Published in final edited form as:

J Neural Eng. 2013 February ; 10(1): 016006. doi:10.1088/1741-2560/10/1/016006.

Classifier-based latency estimation: a novel way to estimate and predict BCI accuracy

David E Thompson¹, Seth Warschausky², and Jane E Huggins^{1,2}

¹Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI

²Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, MI

Abstract

Objective—Brain-computer interfaces (BCIs) that detect event-related potentials (ERPs) rely on classification schemes that are vulnerable to latency jitter, a phenomenon known to occur with ERPs such as the P300 response. The objective of this work was to investigate the role that latency jitter plays in BCI classification.

Approach—We developed a novel method, classifier-based latency estimation (CBLE), based on a generalization of Woody filtering. The technique works by presenting the time-shifted data to the classifier, and using the time shift that corresponds to the maximal classifier score.

Main results—The variance of CBLE estimates correlates significantly ($p < 10^{-42}$) with BCI accuracy in the Farwell-Donchin BCI paradigm. Additionally, CBLE predicts same-day accuracy, even from small datasets or datasets that have already been used for classifier training, better than the accuracy on the small dataset ($p < 0.05$). The technique should be relatively classifier-independent, and the results were confirmed on two linear classifiers.

Significance—The results suggest that latency jitter may be an important cause of poor BCI performance, and methods that correct for latency jitter may improve that performance. CBLE can also be used to decrease the amount of data needed for accuracy estimation, allowing research on effects with shorter timescales.

1. Introduction

Brain-computer interfaces (BCIs) are commonly defined as a technology that allows communication and control without requiring muscle movements by the user [1]. As such, BCIs offer hope of communication for people with the most severe muscle impairments, such as locked-in syndrome.

One of the most common BCIs researched today is the P300 Speller, first proposed by Farwell and Donchin [2]. The P300 or P3 Speller is so named because it utilizes the P300 response, a positive deflection approximately 300 ms post-stimulus that is evident in electroencephalogram (EEG) recordings. In most modern methods, some form of automated data mining or machine learning is used. As a consequence, several event-related potentials (ERPs), including the P300, can be used in the classification process [3, 4]; accordingly, the system will be referred to herein as the ERP-Speller.

In this paradigm, the brain activity of users is recorded while the users are presented with a series of stimuli, often flashing rows and columns in a grid of letters. The system must then identify the stimulus the user is attending to, by determining which brain responses contain the

indicative ERPs. In the signal processing community, this process is known as classification. Particularly in EEG, the signal-to-noise ratio is so low that ERPs are buried within noise and difficult to detect or study from a single stimulus; single-response measurements are of interest in neuroscience and psychophysiology as well as BCI (see, for example, [5, 6, 7, 8]). In most BCI systems, however, each stimuli is presented multiple times before making a decision. Signal averaging is then performed, either explicitly prior to classification, or implicitly through averaging the classifier scores (mathematically equivalent for linear classifiers).

Signal averaging is also commonly used in the neuroscience and psychophysiology fields, where ERPs such as the P300 have been and continue to be studied in detail - a Web of Science search for “P300” and “ERP” returns over 1900 results, 29 of them in the first four months of 2012. Applications range from studies of cognitive mechanisms in neurotypical samples [9] to disease-specific abnormalities such as those noted in attentional processes in early Huntingtons disease [10] and facial affect processing in schizophrenia [11]. One aspect of ERPs which has been acknowledged and studied in these bodies of literature, but is nearly untouched in the BCI research which relies on ERPs, is that of latency jitter, also known as latency variation. Latency jitter is when the latency, the lag between stimulus and ERP, is not constant between trials.

Despite its name, the P300 response does not always appear precisely 300 ms post-stimulus. Latency has been shown to correlate with age, cognitive ability, and other factors (for reviews, see [12, 13]). Even under tightly controlled conditions, within-user variations are known to exist [14, 15]. Large deviations are observed when attention is divided between two tasks, proportional to the perceptual difficulty of the second task [13]. BCI use, particularly in the home, could be accompanied by several secondary tasks requiring varying amounts of attentional and perceptual resources. Additionally, BCI systems often disregard the criteria recommended for clinical ERP measurement. For example, [16] recommends visual stimuli be presented one to two seconds apart, whereas in BCI systems several stimuli are presented per second. Many of the differences between ERP laboratory studies and BCI system use can be expected to increase latency jitter or otherwise have undesirable effects on signal quality. Latency jitter is also far more troubling than amplitude variations for systems relying on signal averaging, as noted in [2].

