter made the task more difficult, as spatial information was removed. The SOA defined the time between the onsets of two consecutive word presentations. For most conditions, it was either 250 ms or 350 ms, but rarely reached up to 1000 ms. Depending on the classification performance of the first two offline sessions, the first online session started with the *6D 250* or the *6D 350* condition.

Changing the task difficulty:

Between sessions and, on a few occasions, within sessions, the experimenters changed the task difficulty with the goal to challenge, but not to overburden the patient. The most important change involved making the task more difficult by switching from the 6D condition to the HP condition, or by shortening the SOA. In addition, single words that had shown very high decoding accuracy on two consecutive sessions when they were in the target role were replaced between sessions. The training intensity was also maintained by using dynamic stopping of ongoing trials [11] that showed a clear decoding of the target word. As a consequence, trials were shorter and more trials—each resulting in feedback to the patient—could be executed in the same amount of time. As, according to the patients, the condition changes had the strongest impact on the task difficulty, we have included this factor into our analysis.

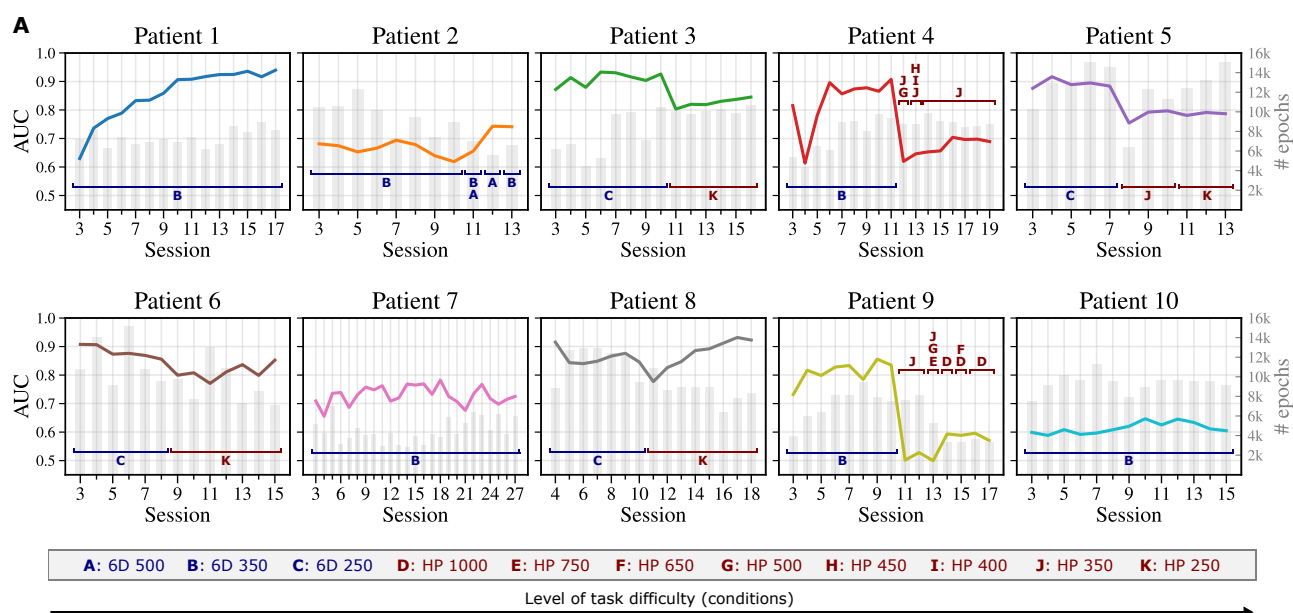


Figure 1: **Individual session-wise learning curves.** Colored lines provide the ROC-AUC 5-fold cross-validation performance per patient and session. The number of sessions differs between patients. Gray bars indicate the number of epochs recorded per session. For condition details please see the main text. Conditions marked in red are substantially more difficult than those in blue.

Data pre-processing: The raw 31-channel EEG data was band-pass filtered between 0.5 Hz and 16 Hz and cut into epochs around each stimulus event. Per epoch and channel, the average amplitude of 18 consecutive time intervals was extracted. The intervals had a width of 50 ms, starting at 100 ms post stimulus onset, and ending 1000 ms after stimulus onset. This resulted in 31×18 average microvolt values, which were flattened into a feature vector of shape 1×558 .

Epochs were rated outliers and removed when exceeding 25 inter-quartile ranges (IQR) based on the absolute magnitude of each feature vector. The procedure was applied independently for each subject and resulted in the removal of 738 epochs in total.

