Bayesian Approach to Dynamically Controlling Data Collection in P300 Spellers

Chandra S. Throckmorton, Kenneth A. Colwell, *Member, IEEE*, David B. Ryan, Eric W. Sellers, and Leslie M. Collins, *Senior Member, IEEE*

***Abstract*—P300 spellers provide a noninvasive method of communication for people who may not be able to use other communication aids due to severe neuromuscular disabilities. However, P300 spellers rely on event-related potentials (ERPs) which often have low signal-to-noise ratios (SNRs). In order to improve detection of the ERPs, P300 spellers typically collect multiple measurements of the electroencephalography (EEG) response for each character. The amount of collected data can affect both the accuracy and the communication rate of the speller system. The goal of the present study was to develop an algorithm that would automatically determine the necessary amount of data to collect during operation. Dynamic data collection was controlled by a threshold on the probabilities that each possible character was the target character, and these probabilities were continually updated with each additional measurement. This Bayesian technique differs from other dynamic data collection techniques by relying on a participant-independent, probability-based metric as the stopping criterion. The accuracy and communication rate for dynamic and static data collection in P300 spellers were compared for 26 users. Dynamic data collection resulted in a significant increase in accuracy and communication rate.**

***Index Terms*—Brain-computer interface, dynamic stopping, P300 speller.**

I. Introduction

ALTHOUGH a variety of augmentative devices are available to assist people with communication disorders, many of these tools require the user to have some motor control in order to operate the device. People affected by severe physical limitations, such as those caused by amyotrophic lateral sclerosis (ALS) or other severe neuromuscular disabilities, may not have the physical ability required to use these devices. Brain-computer interfaces (BCIs) such as the P300 speller provide a noninvasive method of communication that is not reliant on physical movement [1], although the ability to control eye-gaze may impact success with the device (e.g., [2] and [3]). Several studies have indicated that P300 spellers may be viable options for communication for those with ALS (e.g., [4]–[6]).

P300 spellers rely on event-related potentials (ERPs) that occur in scalp-measured electroencephalography (EEG) to determine the character that the BCI user intends to spell (termed the target character). P300 spellers are so named due to the P300 ERP that is elicited by these spellers, although other ERPs may also be used in classification (e.g., [7] and [8]). The P300 is a positive peak in the EEG measurement that occurs approximately 300 ms after an uncommon but relevant stimulus has been presented. In the case of P300 spellers, this stimulus is typically an illumination of the target character. The user attends to the target character while the speller illuminates characters at random. When the target character is illuminated, an ERP is elicited. By detecting the ERP in the recorded EEG responses, the target character can be determined. These ERPs often have low signal-to-noise ratios (SNRs); thus, P300 spellers typically collect multiple repetitions of the EEG responses to all of the character flashes. The repetitions are averaged to improve the SNR of the P300 response and thereby increase spelling accuracy. The number of repetitions collected for averaging is often held constant across participants and target characters, regardless of the SNR of the data. However, numerous offline analyses have suggested that spelling speed could have been greatly improved if the speller had stopped collecting data when the correct character had been selected as the target character (e.g., [9]–[12]).

Several studies have considered methods of adaptively collecting data with significant improvements in spelling speed and/or accuracy [6], [13]–[16]. However, each of these methods has relied in some form on the past performance of the participants to control the data collection. Several of these studies relied on averaging training data across a participant pool to set the threshold for stopping data collection [14], [16]. Using this information for controlling data collection, however, creates the potential for mismatch if the participant pool is changed. This may be a significant issue for BCI users with disabilities since BCI users are likely to differ more from each other, due to the etiology or progression of their disability, than might be expected of a pool of participants without disabilities. While Townsend *et al.* [6] and Jin *et al.* [13] avoided the issue of potential mismatch between participant pools by relying on participant-specific controls to set the stopping criteria, the assumption is made that user performance will remain relatively constant for each target character.

A Bayesian approach to dynamically stopping the data collection process is proposed here. This approach differs from

Manuscript received May 18, 2012; revised October 17, 2012; accepted February 26, 2013. Date of publication March 21, 2013; date of current version May 04, 2013. This work was supported by the NIH under Grant 5R21-DC-010470-02.

C. S. Throckmorton, K. A. Colwell, and L. M. Collins are with the Electrical and Computer Engineering Department, Duke University, Durham, NC 27708 USA (e-mail: chandra.throckmorton@duke.edu; kenneth.colwell@duke.edu; leslie.collins@duke.edu).

D. B. Ryan and E. W. Sellers are with the Department of Psychology, East Tennessee State University, Johnson City, TN 37614 USA (e-mail: ryand1@goldmail.etsu.edu; sellers@mail.etsu.edu).

Digital Object Identifier 10.1109/TNSRE.2013.2253125

previous approaches by basing the stopping criterion on the confidence that the correct character has been selected as the target character rather than relying on a participant-specific metric. The proposed system avoids assumptions about the performance of the participant pool or individual participants; thus providing the flexibility to adjust data collection based on the quality of responses. This flexibility may mitigate issues such as attention shift or increasing fatigue levels.

II. Methods

A. Participants and Equipment

Thirty-one healthy participants were recruited from the student population at East Tennessee State University (ETSU). Four participants were excluded for not completing the study, and the results from one participant were dropped after it was determined that the wrong classifier weights had been used during online testing. Participants volunteered their time, and all data collection occurred at ETSU. Participants gave informed consent, and the use of human participants as described herein was approved by the ETSU Institutional Review Board (IRB). Analysis of the data was approved by the Duke University IRB, and both the use of human participants and data analysis were approved by the National Institute on Deafness and other Communication Disorders.

EEG responses were measured using 32-channel caps from Electro-Cap International, Inc., connected to a computer via two 16-channel GugerTec g·USBAMP Biosignal Amplifiers. Data collected from electrodes Fz, Cz, P3, Pz, P4, PO7, PO8, and Oz were used for classification, referenced to the right ear electrode. These electrodes are commonly used for P300 spellers and have been demonstrated to provide adequate information for communication (e.g., [17]). The EEG responses were sampled at a rate of 256 Hz. The open-source BCI2000 C++ software package developed by Schalk *et al.* [18] was used for stimulus presentation and data collection for the static stopping criterion (SSC). Additional functionality was added to the software, allowing the dynamic stopping criterion (DSC) to be used.