Yet latency variation appears to be unstudied in the BCI field; we were unable to find any previous work in the area. We therefore investigated the role that latency jitter plays in BCI operation. We present a new method of measuring latency jitter, called “classifier-based latency estimation” or CBLE, that exploits the time sensitivity of the classifiers used in BCI research. The following is an report of our findings, including the results from the new method, which will be shown to be a useful way of estimating and predicting BCI accuracy.

2. Methods

2.1. Experimental Design

The data were collected according to the methodology presented in Thompson, et al. [17]. Briefly, each participant completed three sessions, on three separate days, using BCI2000's row-column ERP-Speller. On the first visit, participants were asked to copy a 19-character phrase (“THE QUICK BROWN FOX”), which was used to train the online classifier. The training was performed using 15 intensifications of each row and column as stimuli; the number of stimuli during online operation was determined per participant based on training accuracy, varying from 3 to 15 with a mean of 6.7.

The five-character word “JUMPS” was used as an online test. An accuracy of 80% was required to continue the experiment; if an accuracy of 60% or less was observed, training was repeated until 80% was achieved. If 80% accuracy could not be achieved within an acceptable time, the participant was compensated for their time but not invited to continue the study. This test was performed because our protocol included the correction of errors.

Once the desired accuracy was achieved, participants were asked to copy three 23-character sentences, correcting all errors using a backspace option in the BCI. Each sentence was typed into a different device using the BCI and our Multi-purpose BCI Output Device [18]. The second and third sessions were identical to the first, with the exception that the initial 19-character training was not performed, and the classifier from the first session was used. Device order was counterbalanced both within and between participants.

2.2. Participants

These analyses were conducted on data from 31 of the first 32 participants to successfully complete the protocol for this ongoing study. Twenty participants (ages 18–79, mean 45.3) were recruited from the community as able-bodied controls; the data from three of these users were previously reported in [17]. Ten participants (ages 45–78, mean 61.3) with amyotrophic lateral sclerosis (ALS) were recruited from an ALS outpatient clinic. One participant (age 19) with muscular dystrophy (MD) was recruited from an MD outpatient clinic. Only the three users from [17] had previous BCI experience, the rest of the participants had not used a BCI prior to this study. The excluded participant, an individual with ALS, lost track of his place in the sentence several times during the experiment. As such, the data are unusable, as his intent cannot be unambiguously determined.

All activities were performed under the review of our Institutional Review Board (IRB) and conformed to all institutional guidelines on human subject research. All participants gave informed consent and were offered compensation for their time.

2.3. Classifier-Based Latency Estimation

Our new method works by exploiting the classifier's sensitivity to latency variability. The classifier is applied at different time shifts; in this work, the classifier was applied at each integral number of samples between –100 and 100 ms. The shift that corresponds to the maximum score from the classifier is defined as the classifier-based latency estimate (CBLE) for the current observation. Then the statistical variance of this feature (among attended stimuli) is defined as vCBLE.

This new method can be considered a generalization of the Woody & Nahvi technique to align asynchronous responses [19]. Where Woody & Nahvi [19] used the statistical cross-correlation to find the optimal latency shift between a response and a template, our method uses a classifier score. Any linear classifier, which can be expected to show significant vulnerability to latency variability, should be suitable for this technique. Non-linear classifiers may vary in their vulnerability to latency variability; the benefit of this technique may be reduced for some non-linear classifiers. Note that if the mean response to attended stimuli is used as the classifier weights, the first iteration of the original Woody & Nahvi method is recovered.

To confirm that this method is relatively classifier-independent, this study will incorporate the use of two classifiers: one based on least-squares (LS), and another on stepwise linear discriminant analysis (SWLDA). Throughout this manuscript, the subscripts LS and SW will refer to values as calculated from LS- and SWLDA-based data, respectively.

