Proceedings of the
7th Graz Brain-Computer Interface Conference 2017 DOI: 10.3217/978-3-85125-533-1-36

CHALLENGING THE ASSUMPTION THAT AUDITORY EVENT-RELATED POTENTIALS ARE INDEPENDENT AND IDENTICALLY DISTRIBUTED

D. Hübner, M. Tangermann

University of Freiburg, Dept. Computer Science,
Excellence Cluster BrainLinks-BrainTools, Freiburg, Germany

E-mail: david.huebner@blbt.uni-freiburg.de

ABSTRACT: The majority of brain-computer interface classifiers assumes that repeated events elicit brain potential responses which follow the same class-wise distributions. A few adaptive classifiers can deal with violations of this assumption and compensate for non-stationarities occurring on time scales of minutes to hours. This work reports on non-stationarities observed on much shorter time scales. An electroencephalogram study was conducted with elderly subjects ($N = 20$) using an auditory event-related potential paradigm with bisyllabic words as stimuli and a stimulus onset asynchrony of 250 ms. The collected data reveals three effects within a single sequence of 90 stimuli: (1) habituation: the duration of the ongoing sequence negatively correlates with the P300 amplitude, (2) outliers: stimuli at the start and end of each sequence have a special structure, and (3) order effects: longer target-to-target intervals lead to higher P300 amplitudes. Observing that the performance of linear discriminant analysis, a widely used classifier, suffers from these effects, we propose several mitigation strategies.

INTRODUCTION

The centrepiece of a Brain-Computer Interface (BCI) is the decoder which translates brain signals into meaningful control commands, e.g., to spell text without using muscular pathways [1]. One of the most widely used brain signal features in the electroencephalogram (EEG) are so-called *event-related potentials* (ERPs), transient potential responses elicited by events such as visual or auditory stimuli. In a ERP-based BCI, the decoder is deciding for each stimulus whether it was attended (*target stimulus*) or not (*non-target stimulus*). Generally, this is achieved by training a classification model on calibration data under the assumption of stationarity, i.e. that both, the (labelled) calibration data and any data recorded during online use share the same distribution. For instance, in ERPs, a common assumption is that both classes – targets and non-targets – are multivariate Gaussian distributions which share the same covariance [2].

Even though it is well-known that the distribution of brain signal features can change over the course of a session [3–5] or between calibration phase of a BCI and its online use [5, 6], many classifiers assume that all data points are independent and identically distributed (IID) [2]. Adaptive classifiers exist which can continuously adapt to changing distributions and may not even require label information [3, 7]. However, this adaptation of classifiers typically happens on time scales of minutes to hours. For shorter time scales, adaptive approaches are not feasible if they require the tracking of distributions in order to achieve the adaptive behaviour. In this work, we focus on non-stationarities and violations of the independence assumption in the data distribution, which take place on very short time scales. We systematically analyse (1) the effect of habituation, (2) effects of stimuli at the beginning and the end of a stimulation sequence and (3) order effects, specifically the influence of target-to-target distances. All of them are investigated within the time frame of a single sequence of 90 stimuli, which typically lasts only a few seconds in ERP-BCI paradigms. While these three aspects have been reported in the literature, existing studies either lack a connection to BCIs, have used very long interstimulus durations or have covered only a single aspect of the overall problem [8–14].

The *habituation* effect describes how the repeated presentation of a stimulus affects the ERP response. In two studies [8, 9], Polich and colleagues have studied the habituation of the P300 amplitude in an auditory oddball task – which is to discriminate a high tone from a low tone – with a relatively long stimulus onset asynchrony (SOA) of 1.2 s and 2 s, respectively, which clearly are beyond the fast SOA values utilized in current ERP-BCI paradigms. In the first study [8] it was found that the P300 amplitude decreased only slightly over repeated stimulus presentations, and it was reported to remain constant in the second study [9]. For another oddball study by Murphy and colleagues (SOA=1.2 s−1.6 s) a decrease in amplitude was reported as long as the length of each stimulus sequence was unpredictable for the subjects [10].

The second aspect of our study is the response to stimuli which are located at the beginning and end of

 DOI: 10.3217/978-3-85125-533-1-36

each sequence. From the literature, it is known that brain responses to novelty (P3a ERP component) are different from responses to infrequent, task-relevant stimuli (P3b) in latency, peak position and peak amplitude [11]. We suspected to see outlier responses in the form of P3a ERP components at the beginning and end of the stimulus sequence while observing a P3b within the running sequence.

