J. Neural Eng. **10** (2013) 026014 (11pp) doi:10.1088/1741-2560/10/2/026014

A study of the existing problems of estimating the information transfer rate in online brain–computer interfaces

Peng Yuan[1], Xiaorong Gao[1], Brendan Allison[2], Yijun Wang[3], Guangyu Bin[1] and Shangkai Gao[1]

[1] Department of Biomedical Engineering, Tsinghua University, Beijing, People's Republic of China
[2] Cognitive Science Department, University of California San Diego, San Diego, USA
[3] Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego, San Diego, USA

E-mail: gsk-dea@tsinghua.edu.cn

Received 8 August 2012
Accepted for publication 11 January 2013
Published 28 February 2013
Online at stacks.iop.org/JNE/10/026014

Abstract

Objective. Today, the brain–computer interface (BCI) community lacks a standard method to evaluate an online BCI's performance. Even the most commonly used metric, the information transfer rate (ITR), is often reported differently, even incorrectly, in many papers, which is not conducive to BCI research. This paper aims to point out many of the existing problems and give some suggestions and methods to overcome these problems. *Approach*. First, the preconditions inherent in ITR calculation based on Wolpaw's definition are summarized and several incorrect ITR calculations, which go against the preconditions, are indicated. Then, the issues affecting ITR estimation during the test of online BCI systems are discussed in detail. Finally, a task-oriented online BCI test platform was proposed, which may help BCI evaluations in real-world applications. *Main results*. The guidelines for ITR calculation in online BCIs testing are proposed. The platform executed in the Beijing BCI Competition 2010 shows that it can be used as a common way to compare the online performances (including the ITR) of existing BCI paradigms. *Significance:* The proposed guidelines and task-oriented test platform may reduce the uncertainty and artifacts of online BCI performance evaluation; they provide a relatively objective way to compare different BCI's performances in real-world BCI applications, which is a forward step toward developing standards for BCI performance evaluation.

(Some figures may appear in colour only in the online journal)

1. Introduction

A variety of metrics have been proposed to evaluate the performance of brain–computer interface (BCI) systems, such as classification accuracy, Cohen's Kappa, sensitivity and specificity, positive and negative predictive value, information transfer rate (ITR), the efficiency and the utility (Billinger *et al* 2013, McFarland and Krusienski 2012, Schlögl *et al* 2007, Bianchi *et al* 2007, Quitadamo *et al* 2012, Dal Seno *et al* 2010). The ITR has been the most commonly applied metric to assess the overall performance of BCIs (McFarland and Krusienski 2012). The most popular method for ITR calculation in BCI research was defined by Wolpaw *et al* in 1998, which is a simplified computational model based on Shannon channel theory under several assumptions (Wolpaw *et al* 1998 and 2002, Shannon and Weaver 1964, Pierce 1980, Allison 2010):

$$B = \log_2 N + P \log_2 P + (1 - P)\log_2[(1 - P)/(N - 1)] \tag{1}$$

where B is the ITR in bit rate (bits/symbol), N is the number of possible choices and P is the probability that the desired choice

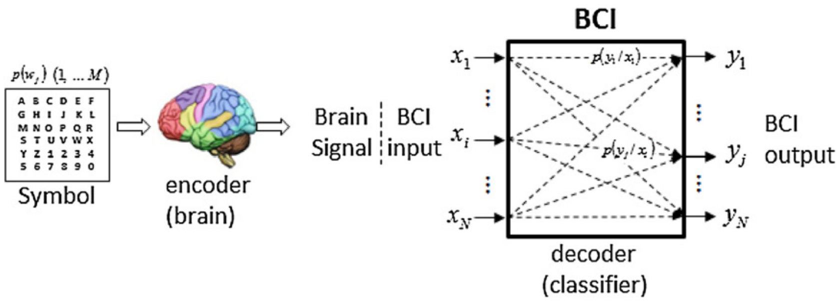

Figure 1. The model of BCI information transfer. Each of the symbols can be encoded to a pattern of brain signals, which are the BCI inputs and will be decoded by BCI. M is the total number of symbols. $p\,(w_i)$ $(i = 1, 2, \ldots M)$ is the probability of ith symbol to be selected. x_i $(i = 1, 2, \ldots N)$ is the ith input. In general, M is equal to N. But, if the system has extra input (e.g. idle state), N will be larger than M. y_i $(i = 1, 2, \ldots N)$ is the ith output of BCI. $p\left(y_j/x_i\right)$ is the probability that the ith input is recognized as jth input. When i is equal to j, the output is correct. $p\,(y_i/x_i)$ is the classification accuracy.

will be selected, also called target identification accuracy or classifier accuracy.

Generally, B_t in bits/min is used to indicate the BCI ITR

$$B_t = B * (60/T) \tag{2}$$

where T (seconds/symbol) is the time needed to convey each symbol.

Sometimes the ITR is calculated based on offline analysis (Meinicke *et al* 2002). However, even if the ITR is correctly calculated based on offline data, it may have little bearing on online performance in field settings (McFarland *et al* 2003). Instability over time, noise sources and distraction from feedback or real world events may impair performance in online operation. Thus, the gold standard for evaluating BCIs is the effectiveness in real-time, closed-loop online performance (McFarland and Krusienski 2012). This paper focuses on the problems of ITR estimation in online BCIs.