2.4. Data Characteristics

EEG data were collected using a 16-channel g.USBamp from Guger Technologies, which sampled at 256 Hz after hardware filtering. Participants wore a cap with electrodes at locations F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, Oz, and PO8 according to the 10–20 system; reference and ground were the right and left mastoids, respectively. LS and SWLDA classifiers were trained on a per-subject basis using the training set described in section 2.1, which included a total of 570 target stimuli and 2850 non-target stimuli. The classifiers were trained using the default behaviors of P300_GUI: The data were broken into epochs consisting of the first 800 ms after each stimulus. These epochs were decimated by a factor of 13 using a moving-average-and-down sample operation, reducing the number of features to prevent overfitting of the classifiers. The classifier weights were then upsampled by the same factor rather than applying the same operation to all data to be classified.

The online test set included three sentences per session from each participant, each nominally 23 or 24 characters in length. Because the protocol required the correction of errors, and the number of stimuli presented varied per participant, the amount of data available varies between runs. The test set thus includes nine sentences with at least 23 characters each, presented with 3–15 full sets of stimuli (2 target and 10 nontarget stimuli).

CBLE values were calculated offline, by applying the classifier over the range $[-100, 900]$ ms. The estimates could therefore range from approximately -100 ms to $+100$ ms, in steps of about 4 ms. Due to rounding and the decimation operation, the range was not perfectly symmetrical around zero offset.

2.5. Hypotheses and Data Analytic Approach

We hypothesized that vCBLE would correlate significantly with BCI accuracy. Initially, we examined the association between vCBLE and BCI accuracy by computing the Pearson correlation coefficient. vCBLE and BCI accuracy were both calculated persentence, using all available sentences.

Using longitudinal data elements, we then investigated vCBLE as a predictor of BCI performance. We hypothesized that the variance of CBLE on the five-character word “JUMPS” would be a good predictor of the first day's performance, as measured by mean first-day accuracy on three sentences. This hypothesis was tested by comparing the prediction based on vCBLE to that based upon accuracy on “JUMPS”. We obtained the prediction from CBLE for each user in two steps:

- (i) Applying a curve fit from vCBLE to accuracy from all other users full-length sentences to create a graph predicting the overall relationship between vCBLE and accuracy (the current user's data were not used in the calculation)
- (ii) Determining the accuracy value on that predictive graph corresponding to the vCBLE observed in the “JUMPS” data for the current user. This value was used as the predictor for the first day's mean BCI accuracy.

Statistical significance of the resulting prediction was tested based on the Pearson correlation coefficient. We also calculated confidence bounds on the difference between the two correlation coefficients, using methods from [20].

We also hypothesized that vCBLE of training data (the 19-character “THE QUICK BROWN FOX”) would be a good predictor of the first day's performance, again measured by mean accuracy on three sentences. This hypothesis was tested by comparing CBLE predictions to those based on training accuracy, and also the gold-standard for training data:

cross-validation accuracy on the training dataset. The prediction was performed in the same manner as the above test, though the CBLE was calculated on the training data. Confidence bounds were calculated on all pairwise difference between the three methods.

3. Results

Figure 1 is a scatter plot of the relationship between latency variation and BCI accuracy. As can be seen from the figure, vCBLE is highly correlated with online BCI accuracy for both classifiers ($r_{LS} = -0.7451$, $p_{LS} < 10^{-49}$; $r_{SW} = -0.7044$, $p_{SW} < 10^{-42}$). Figure 2 shows that the relationship holds even within the data from individual participants, though two participants appear not to follow the overall trend.

Figure 3 shows the outcomes from the first hypothesis test. Predictions based on online accuracy on the five-character word “JUMPS” are not satisfactory ($r_{LS} = 0.15953$, $p_{LS} = 0.39$; $r_{SW} = 0.45217$, $p_{SW} = 0.01$). Accuracy estimates based on CBLE variance produce stronger and more significant correlations from the same data ($r_{LS} = 0.74402$, $p_{LS} < 10^{-5}$; $r_{SW} = 0.68318$, $p_{SW} < 10^{-4}$). The 95% confidence interval for $\rho_{Acc,LS} - \rho_{CBLE,LS}$ is $[-0.94214, -0.23853]$, indicating that CBLE estimates are significantly more correlated than estimates based on accuracy for LS. The analogous confidence bound for SWLDA is $[-0.54903, 0.051919]$, which includes zero, so the result, while suggestive, does not quite reach the usual threshold for statistical significance.