Third and lastly, we focused on how the target-to-target interval (TTI) influences the brain responses. The TTI is defined as the time between the onset of the current and of the preceding target stimulus. Based on the literature, we assumed that longer TTI values yield stronger P300 responses (see [12] for a review). In addition, it has been reported that longer TTIs yield higher amplitudes of the early negativity (with a latency of approx. 150 ms post stimulus onset) in an auditory oddball task with TTI values ranging from 1 to 16 s [13], which is again beyond the TTI range used in current BCI paradigms. Specifically, the first target was found to have a much higher P300 amplitude [14]. Taken together, a confirmation of these three effects in the context of realistic BCI stimulus conditions would clearly violate the assumption that each stimulus elicits an independent and identically distributed brain response and would leave room for improving the classification approach in BCI. So far, only a few attempts have been undertaken to realize this improvement. Citi and colleagues suggested a weighting of the classifier outputs depending on their TTI [12]. A contribution by Martens et al. suggested training one classifier for each TTI [15]. However, both of these studies only focused on TTI, neglecting the other two effects.

The goal of this work is to conduct a comprehensive analysis of violations of the IID assumption under realistic SOA conditions (250 ms) and by using bisyllabic words as stimuli which are more complex and realistic compared to traditional oddball tones. The results are discussed in the context of BCI classifiers for which we will propose possible enhancement strategies.

MATERIALS AND METHODS

An EEG study with $N = 20$ normal hearing subjects (10 female, mean age 60.20 yrs, SD 8.04 yrs) was conducted. It was approved by the Ethics Committee of the University Medical Center Freiburg, and subjects expressed written informed consent prior to participation. EEG signals from 63 passive Ag/AgCl electrodes (EasyCap) were recorded, which were placed approximately equidistantly according to the extended 10–20 system. Impedances were kept below 20 kΩ, and channels were referenced against the nose. The signals were registered by multichannel EEG amplifiers (BrainAmp DC, Brain Products) at a sampling rate of 1 kHz.

Subjects were seated within a ring of 6 loudspeakers (AMUSE paradigm, [16]). Six bisyllabic German words (Drucker, Flasche, Glocke, Knöpfe, Stempel, Trichter; length=300 ms) were chosen as stimuli by the following constraints: Words should have similiar frequency in the German language, should be unambiguous and represent objects which can be depicted. They were played with a 1:1 relation between words and loudspeakers and had an SOA of 250 ms. We define a *trial* as a series of 90 word stimuli. In total, 36 trials were recorded per subject, each consisting of 15 target- and 75 non-target stimuli. The target word/direction was cued at the start of each trial and changed between trials. Within each trail, we grouped 6 consecutive stimuli as one *iteration*, yielding 15 iterations per trial. A target occurred once per iteration. The exact sequences were pseudo-randomized over iterations such that between 2 and 10 non-target stimuli appeared between two targets. The complete stimulus sequence of a single trial took $90 \cdot 0.25\,s = 22.5\,s$ to play.

Data was analysed offline. A third order bidirectional Chebyshev Type II bandpass filter between 0.5 and 12 Hz was applied and data was downsampled to 100 Hz. Eye artefacts were projected out using bipolar EOG recordings [17]. We extracted signal epochs from $[-250, 1000]$ ms relative to each stimulus onset. They were corrected for baseline drifts observed in the interval $[-250, 0]$ ms. Epochs in which the difference between maximum and minimum exceeded 60 μV were treated as outliers and excluded from further analysis. In total, 9.41% of target epochs and 9.03% of non-target epochs were excluded.

Classification was performed using a shrinkage-regularized linear discriminant analysis (*shrinkage-LDA*), a commonly used classification model for ERP signals in BCI [2]. For all 63 channels and 9 intervals per channel located between 100 ms and 1000 ms post stimulus, the average amplitude was computed and used as features for classification, resulting in a 567-dimensional feature vector per epoch.

RESULTS

Overall, the observed ERP responses revealed three kinds of violations of the IID assumption within the course of a single trial (90 stimuli).