B. P300 Speller Paradigm

Participants were presented with a 9×8 grid of characters and functions (see Fig. 1). This study relied on a row/column paradigm for flashing the grid characters [1]. Each row and column was flashed once, in random order, in a sequence. Thus, a target character was flashed twice in a sequence of 17 flashes (once with a row flash and once with a column flash). In the SSC case, five sequences were collected for each target character. This number of sequences was within the range of sequences suggested as optimal for the data collected in the BCI Competition 2003 (e.g., [9] and [10]), and five sequences were used in Townsend *et al.* [6] to determine the optimum number of sequences for each participant, suggesting that a high level of performance can be expected from five sequences. Flash duration was 62.5 ms followed by an interstimulus interval of 62.5 ms before the next flash. After a target character was selected, an intertarget interval of 3.5 s occurred. Thus, the selection of each character in the SSC condition required 14.1

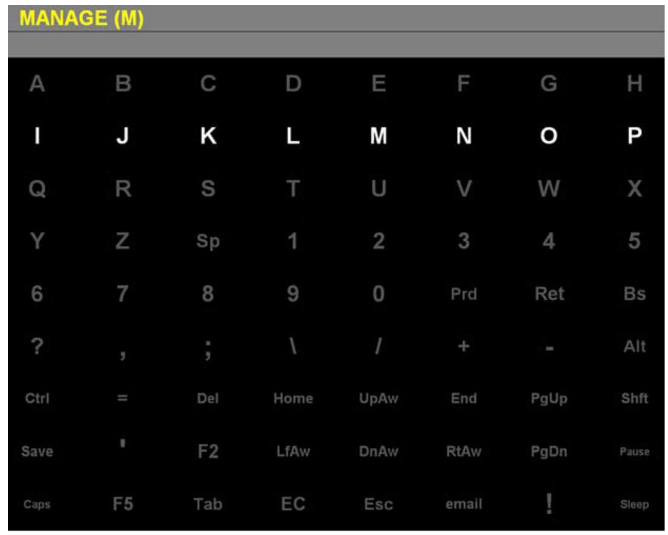

Fig. 1. Screen capture of the interface used in this study. The word to be spelled was displayed in the top gray bar followed by the current target character in parentheses. Feedback was provided by the speller displaying the character it selected below the actual target. Characters were illuminated in rows or columns, and the order was random.

$$\text{seconds: } 3.5 \text{ sec} + \left(\frac{0.125 \text{ sec}}{\text{flash}} * \frac{17 \text{ flashes}}{\text{sequence}} * 5 \text{ sequences} \right).$$

In the DSC case, the number of sequences varied, and the task could end without completing a sequence. The DSC task was allowed to continue for more than five sequences to demonstrate the potential of the algorithm to collect more data, if necessary, to make a confident decision. Ideally, the DSC algorithm would continue until a confident decision was reached; however, a limit of ten sequences was imposed with the assumption that the cost of the time and attentional resources required to select a character with greater data collection would outweigh the potential advantage of getting the character correct with high confidence.

C. Data Collection and Classifier

The words for copy-spelling were randomly drawn from a subset of the available words in the English Lexicon Project [19]. The word set consisted of the 400 six-character words with the highest frequency of occurrence in written communication as measured by HAL corpus frequency [20]. A session consisted of five calibration runs and ten online test runs; five runs for the SSC and five runs for the DSC. Each run consisted of a six-character token.

Participants completed the study in a single session. At the start of the session, participants were asked to copy-spell four words and one random sequence of six numbers. These data were used to train the classifier and provide likelihood estimates for the DSC. The signal preprocessing and the classifier were those provided with BCI2000 [18]. The preprocessing of the

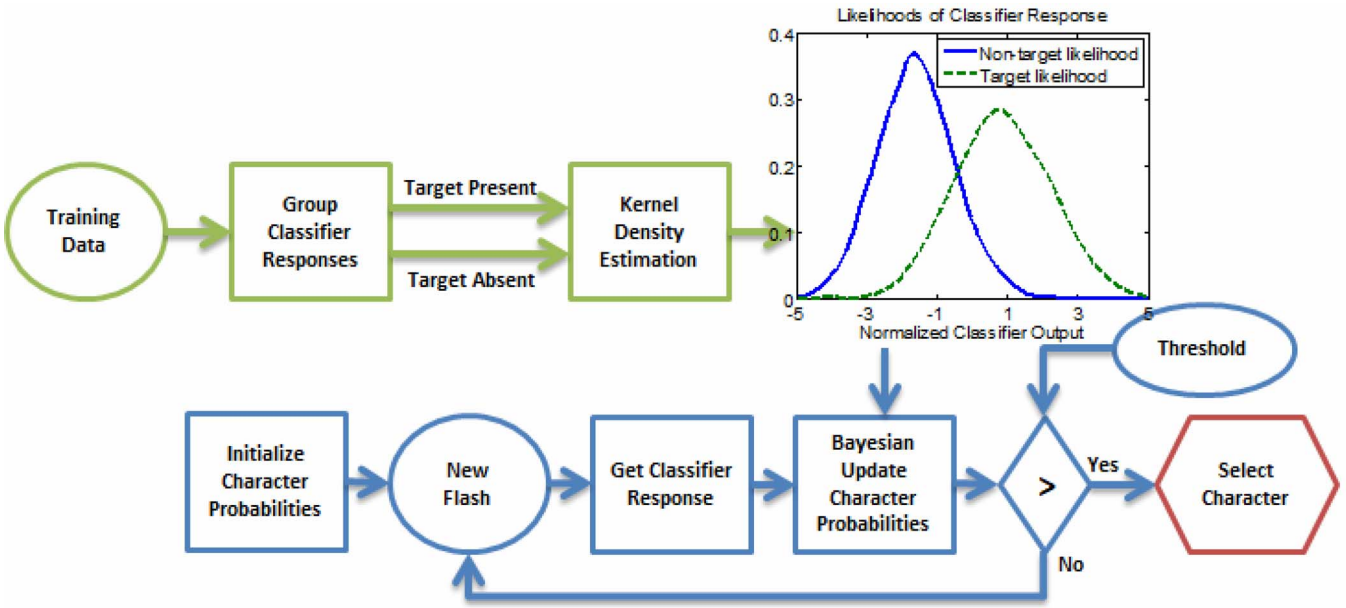

Fig. 2. Flow chart of the DSC algorithm. The upper row indicates the offline calibration necessary to run the algorithm online. Training data is collected and the classifier responses to each flash are grouped by whether a target was present/absent. Kernel density estimation is then used to smooth the histograms of classifier responses and generate likelihood pdfs. The lower row indicates the online processing for dynamically controlling data collection. Before data collection begins, each character is given an initial probability of being the target character. With each new flash, the classifier response is calculated. This response, used with the likelihood pdfs generated from the training data, gives an estimate of the target and nontarget likelihoods. For example, if the classifier response to the flash was −1, then given the pdfs shown, the target likelihood would be estimated as 0.1 while the nontarget likelihood would be estimated as 0.3. The character probabilities are updated with these likelihood values, and if one of the character probabilities exceeds the threshold, that character is selected as the target. If not, a new flash is presented and the process of updating the character probabilities is repeated.