Classification: To track the performance development over time, we deployed a binary classifier. Based on the feature vector of an epoch, it classified the corresponding word as either a target (attended by the patient) or non-target (ignored). To generate a trial-level prediction, i.e., which out of the six words was attended, these epoch-level predictions can be combined at the end of the stimulus sequence. For the purposes of this paper, however, we only consider the epoch-level binary predictions in our post-hoc analysis. These were delivered by the recently introduced block-Toeplitz linear discriminant analysis (btLDA) [12]. It efficiently regularizes the covariance estimation in ERP protocols, and is advantageous specifically with a larger number of temporal features or limited training data.

Evaluation of decoding performance: Per patient, we estimated the classification performance across the training duration in two ways: In the *session-wise* analysis, each session was treated as a separate dataset. This analysis may suffer from performance variances caused by

the varying amount of words, and thus data points, available per session. In the *windows-based* analysis, the on-line data of a patient was cut into overlapping windows of 5000 epochs each. Windows moved with a stride of 1000 epochs. This analysis did not respect the session structure, but instead ensured that each classifier was trained on an equal amount of data. This allows for a meaningful comparison between windows across time. Each window of 5000 epochs was subsequently treated as a single dataset to train and evaluate the classifier on.

For both analyses, we report the area under the receiver-operating characteristic curve (ROC-AUC) of the binary epoch-level classification based on a 5-fold cross-validation, as the ROC-AUC is insensitive to class imbalances and classifier bias. It should be noted that both analyses are adaptive over time, but are not strictly causal due to the within-session or within-window cross-validation procedure.

RESULTS

We present the results in three views, which differ in how the segments of the dataset have been determined, for which the AUC score is computed:

Learning curves over sessions: Figure 1 shows the learning curves for each patient for the *session-based analysis*. The amount of data points contained in each session differs between and within patients, as indicated by the lengths of the vertical gray bars in Figure 1. To check whether sessions with more epochs lead to better classification, a correlation test was conducted. We observed a significant ($p = 0.002$) but weak correlation of $r = 0.26$ between the session size (in epochs) and the AUC obtained. However, experimental condition may act