First, *habituation* of the target P300 amplitude was observed over the trial duration of 22.5 s. On the grand average view (see Fig. 1) the habituation was expressed by a decreased target response in the central channel 'Cz' from 2.8 μV in the first iteration to only 0.8 μV in the second to last iteration. Fitting a

 DOI: 10.3217/978-3-85125-533-1-36

linear regression model yielded a significant influence of the iteration number ($p = 2.28e-08, r = 0.31$):

$$\text{Amp} = 2.49\mu V - 0.17\mu V \cdot \text{Iteration\#}.$$

We chose 580 ms for the evaluation, as the maximum of the grand average target response was located at this latency. The value of the second to last iteration was reported, because ERP responses in the last iteration were subject to another effect which is described in the next paragraph.

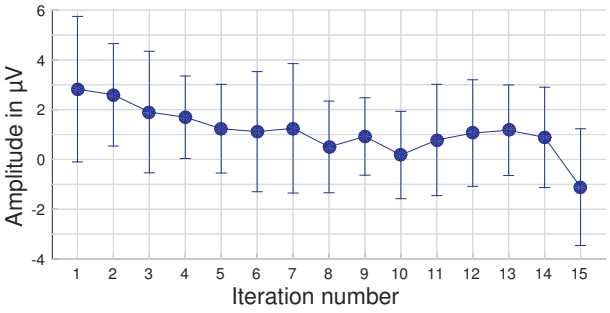

Fig. 1: **Grand average (N=20) target amplitude in Cz at 580 ms post stimulus as a function of the iteration number**. Error bars show the standard deviation across subjects.

This leads to the second type of observed non-stationary behaviour, expressed by the different responses to the first and last stimuli within a sequence. For both, the masking effect due to (missing) neighbouring stimuli is different from stimuli in the middle of a sequence. This effect is visualized via the observed grand average non-target ERP responses in Fig. 2. The top plot reveals strong amplitudes in frontal to central channels, which represent a P1-N2-P3a complex for the average first non-target ERP response. These responses are strongly reduced for stimuli played in the middle of the sequence as shown by depicting an average non-target response observed at the 45th position of sequences (middle plot). The response to the last non-target (bottom plot) shows relatively strong amplitudes after approx. 400 ms, which could indicate an ERP response upon the non-event of a missing 91st stimulus.

The third effect is the influence of the TTI onto the P300 amplitude. We could replicate results from the literature showing a decreased P300 amplitude with shorter distances between two targets, see Fig. 3. This is a clear violation of the independence assumption of target epochs and also shows that the distributions are not identical. The negative peak at around 200 ms was not affected systematically by the TTI. This is especially interesting, as this ERP component shows a class-discriminative amplitude difference between target and non-target stimuli in auditory paradigms [16, 18–20]. The P300 was practically non-existent for a TTI of 750 ms (brown line), which corresponds to exactly two non-target stimuli between target stimuli.

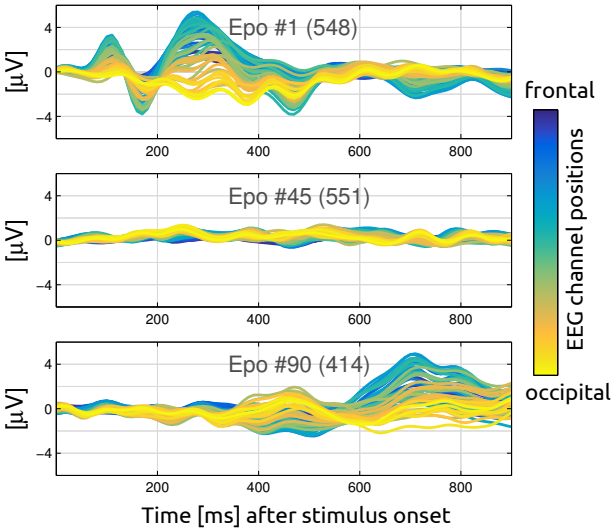

Fig. 2: **Grand average non-target ERP responses for epochs at the beginning (top), the middle (center) and the end of a stimulus sequence (bottom).** Each line depicts the ERP response of one EEG channel, with frontal to occipital channels coloured in blue to yellow. Numbers in parentheses correspond to the number of averaged epochs, with differences caused by artefact removal.