signal consisted of extracting 800 ms of raw EEG signal following each flash, reducing the data dimensionality by partitioning the data into equal lengths of 13 samples and taking the average (reducing the sampling rate to approximately 20 Hz) and concatenating these dimension-reduced features across the channels of interest. This resulted in 120 features per flash (15 features × 8 channels), and these features were used as the input to a stepwise linear discriminant analysis classifier (see Krusienski *et al.* [21] for a description of its use with BCI P300 spellers). Stepwise linear discriminant analysis (SWLDA) has been demonstrated to be effective for discriminating between EEG responses for target and nontarget characters in multiple studies (e.g., [4], [6], [17], and [22]). SWLDA weights the features based on their utility for discrimination. These weights are determined for each participant from the training data. After the training data were collected, participants were then tested with the SSC and the DSC tasks by copy-spelling an additional four words and one random number sequence per task. These tasks were counterbalanced across participants to avoid order effects. While SWLDA was chosen for convenience and its proven functionality, it should be noted that the DSC method is not classifier-dependent and could be used in conjunction with any other classifier as long as the same classifier is used both in training and online spelling.

D. Dynamic Stopping Criterion (DSC)

The DSC algorithm is illustrated in Fig. 2. This algorithm was implemented inside BCI2000 using a row/column P300 speller paradigm and the SWLDA classifier; however, any paradigm and classifier could be used with this algorithm as long as the same paradigm and classifier were used for offline (the collection of the training data) and online processing. During offline processing, illustrated in the top row of Fig. 2, the classifier responses for target and nontarget character flashes were grouped. The training data consisted of 30 target characters for each of which five sequences were measured. Thus, the nontarget group would consist of classifier responses to 2250 nontarget flashes (15 nontarget flashes per sequence × 5 sequences × 30 target characters) while the target group would consist of classifier responses to 300 target flashes. Kernel density estimates (e.g., [23]) using a Gaussian kernel were then calculated for each group to estimate the probability density function (pdf) of the likelihood of the classifier response given that the target character was $(p(x_i|H_1))$ or was not $(p(x_i|H_0))$ included in the current flash, where x_i indicates the classifier response to a flash, and $H_1(H_0)$ indicates the presence (absence) of a target character.

These participant-specific likelihoods are the only requirement for the algorithm to control data collection in real time. The online portion of the algorithm is shown in the bottom row of Fig. 2. The probability of being the target character is initialized for each character, which can be chosen to reflect *a priori* knowledge regarding the characters likely to be chosen. In our implementation, the initial probabilities were set to 1/number of characters, i.e., no prior knowledge was assumed. When a flash occurs, the classifier calculates a confidence regarding whether the flash contained the target character. The likelihood pdfs generated from the training data are used to estimate the likelihood that the classifier response could have occurred if the flash did/did not include the target character, and these likelihoods can be used to update the character probabilities.

The method for updating the probability that a character is the target character based on previous classifier responses is based on Bayes rule (e.g., [23])

$$p(C|X) = \frac{p(X|C)p(C)}{p(X)} \tag{1}$$

where $p(C|X)$ is the current estimate of the character's probability of being the target given all of the classifier responses, X, observed previously; $p(C)$ is the prior probability for the character; $p(X|C)$ is the likelihood of the classifier responses; and $p(X)$ is the probability of the classifier responses. By the rule of total probability (e.g., [24]), the denominator can be replaced as follows:

$$p(C|X) = \frac{p(X|C)p(C)}{\sum_{C} p(X|C)p(C)}. \tag{2}$$

This provides a method for calculating the posterior probability of the character being the target character after all the data has been collected; however, for an online algorithm, the posterior probabilities need to be updated after each flash. If we consider the sequential arrival of classifier responses to be $X = [x_1, \ldots, x_n, \ldots, x_N]$, and we assume that the classifier responses are conditionally independent of each other given the underlying character probabilities, then at time n

$$\begin{aligned} p(C|x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n|C)p(C) \\ &= p(x_n|C)p(x_1, \ldots, x_{n-1}|C)p(C) \\ &= p(x_n|C)p(C|x_1, \ldots, x_{n-1}). \end{aligned} \tag{3}$$

Thus, at time n, the posterior probability of the character being the target character is proportional to the product of the likelihood of the classifier response at time n and the posterior probability of the character being the target character at time $n-1$ [25]. From (3), sequential updating of the character probabilities was carried out using the following:

$$p(C|x_i, S_i, X) = \frac{p(C|X)p(x_i|C, S_i)}{\sum_{i} p(C|X)p(x_i|C, S_i)} \tag{4}$$

where $p(C|X)$ is the current estimate of the character's probability of being the target given all of the classifier responses observed previously; $p(x_i|C, S_i)$ is the likelihood of the current classifier response x_i, given that the character was/was not in the currently flashed set of characters (S_i); and the denominator normalizes the updated probabilities by dividing by the sum over all character probabilities. The likelihood used to update the character probability depends on whether the character was flashed

$$p(x_i|C, S_i) = \begin{cases} p(x_i|H_1) & C \in S_i \\ p(x_i|H_0) & C \notin S_i \end{cases}. \tag{5}$$

TABLE I
EFFECT OF TARGET CHARACTER BEING FLASHED/NOT FLASHED

	Flash Illuminates Target	Flash Does Not Illuminate Target
x_i	Large	Small
$p(x_i / H_1)$	Large	Small
$p(x_i / H_0)$	Small	Large
Target Updated By:	$p(x_i / H_1)$ = large	$p(x_i / H_0)$ = large

This can best be visualized by considering the two cases of S_i including/not including the target character (see Table I). If S_i includes the target character, the classifier response ideally should be large (see first column). A large classifier response results in a high likelihood for $p(x_i|H_1)$ and a low likelihood for $p(x_i|H_0)$. Since a target was flashed, it is updated with the large $p(x_i|H_1)$ value while nontargets that were not flashed are updated with the small $p(x_i|H_0)$ value. Thus, the target character's probability increases while the nontarget characters' probabilities decrease. Similarly, if S_i does not include the target, the classifier response should be small which results in a small $p(x_i|H_1)$ and a large $p(x_i|H_0)$. However, the target character, since it did not flash, is now updated by $p(x_i|H_0)$, and the target probability should increase.

Nontargets are updated with both large and small values of $p(x_i|C, S_i)$, depending on whether or not they flash with the target; however, the variability in these update values leads to the nontarget character probabilities trending towards lower values. After each Bayesian update, the character probabilities are compared to a threshold to determine if the target character can be selected. The threshold indicates the confidence that the correct character has been chosen, and in this study, it was set to 90%. This threshold is a tradeoff between accuracy and speller rate—a higher threshold results in fewer errors but requires more data to be collected while a lower threshold requires less data but results in more errors. The threshold was chosen to encourage a high level of accuracy; however, it is possible that a lower level of accuracy would lead to a higher communication rate [16]. Once a target character is selected, the process begins again for the next target character by reinitializing all of the character probabilities.