Figure 4 shows the predictions from the second hypothesis test. Training accuracy was correlated with online accuracy ($r_{LS} = 0.20049$, $r_{SW} = 0.26146$), but the relationship was not significant ($p_{LS} = 0.28$, $p_{SW} = 0.16$). Cross-validation performs better ($r_{LS} = 0.53117$, $r_{SW} = 0.54158$), and is a significant predictor (p_{LS} , $p_{SW} < 0.01$). Predictions based upon CBLE were equivalent or better than both ($r_{LS} = 0.64836$, $r_{SW} = 0.63282$), and are significant (p_{LS} , $p_{SW} < 10^{-4}$). The 95% confidence intervals for each pairwise difference are presented in table 1. The confidence intervals indicate that CBLE performs significantly better than training accuracy for both methods; other differences are non-significant, though the difference between predictions from training accuracy and cross-validation is very close to the significance threshold.

4. Discussion

The method we have proposed, CBLE, has been shown to be a powerful way to estimate and predict BCI accuracy. The ability to predict performance from training data is of interest from a clinical viewpoint: this technique could be used to reduce the time needed to determine if a BCI could be useful for a particular patient. Alternatively, this method may be useful in indicating if additional training data is required - before performing online feedback. This could have important implications in patient morale and trust of a system.

Of particular interest to the BCI community is the ability to predict performance from small datasets, such as 5 characters (see [21], where both training and test data sets were 5 characters long). Statistically, measuring accuracy is a binomial parameter estimation problem. In addition to granularity issues (a test of 5 characters can produce only 6 possible accuracy estimates), the confidence bounds are extremely large; the smallest 95% confidence bounds are for 0 and 5 correct trials, and cover 52% of the accuracy space. The confidence bounds for the other values cover 71% and 80% of the accuracy space. While [21] is an extreme example of data scarcity, some metrics suggested for use with BCIs (e.g. [22]) require accuracy estimates for every class, increasing the data requirements by at least an order of magnitude. Few BCI studies have the data required for such estimates to be

useful. CBLE could be used in larger studies to calculate the more powerful metrics, or in small studies to improve the reliability of accuracy estimates.

Making accurate estimates from these small datasets could improve the value gained from typical BCI experiments. Also, smaller data requirements could be particularly important for the study of non-stationarities in accuracy, such as fatigue effects or distractions. As it stands, state-of-the-art accuracy estimates simply are not precise enough to study these effects on short timescales. This study has shown that CBLE estimates are a better predictor of accuracy from small data sets than existing methods, indicating that the time resolution of such studies could be increased by using CBLE estimates rather than e.g. a windowed 5-character accuracy.

As CBLE variability is so highly correlated with accuracy, it is possible that CBLE might be used to improve BCI performance. Some care must be taken when doing so, however. Naïve approaches, such as simply using the maximal (shifted) classifier score from each observation, are likely to increase the score of negative observations much more than positive observations, since negative observations typically have a much lower unshifted score. Since low variance is associated with high BCI accuracy, one might choose the character with minimal vCBLE; this approach yields classifier accuracies around 70%, though this is considerably lower than the online performance in the original study. Early investigations into an iterative method of building a better classifier, drawing further inspiration from [19], have met with convergence issues.

A second-level classifier may be the easiest effective approach; the use of two classifiers is nearly unavoidable given that CBLE itself requires one classifier. Using the unshifted score, the maximal shifted score, and the square of the estimate of latency offset as input features to a second-level linear classifier has shown some initial success, but has yet to clearly outperform the original classifier. The following characteristics of the CBLE estimates may be of use to other researchers in developing second-level classifiers:

- (i) Target characters (those that should contain a P300 response) typically have low latency variability relative to non-target characters.
- (ii) The classifier output from target characters, as a function of time shift, tends to have a global maxima near, but not at, zero time shift. The output from non-target characters tend to have a global minima near zero shift.
- (iii) If latency shifts are allowed, the classifier output for non-target characters tends to increase more than non-target characters. This is due to the minima/maxima characteristics mentioned above. This effect is somewhat mitigated by choosing narrow windows for allowable time shifts.