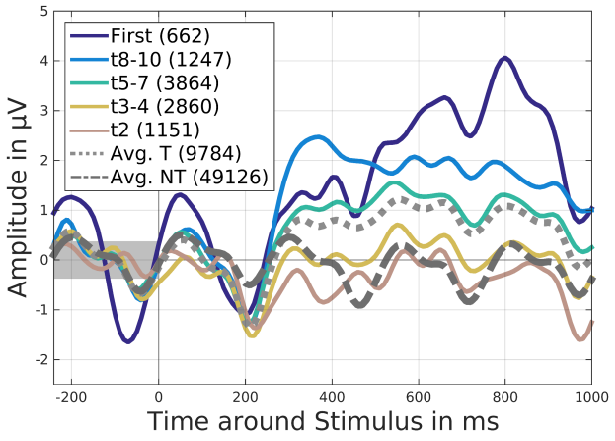

Fig. 3: **Grand average target amplitudes in channel Cz** sorted by the number of non-target stimuli appearing prior to the target stimulus, e.g., t5-7 indicates that since the last target stimulus a number of five to seven non-target stimuli had been played before the next target stimulus was presented. *First*: First epoch per trial, t: Number of preceding non-targets of each target. T and NT: average target and non-target responses over all possible TTIs. Numbers provided in parentheses indicate the number of averaged epochs.

The effect of the TTI upon the classifier has been described previously by Citi et al. [12] for a visual

 DOI: 10.3217/978-3-85125-533-1-36

paradigm. We show how the two other effects, habituation and stimulus position, can affect the classifier performance as well. We chose to test a regularized LDA classifier [2], as a state-of-the-art classifier in BCI. The classifier was rescaled such that the mean target and non-target classifier outputs of the training data are mapped to $+1$ and -1. The classifier performance was estimated by 5-fold chronological crossvalidation, an approach in which the epochs are divided in 5 consecutive blocks, from which 4 blocks are always used for training and one for testing the classifier. Classifier outputs of all test epochs were sorted according to their positions within the sequence of 90 stimuli, averaged over all trials and subjects and plotted in Fig. 4.

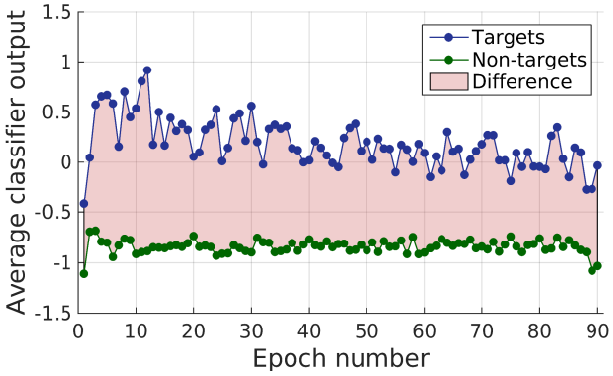

Fig. 4: **Grand average classifier outputs** for target- and non-target epochs plotted as a function of their position within a trial's sequence.

It can be observed, that the non-target classifier outputs for this unseen test data remain relatively stable around -0.85 over the trial, while the target outputs decrease over the duration of a trial's sequence. In addition, target epochs located at the first sequence position may appear as outliers, as their classifier outputs are similar to those of non-target epochs. These two effects show that the discriminatory power of the classifier suffers especially in the beginning and with the ongoing length of a trial.

DISCUSSION

We showed how the stimulus position within a sequence and the preceding stimuli can influence the ERP responses, and that these effects lead to systematic variations during a single sequence of 90 stimuli. Most findings were coherent with the literature. However, we observed no changes in the amplitude of the early negativity as a function of the TTI which was previously reported in [13]. We also found that habituation was more pronounced than previously reported in the literature. We observed a reduction in mean amplitude from 2.8 μV to 0.8 μV corresponding to a drop of 71 %. In contrast, Polich et al. [9] observed no difference in amplitude values for any ERP component as a function of the epoch number. In a second study by Polich [8], a significant decrease was found, however it was rather weak and concluded to be "more spurious than real". Concerning TTI, we observed almost a complete extinction of the P300 response for short TTIs. This is also surprising as an auditory oddball experiment by Höhne et al. [18] showed P300 components for SOA values as short as 125 ms.

We believe that three effects contribute to these observations: (1) Using words instead of simple tones can lead to delayed ERP responses [21], (2) the short SOA of 250 ms may reduce the amplitude of ERP responses [18] and (3) elderly subjects have been reported to show weaker and later P300 amplitudes [22] compared to many BCI offline studies performed with young subjects.