III. RESULTS

By basing data collection on the quality of the data received, the DSC method was expected to impact accuracy as well as communication rate. In Fig. 3, accuracy, time to complete the task, bit rate and theoretical bit rate are plotted for both the DSC and SSC conditions. Bit rate provides a measure of communication rate that considers accuracy, the number of possible target characters, and the time required to complete the task (e.g., see [26]). Theoretical bit rate differs from bit rate by not including intervals between target selection tasks in the calculation of the time required to complete the task. The DSC condition provided significant improvements in accuracy and communication rate

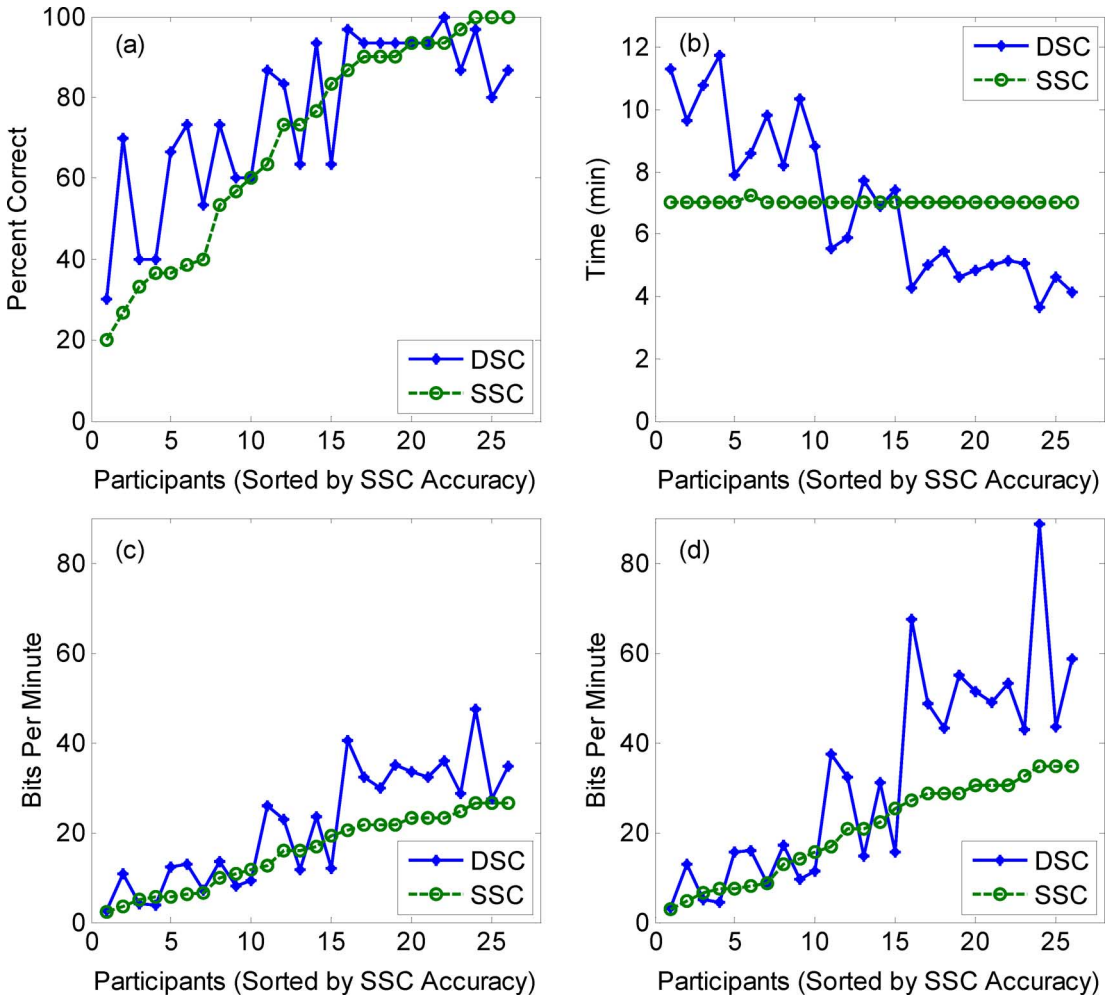

Fig. 3. Comparison of DSC and SSC in terms of (a) accuracy, (b) time to complete the task, (c) bit rate, and (d) theoretical bit rate. Bit rate is a function of accuracy, the number of possible target characters, and time required to complete the task (e.g. see [26]). Theoretical bit rate differs from bit rate by excluding the time spent in the task that is not devoted to spelling. Note that one participant spelled an extra character under the SSC condition, resulting in a slight increase in time to complete the task. Participants are sorted by their SSC accuracy.

TABLE II
GROUP MEANS OF PERFORMANCE

	SSC	DSC	p <
Accuracy	69.4%	75.8%	0.03
Bit Rate	15.6 bits/min	21.4 bits/min	0.0006
Theoretical Bit Rate	20.6 bits/min	32.2 bits/min	0.0004

Group means on three performance measures for SSC and DSC. The DSC method resulted in statistically significant improvements over the SSC method using a Wilcoxon signed-rank test.

(see Table II). The data in Fig. 3(a) indicate that the improvement in accuracy was especially high for participants whose accuracy with the SSC was below 60% correct. Improving accuracy for these participants is critical since accuracy below 50% makes effective communication with the BCI system unlikely. At 50% accuracy, any effort to correct an error through use of the backspace is equally likely to lead to another error; thus, an improvement in accuracy to above 50% correct potentially transitions the BCI system from unusable to usable.

Fig. 3(b) plots the time to complete the task, including intertarget intervals, for each of the participants. For the SSC, participants finished spelling five six-character tokens in seven minutes (participant 6 spelled one extra character, resulting in a slightly higher completion time). The increase in DSC accuracy for participants with low SSC accuracy was achieved by increasing the amount of data collected before making a decision about the target character, thereby increasing the time required to complete the task. However, the corresponding increase in accuracy outweighed the increase in task completion time, resulting in an increase in communication rate for many of the participants with low SSC accuracy [Fig. 3(c) and (d)]. For participants with high accuracy, the DSC reduced the amount of data collected, thereby reducing the time required to complete the task. The reduced task completion time coupled with relatively unchanged accuracy for these participants resulted in large improvements in communication rate.