4.1. Limitations

CBLE has not yet been validated against existing methods for latency estimation, though the results are similar to prior investigations into latency jitter: 20–80 ms as compared to 15–50 ms reported in [15] for P3b. Our task is comparatively complex, so some deviation is to be expected; additionally, CBLE attempts to model a single latency shift for the entire ERP complex, which is a substantial simplification. Our next step is to implement a method for estimating latency of individual components and confirm if the measures of latency jitter correlate. Without this validation, we are unable to say with certainty that latency jitter is a substantial cause of poor BCI operation, though the CBLE results are suggestive. Regardless of how well CBLE measures latency jitter, the accuracy relationships explored in this work can be of use to the BCI field.

5. Conclusion

We have developed a new method that we have termed CBLE. This method produces a feature that is significantly and strongly correlated with BCI accuracy in the ERP-Speller. Additionally, CBLE is strongly predictive of daily performance, even from small datasets or datasets that have already been used to train the classifier. The technique should be relatively classifier-independent, and the results were confirmed on two linear classifiers. CBLE may have uses in improving classification, as well as predicting performance from smaller datasets. Additional investigations into methods for correcting for latency jitter are also indicated.

Acknowledgments

This work was supported in part by the National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) under Grant R21HD054697 and by the National Institute on Disability and Rehabilitation Research (NIDRR) in the Department of Education under Grant H133G090005 and Award Number H133P090008. The content is solely the responsibility of the authors and does not necessarily represent the official views of NICHD, NIH, NIDRR or the Department of Education. This material is based upon work supported by the National Science Foundation Graduate Student Research Fellowship under Grant DGE0718128. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF.

References

- [1]. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 2002; 113(6):767–91. [PubMed: 12048038]
- [2]. 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):510–23. [PubMed: 2461285]
- [3]. Treder MS, Blankertz B. (C)overt attention and visual speller design in an ERP-based brain-computer interface. *Behav. Brain Func.* 2010; 6(1):28.
- [4]. Bianchi L, Sami S, Hillebrand A, Fawcett I, Quitadamo L, Seri S. Which Physiological Components are More Suitable for Visual ERP Based Brain-Computer Interface? A Preliminary MEG/EEG Study. *Brain Topography.* 2010; 23(2):180–5. [PubMed: 20405196]
- [5]. D'Avanzo C, Schiff S, Amodio P, Sparacino G. A Bayesian method to estimate single-trial event-related potentials with application to the study of the P300 variability. *J. Neurosci. Methods.* 2011; 198(1):114–24. [PubMed: 21439324]
- [6]. Jarchi D, Sanei S, Principe JC, Makkiabadi B. A New Spatiotemporal Filtering Method for Single-Trial Estimation of Correlated ERP Subcomponents. *IEEE Trans. Biomed. Eng.* 2011; 58(1): 132–43. [PubMed: 20923728]
- [7]. Jarchi D, Sanei S, Mohseni HR, Lorient MM. Coupled particle filtering: A new approach for P300-based analysis of mental fatigue. *Biomed. Signal Process. Control.* 2011; 6(2):175–85.
- [8]. Li R, Keil A, Principe JC. Single-trial P300 estimation with a spatiotemporal filtering method. *J. Neurosci. Methods.* 2009; 177(2):488–96. [PubMed: 19041343]
- [9]. Kida T, Kaneda T, Nishihira Y. Dual-task repetition alters event-related brain potentials and task performance. *Clin. Neurophysiol.* 2012; 123(6):1123–30. [PubMed: 22030141]
- [10]. Hart EP, Dumas EM, Reijntjes RH a. M, van der Hiele K, van den Bogaard SJA, Middelkoop H a. M, Roos R a. C, van Dijk JG. Deficient sustained attention to response task and P300 characteristics in early Huntingtons disease. *J. Neurol.* 2012; 259(6):1191–8. [PubMed: 22143614]
- [11]. Lee S-H, Kim E-Y, Kim S, Bae S-M. Event-related potential patterns and gender effects underlying facial affect processing in schizophrenia patients. *Neurosci. Res.* 2010; 67(2):172–80. [PubMed: 20214929]