Not surprisingly, we found indications, that an LDA classifier, which assumes IID data points, is suffering from these effects. In the following, we propose different mitigation strategies to overcome these non-stationarities and improve current classifiers.

Adjusting the stimulus order

An easy-to-implement solution is to change the order of stimuli. Instead of allowing for a wide range of TTIs, it might be beneficial to limit them to a narrow range of possibilities, e.g., 4-7 non-targets between two targets in our paradigm. Following the observations of Tangermann et al. [23], it may not even be necessary to retain uncertainty in the sequence. To some extent, this concept is already implemented by the pseudo-randomization of the stimuli order, which at least avoids the subsequent highlighting of the same symbol in visual speller and is used by many groups [24–26].

Weighting individual epoch

Citi et al. proposed an approach in which classifier outputs of each epoch were weighted according to their TTI [12]. To select a target at the end of a trial, this approach should give a higher relevance to more informative epochs. This approach could also be used to deal with the special brain responses in the beginning and end of each trial, e.g., by reducing their influence. A downside of this approach is that it does not actually solve the underlying problem of the violation of the IID assumption, but rather fights the symptoms of bad classifier outputs for some epochs.

Training of sub-classifiers

In contrast, Martens and colleagues outlined an approach in which an individual classifier is trained for each TTI [15]. They showed that specifically those

 DOI: 10.3217/978-3-85125-533-1-36

targets with small TTI can benefit. Considering a bias-variance trade-off, this approach will have a smaller bias, as the individual classifiers are able to capture the characteristics of the epoch-wise brain responses and their dependency on TTI more accurately. However, it will have a larger variance as fewer data points can be used to train each of these individual classifiers.

A similar idea was previously applied in another context by Höhne et al. [27], who observed that ERP responses vary for each of the individual stimuli due to different stimulus properties, e.g., length, pitch or loudness. They exploited this observation by creating individual LDA classifiers for each of the stimuli which give higher weight to the mean estimation of that specific stimuli and thus, reduce the influence of the other stimuli on the mean estimation. Their results show that this approach can improve performance in auditory ERP data and could be easily transferred to deal with habituation or TTI effects.

Adding additional features

The TTI and epoch number can be given as additional features to the classifier enabling it to learn dependencies on those parameters as well and thus, to partly overcome independence violations and non-stationarities in the data. However, one has to be careful whether the classifier model is suited for discrete features or not. Linear discriminant analysis (LDA), for example, assumes multivariate normally distributed features and may perform suboptimally with discrete features.

Data correction

To account for the observation that the first epochs are not influenced from preceding epochs and that they include a novelty P3a, one could add a template of an average non-target and remove a template of a P3a to those ERP responses. The templates could be learned based on data from the same or other subjects. A similar procedure may mitigate the problems observed for epochs at the end of a trial.

CONCLUSION

We showed three different effects – habituation, outlier effects of first and last stimuli, and effects based on the target-to-target interval – which influence the event-related potential responses within a single sequence of 90 stimuli. They clearly violate the assumption that brain responses to single stimuli are class-wise independent and identically distributed (IID). We showed how the decoding performance of a state-of-the-art classifier, regularized linear discriminant analysis, varies within a sequence of 90 stimuli as a result of this violation. To overcome this loss in discriminatory power, we proposed several mitigation strategies, partly by modifying the stimulus presentation and partly by changing the data processing and classification. The next step will be to implement and compare these strategies to ultimately enhance the decoding quality in ERP-based BCIs.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support by BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG), grant number EXC 1086 and by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 39/963-1 FUGG. The authors would also like to thank Simone Denzer for her involvement in recording the data. We also thank the reviewer for their helpful feedback.

REFERENCES

[1] Farwell L A and Donchin E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[2] Blankertz B, Lemm S, Treder M, Haufe S, and Müller K-R. Single-trial analysis and classification of ERP components, a tutorial. *NeuroImage*, 56(2):814 – 825, 2011.

[3] Vidaurre C, Kawanabe M, von Bünau P, Blankertz B, and Müller K-R. Toward unsupervised adaptation of LDA for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 58(3):587–597, 2011.

[4] Kindermans P-J, Schreuder M, Schrauwen B, Müller K-R, and Tangermann M. True zero-training brain-computer interfacing–an online study. *PLOS ONE*, 9(7):e102504, 2014.