The maximum time for task completion that could occur for the DSC was 12 min 19 s based on the imposed limit of ten sequences. Interestingly, this limit was not reached despite the continued poor accuracy of some participants (e.g., participants 1 and 4). In order to investigate this effect further, the distribution of the number of flashes to spell each character for the DSC

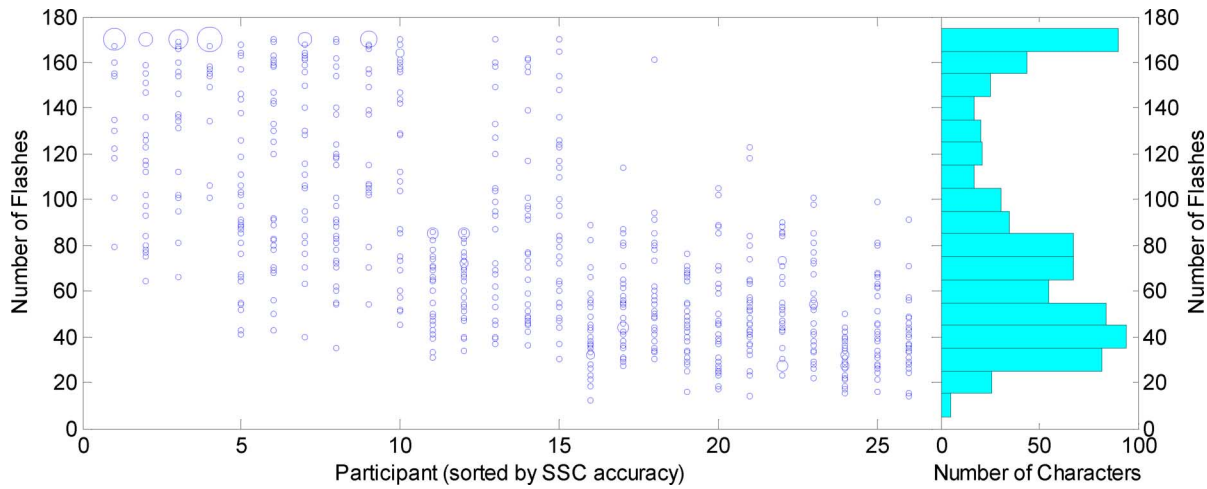

Fig. 4. Number of flashes used to spell each target character for the DSC. On the left, the distribution of flashes that occurred during spelling for each participant is plotted. Note that data are presented in terms of flashes rather than sequences since the DSC could end data collection after any flash rather than having to complete a sequence. The size of each symbol indicates the number of target characters spelled with a particular number of flashes. For example, the large symbols at 170 flashes for participants 1–4 indicate that many target characters were spelled after the limit in the number of flashes had been reached. On the right, the results have been summed across participant.

is plotted by participant in Fig. 4. Note that the distribution is plotted by flash rather than sequence since data collection with the DSC could stop before a sequence had been completed. On the left, the size of each symbol indicates the number of target characters that were selected by a particular number of flashes, e.g., it is apparent that for the first four participants, many of the target characters were selected only after the limit of 170 flashes (10 sequences $\times$ 17 flashes/sequence) had been reached. This may suggest that a higher limit could have led to further improvements in accuracy. It should be noted, however, that even for these participants, many target characters were selected with far fewer flashes. While generally participants with low spelling accuracy required more flashes across all target characters than higher performing participants, the number of flashes required to select target characters varied widely for each participant. Thus, methods that rely on past participant performance to set algorithm parameters may be limiting the potential accuracy and rate that can be achieved.

On the right in Fig. 4, the numbers of target characters selected by each flash count are pooled across participants. The distribution appears to be bimodal with the majority of target characters requiring either 1–5 sequences (17–85 flashes total) for selection or ten sequences. Note that no target character was selected in less than 11 total flashes, but selection of a target character with less than one complete sequence was possible and did occur. At the other extreme, a number of target characters were selected only after the limit of ten sequences were reached.

While it might be expected that accuracy would be poor for target characters selected only after the maximum number of flashes was reached, it is also possible that accuracy would be poor for characters selected with a small number of flashes. Theoretically, this should not be the case since the threshold is a measure of confidence in the response and should not halt data collection until enough data have been collected to make a highly confident decision. However, one advantage of collecting a larger number of sequences is that if an erroneous response is measured (e.g., if a P300 is elicited for a character other than a target character), its effect is minimized by the large collection of data that correctly contradicts it. Thus, a smaller data collection could be dominated by a small amount of misleading data. To investigate whether this was the case, the target characters were grouped by the number of flashes used to select them. For each group, the proportion of correctly selected characters was calculated and plotted in Fig. 5. The circle symbol indicates the proportion of correct characters selected with 85 flashes (five sequences) under the SSC condition, while the diamond symbol indicates the proportion of correct characters selected after the limit of 170 flashes was reached under the DSC condition. As expected, accuracy was poor for these target characters. However, performance appears to be relatively constant otherwise. Thus, decisions based on fewer flashes do not appear to be inherently more erroneous than decisions based on more flashes. Further, DSC performance is similar to or greater than that achieved under the SSC condition.

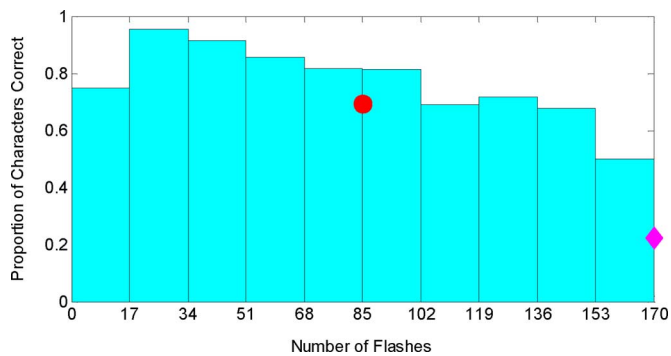

Fig. 5. Proportion of characters correctly spelled for a particular number of flashes for the DSC. Note that each proportion is based on a different number of target characters. The circle symbol indicates the proportion of characters correctly spelled under the SSC condition, and the diamond symbol indicates the proportion of characters correctly spelled after the limit of flashes under the DSC condition had been reached. Performance tends to be independent of the number of flashes with the exception of characters selected after the limit in the number of flashes was reached (170 flashes).

IV. Discussion

Previous dynamic data collection methods have relied on past participant performance either as a group or on an individual basis to set the stopping criterion. Measurements from this study suggest that such assumptions may limit optimization of accuracy and communication rate. The number of sequences required to select target characters varied widely for each participant, regardless of their average accuracy. By relying on the quality of the data to control data collection, the DSC significantly improved accuracy and communication rate. These improvements were often large. Bit rate was increased by greater than 30% and theoretical bit rate by greater than 50% for more than half of the participants. In terms of accuracy, of the seven participants whose SSC performance was below 50% correct, making effective communication unlikely, four of the participants had an improvement in accuracy to above 60% correct with the DSC.