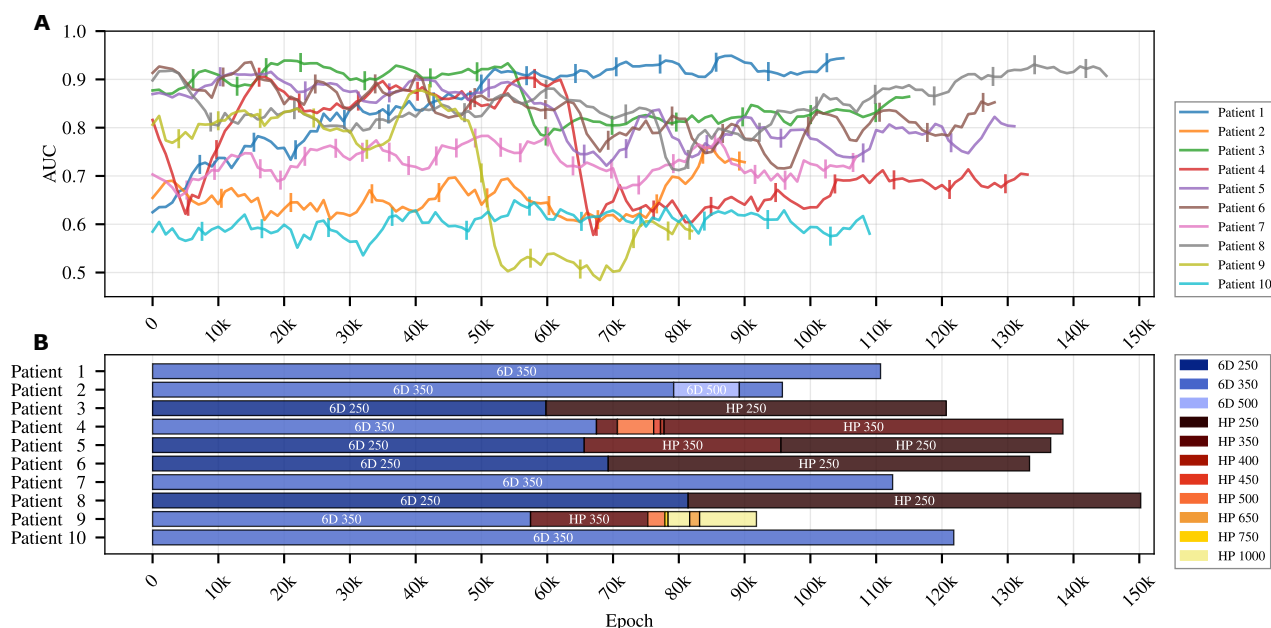


Figure 2: **Windows-based performance curves.** (A) Five-fold cross-validation ROC-AUC for the moving window starting at that epoch. Per curve, vertical bars indicate the session boundaries. (B) Overview of experimental conditions used for each patient across all epochs. ‘6D’ refers to the stereo setup with 6 speakers surrounding the patient, and ‘HP’ refers to the headphones condition. The subsequent number indicates the SOA in milliseconds. Please note that due to the data windowing in the training and validation process, the curves in the upper plot end between 5000 and 5999 epochs earlier than the corresponding bar in the lower plot.

as a confounding variable. For example, shorter SOAs are often used with high-performing patients, and sessions with shorter SOAs tend to have more epochs.

Learning curves over constant window sizes: Figure 2 shows the learning curves for each patient over time for the *window-based* analysis (top), and indicates which experimental conditions have been used (bottom). Unlike Figure 1, per patient a classifier slides with a window size of 5000 epochs over the data and treats each *window* as a separate dataset, eliminating the effect that varying dataset sizes across sessions may have on the classification performance. However, this approach does not respect session boundaries, which can result in windows that contain data from multiple sessions.

Learning curves over constant experimental conditions: Figure 3 shows the performance development relative to the first window, and separately for data *segments* from consecutive sessions (of a single patient) that share the same experimental condition. Only segments containing at least two sessions and 20,000 epochs are considered. We observed positive slopes for the HP conditions (right plot), while for the 6D conditions (left plot) no dominating slope direction could be identified, as depicted in the corresponding histograms.

DISCUSSION

We have derived the BCI performance curves for data obtained from stroke patients who received a longitudinal ERP-based language training protocol. To our knowledge, this data is unique in that task difficulty was continuously adjusted to remain challenging throughout the 30

hours of training. This design characteristic of the original study may influence the shape of learning curves compared to a mere long-term *usage* of an ERP application—a well-working application would not be made more difficult on purpose.

In the following, we use the term *learning* to denote changes in classification performances over time. Please note that the clinically targeted language improvement has been discussed in [9] and will not be regarded here.

Overall learning effects: On the group level, Musso and colleagues reported training induced stronger ERP components and reduced latencies [9]. Not unexpected due to their different lesions and stroke-induced deficits (see [9]), our offline post-hoc analysis revealed substantial differences between patients, as presented in the learning curves. The data of patient P1 exemplifies a continuously improved binary classification performance, starting with an AUC of just over 0.6 and reaching an AUC of more than 0.9 after 30 hours of training. As a clear learning curve is observed not only in the session-wise analysis, but also using the moving window classifier in Figure 2, P1’s improvement cannot solely be caused by larger training data sets per session, but more likely represents a proof of concept that learning in this patient group is feasible. On the other end of the spectrum is the learning curve of patient P10. It shows a relatively stable performance over 30 hours without a major improvement in classification.

Influence of conditions: We found that learning effects can be minimal in one experimental condition, but large in another, as exemplified by patient 8: While P8’s performance in the 6D 250 condition remained rather sta-