- [12]. Picton TW. The P300 wave of the human event-related potential. *J Clin Neurophysiol.* 1992; 9(4):456–79. [PubMed: 1464675]
- [13]. Polich J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology.* 2007; 118(10):2128–48. [PubMed: 17573239]
- [14]. Polich J. P300 development from auditory stimuli. *Psychophysiology.* 1986; 23(5):590–7. [PubMed: 3809365]
- [15]. Fjell AM, Rosquist H, Walhovd KB. Instability in the latency of P3a/P3b brain potentials and cognitive function in aging. *Neurobiol. Aging.* 2009; 30(12):2065–79. [PubMed: 18339453]
- [16]. Picton TW, et al. Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology.* 2000; 37(2):127–52. [PubMed: 10731765]
- [17]. Thompson, DE.; Baker, JJ.; Sarnacki, WA.; Huggins, JE. Plug-and-play Brain-Computer Interface Keyboard Performance. 2009 Proc. 4th Int. IEEE/EMBS Conf. on Neural Eng. (Antalya); New York: IEEE; 2009. p. 426-8.
- [18]. Thompson DE, Huggins JE. A Multi-Purpose Brain-Computer Interface Output Device. *Clin. EEG Neurosci.* 2011; 42(4):230–5. [PubMed: 22208120]
- [19]. Woody C, Nahvi M. Application of optimum linear filter theory to detection of cortical signals preceding facial movement in cat. *Exp. Brain Res.* 1973; 16(5):455–65. [PubMed: 4695775]
- [20]. Zou GY. Toward using confidence intervals to compare correlations. *Psychol. Methods.* 2007; 12(4):399–413. [PubMed: 18179351]
- [21]. Guger C, Daban S, Sellers E, Holzner C, Krausz G, Caraballona R, Gramatica F, Edlinger G. How many people are able to control a P300-based brain-computer interface (BCI)? *Neurosci. Lett.* 2009; 462(1):94–8. [PubMed: 19545601]
- [22]. Bianchi L, Quitadamo L, Garreffa G, Cardarilli G, Marciani M. Performances evaluation and optimization of brain computer interface systems in a copy spelling task. *IEEE Trans. Neural Sys. Rehab. Eng.* 2007; 15(2):207–16.

Figure 1.

Accuracy plotted against classifier-based latency jitter estimates (CBLE) using two classifiers. Left: results from a classifier using least-squares (LS) regression; right: the popular step-wise linear discriminant analysis (SWLDA) classifier.

Figure 2. Accuracy vs. variance of CBLE by participant. The data used are the same as in the left panel of Figure 1.

Figure 3.

Predicting first-day accuracy based on a 5-character dataset, by classifier type and estimation method. Left: least-squares classification (LS); right: classification using step-wise linear discriminant analysis (SWLDA). Top: prediction based on 5-character accuracy; bottom: prediction based on classifier-based latency jitter estimates (CBLE).

Figure 4.

Predicting first-day accuracy based on training data, by classifier type and estimation method. Left: least-squares classification (LS); right: classification using step-wise linear discriminant analysis (SWLDA). Top: prediction based on training accuracy; middle: prediction based on leave-one-out cross-validation accuracy on training data; bottom: prediction based on classifier-based latency jitter estimates (CBLE).

Table 1

Confidence Intervals for Correlation Coefficients.

	LS		SWLDA	
	Lower Bound	Lower Bound	Upper Bound	Lower Bound
$\rho_{ACC} - \rho_{CBLE}$	-0.803	-0.238	-0.697	-0.067
$\rho_{ACC} - \rho_{XVAL}$	-0.664	0.003	-0.393	0.065
$\rho_{XVAL} - \rho_{CBLE}$	-0.379	0.115	-0.543	0.093