It is worth considering whether similar levels of performance improvement could have been achieved by the SSC if the number of sequences for the SSC had been set on an individual participant basis, such as was proposed by Townsend *et al.* [6], rather than selecting a constant number for all participants. Recent research by Schreuder *et al.* [27] suggests that this would not be the case. In their study, setting the number of sequences based on participant performance during training rarely resulted in significant increases in performance over a constant number of sequences. To consider this point further, a *post hoc* analysis of the training data from this study was conducted using written symbol rate (WSR) to estimate the optimal number of sequences per participant [6], [28]. WSR is a measure of the number of correct characters that can be spelled over time and can be estimated offline from training data for increasing numbers of sequences [6], [28]. The number of sequences that maximizes WSR for each participant in the offline training data can then be used to set the number of sequences used in the online static measurement paradigm. In the *post hoc* analysis, the number of sequences that maximized WSR was determined for each participant. For all but one of the participants (participant 26), WSR was maximized by five sequences. Thus, relying on WSR to set the sequences for the SSC would have left the results of this study relatively unchanged.

Since only five sequences of data were collected during training, the WSR analysis could not estimate whether more than five sequences would have been considered optimal for some participants, as is likely for those with lower SSC accuracy. However, an increase in sequences cannot necessarily be assumed to result in an increase in accuracy for these participants. In Krusienski *et al.* [17], two of the participants with the lowest accuracy experienced little to no benefit from increasing the number of sequences above five. Further, several studies have observed plateaus in performance as the number of sequences increases [13], [29]. However, based on Fig. 4, it may be possible to hypothesize under what conditions the SSC would have resulted in accuracy similar to that of the DSC and consider whether the DSC would still be expected to improve communication rate over the SSC. One hypothesis might be that to achieve the DSC accuracy, each target under the SSC would need, at minimum, the number of sequences used by the DSC. Since the SSC by definition uses a constant number of sequences across all targets, the SSC number of sequences would have to be set to the maximum number of sequences used by the DSC to ensure that all targets are classified by at least their minimum number of sequences. For the majority of lower accuracy participants, the maximum number of sequences used by the DSC was ten. Assuming the same accuracy between the SSC and the DSC, using ten sequences for the SSC would result in much lower communication rates than those observed for the DSC.

Another hypothesis might be that the DSC accuracy would be achieved if the number of sequences for the SSC had been set to the average number of sequences used by the DSC. The reasoning might be that while some targets would be classified with too few sequences, the classification of others would be improved by the increased data collection. The number of flashes that would have been collected by the SSC if the number of sequences were increased to the average number of sequences used by the DSC was calculated for each participant and compared to the actual number of flashes used by the DSC. In all cases, the DSC used fewer flashes total. Thus, while this issue deserves further investigation, initial results suggest that increasing the number of sequences for the SSC would not result in the same level of improvement seen by the DSC.

Direct comparisons between the improvements resulting from the DSC and those observed in the literature for other proposed dynamic data collection techniques are difficult given differences in paradigm and subject pool. Given those caveats, however, some comparisons are considered. Schreuder *et al.* [27] compared four dynamic stopping methods to a static method with and without preselection of the number of sequences per participant. They considered five ERP-based spellers, two of which used auditory stimuli and three of which used visual stimuli. The rate at which symbols were correctly spelled for the visual-stimuli-based spellers using static data collection ranged between 1 and 1.5 symbols/min. Dynamic methods mostly ranged between 1 and 2 symbols/min, although some performed worse with specific paradigms. Calculating symbol rate as defined in [27], the SSC had a rate of 1.7 symbols/min and the DSC had a rate of 3 symbols/minute. This suggests that the DSC may perform as well as or better than previously considered techniques. The dynamic stopping technique used in Jin *et al.* [13] resulted in an improvement in theoretical bit rate from approximately 20 to 40 bits/min. The average accuracy of the subject pool for the static method was 90% which is much higher than average accuracy of the subject pool used in this study. However, if results are restricted to subjects scoring 70% correct or better with the SSC, then a similar improvement is observed: 29 bits/min to 46 bits/min. Thus, the DSC seems comparable in terms of performance gains and may offer some advantages over current dynamic stopping techniques, specifically that subject-specific data are not required to set the stopping criterion and that implementation of the algorithm is independent of classifier or paradigm.

Schreuder *et al.* [27] noted in their comparison of dynamic stopping techniques that performance improvements tended to

be limited to those participants with good performance while the DSC algorithm presented in this study improved accuracy for all of the participants with SSC accuracy below 60%. However, one limitation of this algorithm was that for some participants, the improved accuracy level still remained well below the effective communication level (e.g., participants 1, 3, and 4). These participants also had the greatest number of target characters selected only after the limit on the number of flashes had been reached. While it is possible that for these participants accuracy would have been further increased had the limit for the DSC been set to a higher number of flashes, there may be other underlying causes for the poorest performance that cannot be corrected by increasing data collection to improve SNR. Research suggests that the electrodes used for controlling the BCI (e.g., [21]) and the paradigm (e.g., [6], [13], [30], and [31]) can have a significant impact on participants' performance. It is possible that electrode and paradigm selection coupled with the proposed algorithm might jointly improve performance to an effective communication level for the participants with the lowest accuracy.

In addition to these options, the algorithm itself might be optimized in several ways. The choice of threshold is a tradeoff between accuracy and speller speed, with higher thresholds reducing speller speed but tending to increase accuracy. In this study, the threshold was set to a constant, participant-independent high value; however, it might be possible to optimize the threshold to maximize bit rate for each participant. In principle, the optimal threshold can be estimated from training data if the training data can provide an accurate estimate of the function relating threshold to spelling accuracy. In practice, offline simulations conducted prior to this study using previously collected data [32] suggested that using the estimates of optimal threshold derived from the training data did not lead to the best communication rates in the test data. This might be due to inadequate training data or it might be due to the effects observed here and in Schreuder *et al.* [27]; user performance is variable and training data may not adequately reflect future performance. In either case, further investigation is warranted to determine how the threshold should be set for this algorithm in order to maximize performance.

Another technique for optimization of the algorithm would be to use language models to initialize the character probabilities. Before data collection, each character is given an initial probability of being the target character. In this study, the character probabilities were set with no prior knowledge assumed. However, knowledge about the language could be incorporated. For example, if a "q" has been spelled, then in English it is highly likely it will be followed by a "u". Thus, the initial probabilities could be set based on previously spelled characters and word frequencies. The use of language models has been proposed for P300 spellers [1], and several studies have investigated the incorporation of language models into speller systems with positive results [33], [34].