[5] Shenoy P, Krauledat M, Blankertz B, Rao R P, and Müller K-R. Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13, 2006.

[6] Blankertz B, Dornhege G, Krauledat M, Müller K-R, and Curio G. The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

 DOI: 10.3217/978-3-85125-533-1-36

[7] Hübner D, Verhoeven T, Schmid K, Müller K-R, Tangermann M, and Kindermans P.-J. Learning from label proportions in brain-computer interfaces: online unsupervised learning with guarantees. *PLOS ONE*, 12(4):e0175856, 2017.

[8] Polich J. P300 development from auditory stimuli. *Psychophysiology*, 23(5):590–597, 1986.

[9] Polich J. Habituation of P300 from auditory stimuli. *Psychobiology*, 17(1):19–28, 1989.

[10] Murphy T I and Segalowitz S J. Eliminating the P300 rebound in short oddball paradigms. *International Journal of Psychophysiology*, 53(3):233–238, 2004.

[11] Polich J. Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10):2128–2148, 2007.

[12] Citi L, Poli R, and Cinel C. Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in Donchin's speller. *Journal of Neural Engineering*, 7(5):056006, 2010.

[13] Gonsalvez C J, Barry R J, Rushby J A, and Polich J. Target-to-target interval, intensity, and P300 from an auditory single-stimulus task. *Psychophysiology*, 44(2):245–250, 2007.

[14] Ganin I, Shishkin S, Kochetova A, and Kaplan A Y. P300-based brain-computer interface: The effect of the stimulus position in a stimulus train. *Human Physiology*, 38(2):121–128, 2012.

[15] Martens S, Hill N, Farquhar J, and Schölkopf B. Impact of target-to-target interval on classification performance in the P300 speller. In *Applied Neuroscience Conference*, 2007.

[16] Schreuder M, Blankertz B, and Tangermann M. A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue. *PLOS ONE*, 5(4):e9813, 2010.

[17] Parra L C, Spence C D, Gerson A D, and Sajda P. Recipes for the linear analysis of EEG. *NeuroImage*, 28(2):326–341, 2005.

[18] Höhne J and Tangermann M. How stimulation speed affects event-related potentials and BCI performance. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1802–1805. IEEE, 2012.

[19] Höhne J, Schreuder M, Blankertz B, and Tangermann M. A novel 9-class auditory ERP paradigm driving a predictive text entry system. *Frontiers in Neuroscience*, 5:99, 2011.

[20] Gao S, Wang Y, Gao X, and Hong B. Visual and auditory brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 61(5):1436–1447, 2014.

[21] Tangermann M, Schnorr N, and Musso M. Towards aphasia rehabilitation with BCI. In *Proceedings of the 6th International Brain-Computer Interface Conference*, pages 65–68. Verlag der Technischen Universität Graz, 2014.

[22] van Dinteren R, Arns M, Jongsma M L, and Kessels R P. P300 development across the lifespan: a systematic review and meta-analysis. *PLOS ONE*, 9(2):e87347, 2014.

[23] Tangermann M, Höhne J, Stecher H, and Schreuder M. No surprise-fixed sequence event-related potentials for brain-computer interfaces. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2501–2504. IEEE, 2012.

[24] Townsend G, LaPallo B, Boulay C, Krusienski D, Frye G, Hauser C, Schwartz N, Vaughan T, Wolpaw J, and Sellers E. A novel P300-based brain–computer interface stimulus presentation paradigm: moving beyond rows and columns. *Clinical Neurophysiology*, 121(7):1109–1120, 2010.

[25] Tangermann M, Schreuder M, Dähne S, Höhne J, Regler S, Ramsay A, Quek M, Williamson J, and Murray-Smith R. Optimized stimulation events for a visual ERP BCI. *Int. J. Bioelectromagn*, 13(3):119–120, 2011.

[26] Verhoeven T, Buteneers P, Wiersema J, Dambre J, and Kindermans P-J. Towards a symbiotic brain–computer interface: exploring the application–decoder interaction. *Journal of Neural Engineering*, 12(6):066027, 2015.

[27] Höhne J, Blankertz B, Müller K-R, and Bartz D. Mean shrinkage improves the classification of ERP signals by exploiting additional label information. In *Pattern Recognition in Neuroimaging, 2014 International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4. IEEE, 2014.