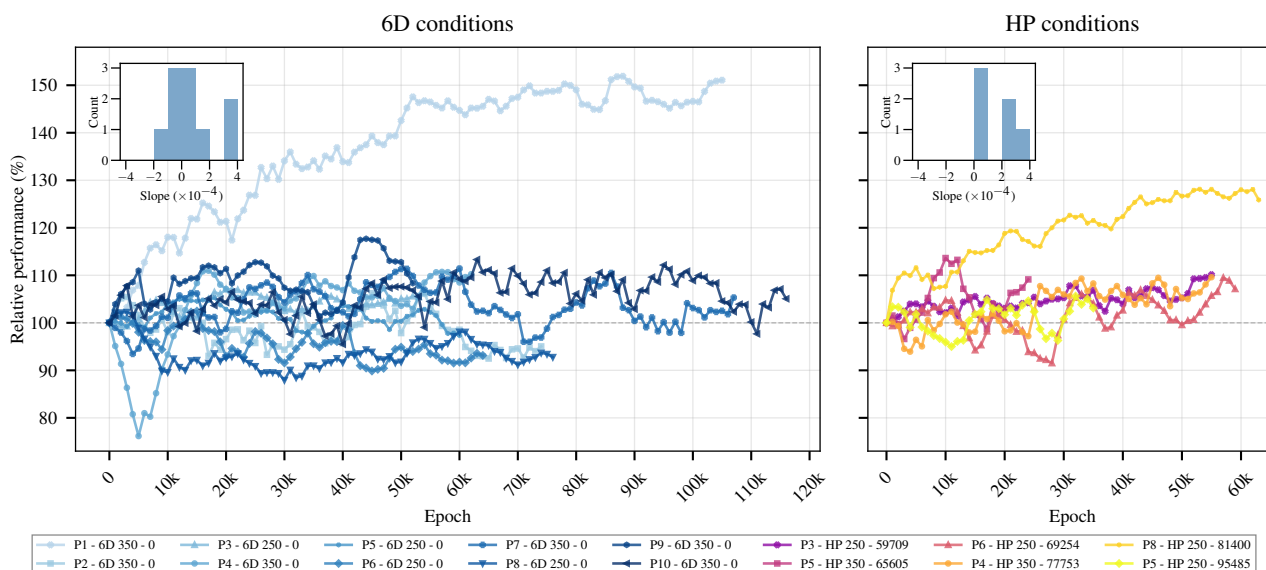


Figure 3: **Relative performance curves for data segments of stable conditions.** Each line represents a segment, that shares the same experimental condition and contains data of consecutive sessions. Multiple lines (segments) are shown per patient, if multiple conditions have been used during his/her training. Conditions are denoted as in Figure 1. The segment legend denotes the patient number combined with the condition information and the epoch number the segment started at. **Left:** Performance curves of data segments that used the 6D condition. **Right:** Performance curves for HP segments. Note that no patient started the online training with the HP condition. The curves depict the *relative* performances with respect to the 5-fold AUC score obtained at the first data window of each segment. For both conditions, histograms show the distribution of linear slopes for the segments contained.

ble at an AUC between 0.8 and 0.9, the change to the more challenging *HP 250* condition started an interesting process: After an initial drop to AUC values below 0.7, the learning curve recovered over the following 70.000 epochs (8 sessions) and eventually surpassed the initial 6D performance. This suggests that increasing task difficulty may stimulate the learning for this patient, while the original 6D condition may not have provided a sufficiently large learning pressure or motivation to further improve the ERP classification.

While we have not analyzed a potential influence of SOA values on the learning curves, the results of P8 exemplify our finding that increasing learning curves were observed for all HP segments, but only for six out of ten 6D segments, see the histograms in Figure 3. Certainly, it would be premature to claim that the harder HP conditions are more beneficial for learning, as the unbalanced histogram finding could be caused by ceiling- or floor effects as well, or by the need to first adjust to the harder HP condition after the easier 6D condition.

Comparison with ERP learning curves from other studies: The studied patient data may be special in the sense that the patients' initial performances were probably lower than those of healthy controls. While healthy controls may quickly run into a ceiling performance, patients in principle could have increasing learning curves.

This is similar to what is seen in [7] and [8]. In the study by Klobassa and colleagues, participants using a P300 speller with relatively easy audio-visual stimuli show good performance from the start, followed by little to no learning over sessions. However, participants using

the more challenging auditory-only P300 speller initially had a much lower performance, but do show meaningful learning over time. In the study by Baykara and colleagues, significant learning is only observed in the first three sessions, possibly indicating that a ceiling performance is reached rather quickly.

Session-based and windows-based analysis: Our *session-based* analysis highlights the influence of different sessions on the decoding performance. Since sessions were conducted on different days, this analysis best reflects day-to-day differences caused by factors such as sleep quality, mood, cap fitting, and any learning that occurs between sessions. This session-based analysis, however, has the disadvantage that results are confounded by varying dataset size. Our *window-based* analysis used a constant amount of 5000 epochs for training and evaluating classifier performances, which is not a given for the session-based analysis. While this makes btLDA performances comparable, the window size is a critical trade-off between training set size (and thus classification performance), the ability to follow non-stationary EEG distributions, inter-session effects and inter-condition effects. The latter may be specifically critical, as a window that bridges between an easier and a more difficult condition smears out and dampens any condition-induced performance differences over time. Around these transitions, the moving window classifier mixes data from two experimental conditions, for which different features may be informative. This can result in a seemingly reduced classifier performance for that bridging window, even if the true performances are equal and overall higher in both

conditions, if analyzed separately. Due to these differences in the two analyses, we decided to include both the session-based and windows-based flavour in this work.