Another adaptation of the algorithm that might improve its performance would be to calculate posterior probabilities of rows and columns containing the target rather than probabilities for each character. By updating the posterior probabilities on a character basis, the nontarget characters that are flashed with the target character are "incorrectly" updated with a large likelihood of being the target character since there is only one response for the entire set. These characters are also updated "correctly" whenever they are flashed without the target, ultimately resulting in their rejection as targets; however, this technique of updating with both large and small likelihoods of being the target may slow the process of determining the target character. Calculating posterior probabilities based on rows and columns, the correct row and column (those containing the target) would be updated with a large likelihood of containing the target while all other rows and columns would be updated only with small likelihoods of containing the target. The selection of a row and column would then define the target character. This technique was not considered here since it would limit the algorithm to only those paradigms that incorporate an intersecting-set approach to target classification; however, this technique might provide further performance improvements for row/column paradigms.

The information in the algorithm might also be used to improve the speller paradigm itself. Since the posterior probabilities of the characters being the target character are available through the dynamic stopping algorithm, it might be possible to use this information to select the next stimulus to maximize information gain rather than rely on random presentation of stimuli (e.g., [35]). However, converting the paradigm from random stimulus presentation will necessarily require consideration of the factors that influence the elicitation of P300 responses such as the need for the target to be rare and unexpected (e.g., [22] and [36]–[38]).

While the algorithm was demonstrated with a row/column speller paradigm and a SWLDA classifier, the algorithm itself is paradigm- and classifier-independent and could easily be adapted to other decision-based BCI. This study has demonstrated the potential for an algorithm that adapts data collection based on the quality of data being collected rather than relying on assumptions based on participants' prior performance resulting in significant improvements in accuracy and communication rate.

ACKNOWLEDGMENT

The authors would like to thank all of the participants who generously volunteered their time as well as B. Hamner, Dr. K. Morton, and Dr. P. Torrione for their input on this study. The authors would also like to thank two anonymous reviewers for their insightful and helpful comments.

REFERENCES

[1] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, pp. 510–523, 1988.

[2] P. Brunner, S. Joshi, S. Briskin, J. R. Wolpaw, H. Bischof, and G. Schalk, "Does the 'P300' speller depend on eye gaze?," *J. Neural Eng.*, vol. 7, pp. 1–9, 2010.

[3] M. S. Treder and B. Blankertz, "(C)overt attention and visual speller design in an ERP-based brain-computer interface," *Behavioral Brain Functions*, vol. 6, pp. 1–13, 2010.

[4] E. W. Sellers and E. Donchin, "A P300-based brain-computer interface: Initial tests by ALS patients," *Clin. Neurophysiol.*, vol. 117, pp. 538–548, 2006.
[5] E. W. Sellers, A. Kubler, and E. Donchin, "Brain-computer interface research at the University of South Florida cognitive psychophysiology laboratory: The P300 speller," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 3, pp. 221–224, May 2006.
[6] G. Townsend, B. K. LaPallo, C. Boulay, D. J. Krusienski, G. E. Frye, C. K. Hauser, N. E. Schwartz, T. M. Vaughan, J. R. Wolpaw, and E. W. Sellers, "A novel P300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns," *Clin. Neurophysiol.*, vol. 121, pp. 1109–1120, 2010.
[7] B. Z. Allison and J. A. Pineda, "ERPs evoked by different matrix sizes: Implications for a brain computer interface (BCI) system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 1, pp. 110–113, Jan. 2003.
[8] L. Bianchi, S. Sami, A. Hillebrand, I. P. Fawcett, L. R. Quitadamo, and S. Seri, "Which physiological components are more suitable for visual ERP based brain-computer interface? A preliminary MEG/EEG study," *Brain Topogr.*, vol. 23, pp. 180–185, 2010.
[9] V. Bostanov, "BCI competition 2003—Data sets Ib and IIb: Feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram," *IEEE Trans. Biomed. Eng.*, vol. 5, no. 7, pp. 1057–1061, Jul. 2004.
[10] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "BCI competition 2003—Data set IIb: Support vector machines for the P300 speller paradigm," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1073–1076, Jul. 2004.
[11] A. Rakotomamonjy and V. Guigue, "BCI Competition III: Dataset II—Ensemble of SVMs for BCI P300 speller," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 6, pp. 1147–1154, Jun. 2008.
[12] N. Xu, X. Gao, B. Hong, X. Miao, S. Gao, and F. Yang, "BCI competition 2003—Data set IIb: Enhancing P300 wave detection using ICA-based subspace projections for BCI applications," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1067–1072, Jul. 2004.
[13] J. Jin, B. Z. Allison, E. W. Sellers, C. Brunner, P. Horki, X. Wang, and C. Neuper, "An adaptive P300-based control system," *J. Neural Eng.*, vol. 8, pp. 1–14, 2011.
[14] A. Lenhardt, M. Kaper, and H. Ritter, "An adaptive P300-based online brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 1, pp. 121–130, Jan. 2008.
[15] M. Schreuder and M. Tangermann, "Online spelling using the new spatial auditory BCI," in *Proc. Fourth Int. BCI Meeting*, Asilomar, CA, USA, 2010.
[16] H. Serby, E. Yom-Tov, and G. F. Inbar, "An improved P300-based brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 1, pp. 89–98, Jan. 2005.
[17] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the P300 speller," *J. Neural Eng.*, vol. 3, pp. 299–305, 2006.
[18] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain-computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 7, pp. 1034–1043, Jul. 2004.
[19] D. A. Balota, M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman, "The English lexicon project," *Behavior Res. Methods*, vol. 39, pp. 445–459, 2007.
[20] C. Burgess and K. Livesay, "The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis," *Behavior Res. Methods, Instruments Computers*, vol. 30, pp. 272–277, 1998.
[21] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *J. Neurosci. Methods*, vol. 167, pp. 15–21, 2008.
[22] E. W. Sellers, D. J. Krusienski, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A P300 event-related potential brain-computer interface (BCI): The effects of matrix size and interstimulus interval on performance," *Biological Psychol.*, vol. 73, pp. 242–252, 2006.
[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Science/Business Media, 2006.
[24] J. E. Freund, *Mathematical Statistics*, 5th ed. Englewood Cliffs, NJ, USA: Prentice Hall, 1992.
[25] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. vonLuxburg, and G. Ratsch., Eds. New York, NY, USA: Springer, 2004 [Online]. Available: http://www.miketipping.com/papers.htm
[26] D. J. McFarland, W. A. Sarnacki, and J. R. Wolpaw, "Brain-computer interface (BCI) operation: Optimizing information transfer rates," *Biological Psychol.*, vol. 63, pp. 237–251, 2003.
[27] M. Schreuder, J. Hohne, M. S. Treder, B. Blankertz, and M. Tangermann, "Performance optimization of ERP-based BCIs using dynamic stopping," in *Proc. 33rd Annu. Int. Conf. IEEE EMBS*, Boston, MA, USA, 2011.
[28] A. Furdea, S. Halder, D. J. Krusienski, D. Bross, F. Nijboer, N. Birbaumer, and A. Kubler, "An auditory oddball (P300) spelling system for brain-computer interface," *Psychophysiology*, vol. 46, pp. 617–625, 2009.
[29] C. E. Lakey, D. R. Berry, and E. W. Sellers, "Manipulating attention via mindfulness induction improves P300-based brain-computer interface performance," *J. Neural Eng.*, vol. 8, pp. 1–7, 2011.
[30] K. Takano, T. Komatsu, N. Hata, Y. Nakajima, and K. Kansaku, "Visual stimuli for the P300 brain-computer interface: A comparison of white/gray and green/blue flicker matrices," *Clin. Neurophysiol.*, vol. 120, pp. 1562–1566, 2009.
[31] M. Salvaris and F. Sepulveda, "Visual modifications on the P300 speller BCI paradigm," *J. Neural Eng.*, vol. 6, pp. 1–8, 2009.
[32] C. S. Throckmorton, D. B. Ryan, B. Hamner, K. Caves, K. Colwell, E. W. Sellers, and L. M. Collins, "Towards clinically acceptable BCI spellers: Preliminary results for different stimulus selection patterns and pattern recognition techniques," presented at the 4th Int. Meeting Brain Computer Interfaces, Asilomar, CA, USA, 2010.
[33] B. Blankertz, M. Krauledat, G. Dornhege, J. Williamson, R. Murray-Smith, and K.-R. Muller, "A note on brain actuated spelling with the Berlin brain-computer interface," in *Universal Access in Human-Computer Interaction. Ambient Interaction*, C. Stephanidis, Ed. Berlin, Germany: Springer, 2007, vol. 4555, pp. 759–768.
[34] D. B. Ryan, G. E. Frye, G. Townsend, D. R. Berry, S. Mesa-G, N. A. Gates, and E. W. Sellers, "Predictive spelling with a P300-based brain-computer interface: Increasing the rate of communication," *Int. J. Human-Computer Interaction*, vol. 27, pp. 69–84, 2011.
[35] K. Kastella, "Discrimination gain to optimize detection and classification," *IEEE Trans. Syst., Man, Cybern.—Part A: Systems and Humans*, vol. 27, pp. 112–116, 1997.
[36] J. Polich, "Attention, probability, and task demands as determinants of P300 latency from auditory stimuli," *Electroencephalogr. Clin. Neurophysiol.*, vol. 63, pp. 251–259, 1986.
[37] J. Polich, "Task difficulty, probability, and inter-stimulus interval as determinants of P300 from auditory stimuli," *Electroencephalogr. Clin. Neurophysiol.*, vol. 38, pp. 311–320, 1987.
[38] J. Polich, L. Howard, and A. Starr, "Stimulus frequency and masking as determinants of P300 latency in event-related potentials from auditory stimuli," *Biological Psychol.*, vol. 21, pp. 309–318, 1985.

Chandra S. Throckmorton received the B.S. degree in electrical engineering from the University of Texas at Arlington, Arlington, TX, USA, in 1995, and the M.S. and Ph.D. degrees in signal processing from the Electrical Engineering Department, Duke University, Durham, NC, USA, in 1998 and 2001, respectively.

She served as a Postdoctoral Research Associate for two years at Duke University. She was promoted to Research Scientist in 2004 and Senior Research Scientist in 2008. Her research focuses on improving systems through the application of signal processing techniques, pattern recognition, and machine learning. She is currently involved in developing improvements for cochlear implants, brain-computer interfaces, and EEG-based seizure prediction systems.

Dr. Throckmorton is a member of the Society for Neuroscience, the Association for Research in Otolaryngology, the Acoustical Society of America, and the American Auditory Society.

Kenneth A. Colwell (M'11) was born in New Haven, CT, USA, in 1987, and raised in Portland, OR, USA. He received the B.S. degree in engineering physics from Cornell University, Ithaca, NY, USA, in 2009, and the M.S. degree in electrical and computer engineering from Duke University, Durham, NC, USA, in

2012. Currently he is working toward the Ph.D. degree in electrical and computer engineering at Duke University.

He has worked as a Research and Development Intern at the electron microscopy firm FEI Company, and at the National Ignition Facility at Lawrence Livermore National Laboratory in Livermore, CA, USA, researching noninvasive brain-computer interfaces and machine learning.

Mr. Colwell is also a member of the Society for Neuroscience and the American Statistical Association.

David B. Ryan received the B.A. degree in psychology from University of Tennessee, Knoxville, TN USA, in 2006, and the M.A. degree in psychology from East Tennessee State University, Johnson City, TN, USA, in 2011, where he is currently working toward the Ph.D. degree in the Brain-Computer Interface Laboratory.

His research is focused on noninvasive EEG-based communication and the neural correlates of attention and perceptual processing.

Mr. Ryan is a member of the Society of Neuroscience.

Eric W. Sellers received the Ph.D. degree in cognitive and neural science from the University of South Florida, Tampa, FL, USA, in 2004.

He then served as a Postdoctoral Research Fellow for two years in the Laboratory of Neural Injury and Repair, New York State Department of Health's Wadsworth Center, Albany, NY. In 2006, he was promoted to a Research Scientist and named Clinical Director of the Laboratory. He is now the Director of the ETSU Brain-Computer Interface Laboratory and an Associate Professor at East Tennessee State University, Johnson City, TN, USA. His research is focused on developing EEG-based communication systems for severely disabled people using methods broadly adapted from the areas of cognitive psychophysiology and attention. His basic research interests lie in the areas of high-level cognitive function.

Dr. Sellers is a member of the Society for Neuroscience, Electroencephalography and Clinical Neuroscience Society, and the Association for Psychological Science.

Leslie M. Collins (M'96–SM'01) was born in Raleigh, NC, USA. She received the B.S.E.E. degree from the University of Kentucky, Lexington, and the M.S.E.E. and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, MI, USA.

She was a Senior Engineer with the Westinghouse Research and Development Center, Pittsburgh, PA, USA, from 1986 to 1990. In 1995, she became an Assistant Professor in the Department of Electrical and Computer Engineering (ECE), Duke University, Durham, NC, USA, and became an Associate Professor in 2002 and has been a Professor in the ECE Department since 2007. Her current research interests include incorporating physics-based models into statistical signal processing algorithms, and she is pursuing applications in subsurface sensing, brain computer interfaces, as well as enhancing speech understanding by hearing-impaired individuals.

Dr. Collins is a member of the Tau Beta Pi, Eta Kappa Nu, and Sigma Xi honor societies.