Adaptive, non-causal analysis: Our post-hoc analysis used sessions or sliding time windows to both train and evaluate a classifier. This allows the classifier to adapt to long-term changes of the signals, which are an intended consequence of the rehabilitation procedure. As each session or window was effectively treated as a separate dataset, the reported performance mainly reflects the internal data quality within that specific session or window. This design choice, however, did not allow to evaluate sessions or windows relative to a common reference dataset, which means that a similar performance of two windows does not necessarily imply similarity in the underlying features.

Comparison of online and offline sessions: Our analysis was restricted to the online sessions of the ERP training and has excluded the offline sessions which did not provide feedback and were conducted before and after the 30 hours of online feedback training. Musso and colleagues [9] reported a significant grand average improvement of the binary classification performance when comparing pre- (AUC approx. 0.64) and post-training (AUC approx. 0.76) offline sessions. Please note, however, that these values were obtained using a shrinkage-regularized LDA classifier, which may perform less than the btLDA classifier in our study [12].

Limitations: The interpretation of our results is challenged by the diversity in patients and the—despite a constant task—changing difficulty. As sessions were of differing length, and the time between sessions also varied, it is difficult to compare performances between patients. We are aware that improvements in BCI decoding do not necessarily result in improved language performance. This question will be addressed in future work with larger patient numbers.

CONCLUSION

We analyzed a rare experimental paradigm consisting of multi-session auditory ERP training with trial-wise feedback about the ERP classification and a continuously challenging task difficulty. The study was conducted with stroke patients, who at the beginning of the training showed, compared to healthy participants in a similar age range, a lower ERP classification performance.

Overall, we observed that the stimulus conditions determined most of the observed performance changes. Breaking down the data into multi-session segments of constant stimulus conditions revealed high variability between patients in their ability to improve the ERP decoding performance over sessions. Increasing learning curves were observed systematically for the more demanding headphone mono condition, compared to the easier 6D stereo condition, suggesting that higher task difficulty may promote learning in ERP-based training.

ACKNOWLEDGMENTS

Authors received support from the Dutch Research Council (NWO) via the Dutch Brain Interface Initiative (DBI2), grant 024.005.022 and from BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG), grant EXC 1086.

AUTHOR CONTRIBUTIONS

MT, SR, AdR: Conceptualization, Methodology, Writing - original draft - review & editing; SR, AdR: Formal Analysis MT: Supervision; MM: Writing - review & editing MT, MM: Funding acquisition

REFERENCES

- [1] Wolpaw JR et al. Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis. *Neurology*. 2018;91(3):e258–e267.
- [2] Vansteensel MJ et al. Longevity of a brain–computer interface for amyotrophic lateral sclerosis. *New England J of Medicine*. 2024;391(7):619–626.
- [3] Milekovic T et al. Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. *J of neurophysiology*. 2018;120(7):343–360.
- [4] Nijboer F et al. A P300-based brain–computer interface for people with amyotrophic lateral sclerosis. *Clin. neurophysiology*. 2008;119(8):1909–1916.
- [5] Botrel L, Holz EM, Kübler A. Using brain painting at home for 5 years: Stability of the P300 during prolonged BCI usage by two end-users with ALS. In: *Int. Conference on Augmented Cognition*. 2017, 282–292.
- [6] Pichiorri F et al. Brain–computer interface boosts motor imagery practice during stroke recovery. *Annals of neurology*. 2015;77(5):851–865.
- [7] Klobassa DS et al. Toward a high-throughput auditory P300-based brain–computer interface. *Clin. neurophysiology*. 2009;120(7):1252–1261.
- [8] Baykara E et al. Effects of training and motivation on auditory P300 brain–computer interface performance. *Clin. Neurophysiology*. 2016;127(1):379–387.
- [9] Musso M et al. Aphasia recovery by language training using a brain–computer interface: A proof-of-concept study. *Brain communications*. 2022;4(1):fcac008.
- [10] Schreuder M, Rost T, Tangermann M. Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. *Frontiers in neuroscience*. 2011;5:112.
- [11] Schreuder M, Höhne J, Blankertz B, Haufe S, Dickhaus T, Tangermann M. Optimizing event-related potential based brain–computer interfaces: A systematic evaluation of dynamic stopping methods. *J of neural eng.* 2013;10(3):036025.
- [12] Sosulski J, Tangermann M. Introducing block-toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain–computer interfaces. *J of neural eng.* 2022;19(6):066001.