The following problems with online BCIs' ITR estimation still exist:

(1) Equation (1) is valid under several strict assumptions. Unfortunately, those required preconditions are sometime ignored in literature, which leads to incorrect ITR estimation.
(2) During online tests, some factors (e.g. the small number of test trials and the time that users need to shift between targets) have an effect on estimating parameters P and T, and hence on estimating the ITR. These details have not been adequately addressed in many articles that estimate the ITR in online BCIs (see, for example, comments in Sellers and Donchin (2006) and Allison and Neuper (2010)).
(3) A general test platform for online BCI performance evaluation (including ITR estimation) is lacking. The platform should be effective for the online implementation of different BCI paradigms.

This paper aims to solve the above problems. In this paper, first the preconditions of using equation (1) for ITR calculation are summarized and the types of BCIs that cannot use equation (1) for ITR calculation directly are discussed. Second, a number of factors that can affect parameter estimation during online tests are analyzed, leading to a proposed guideline for online parameter (P and T) estimation. Ultimately, based on these issues, a task-oriented BCI test platform was developed and used in the Beijing BCI Competition 2010. This platform can compare different parameters (including the ITR) of different BCIs in online operation.

2. ITR calculation using Wolpaw's definition

2.1. Preconditions

The model of BCI information transfer is illustrated in figure 1.

A number of papers have discussed BCI ITR calculation based on equation (1) (Wolpaw *et al* 1998, 2002, Kronegg *et al* 2003, 2005, Fatourechi e*t al* 2006). The preconditions of using equation (1) are summarized as follows.

(1) BCI systems are memory-less and stable discrete transmission channels.
(2) All the output commands are equally likely to be selected $(p\,(w_i) = 1/N)$.
(3) The classification accuracy is the same for all the target symbols $(p(y_i|x_i) = p(y_j|x_j))$.
(4) The classification error is equally distributed among all the remaining symbols $(p(y_j|x_i)_{j\neq i} = (1 - p(y_i|x_i))/(N - 1))$.

Actually, precondition (1) is the basic one and precondition (2) suggests that BCI systems do not consider an idle state $(N = M)$, because the probability of selecting an idle state is not the same as selecting other symbols. In addition, to ensure that the ITR increases monotonously with P, the classifier accuracy P should be above the chance level. Normally, most BCIs meet these conditions in practice.

The important preconditions listed above should not be ignored before calculating the ITR using Wolpaw's definition. In the following section, the problems involving ITR calculation using Wolpaw's definition in different types of online BCI systems were discussed in detail.

2.2. Problems involving ITR calculation using Wolpaw's definition in online BCI systems

2.2.1. Synchronous BCI. In synchronous BCI systems, the timing of the BCI operation is determined by the system. The BCI provides cues that instruct the user when to choose a target character, when to perform mental tasks to send a message or command and perhaps when to rest or perform other actions (Birbaumer *et al* 1999, Wolpaw *et al* 2002, Pfurtscheller and Neuper 2001, Boostani *et al* 2007, Bin *et al* 2011). Some synchronous BCIs may nonetheless allow different numbers of selections per minute, such as when variable numbers of trials are averaged together to identify target characters (Jin *et al* 2011). According to the preconditions above, synchronous BCIs can use equation (1) for ITR calculation.

However, in *ITR* calculation for online synchronous BCIs, some uncertainty affects the estimation of parameters in equation (1) (e.g. the number of test trials affects the estimation of P, and the target shifting time affects the estimation of T). In section 3, we will discuss the issues that affect parameter estimation in online synchronous BCIs and propose some guidelines.

2.2.2. Asynchronous BCI. Many BCIs operate in asynchronous (or self-paced) mode (Townsend *et al* 2004, Birch *et al* 2002, Fatourechi *et al* 2008, Millán and Mouriño 2003, Scherer *et al* 2007, Mason and Birch 2000, Roberts and Penny 2000, Tsui *et al* 2009, Krauledat *et al* 2004). In this mode, users can choose to control BCIs whenever they want. The timing of system operation, including the number of selections per minute, may vary dramatically depending on the user. Also, in an asynchronous BCI, users may choose not to send any messages or commands (an idle state) for long periods. Any message or command sent during such periods reflects a false positive. Hence, the probability of selecting the idle state may be very different from the probability of selecting specific commands. Therefore, asynchronous BCIs do not meet the preconditions of equation (1) (against precondition (2)).

Theoretically, the ITR for asynchronous BCIs can be calculated using general equations of the mutual information (Townsend *et al* 2004, Birch *et al* 2002, Fatourechi *et al* 2008, Millán and Mouriño 2003, Scherer *et al* 2007, Mason and Birch 2000, Roberts and Penny 2000, Tsui *et al* 2009, Krauledat *et al* 2004). However, in practice, it is difficult to know the prior probability and the information transfer matrix exactly. Asynchronous BCIs often report performance without using an ITR at all. For example, authors might report the time required to complete a sequence of actions, such as navigating through a virtual environment (Scherer *et al* 2008). Ideally, measures of asynchronous BCI performance should include a non-control state to evaluate how well a BCI system can "sleep" when users do not want to use it (Ortner *et al* 2011). Indeed, recent surveys of severely disabled BCI users have confirmed that an effective "standby" or "sleep" mode is very important (Huggins *et al* 2011, Blain-Moraes *et al* 2012).

2.2.3. Special types of BCIs. One of the assumptions in Wolpaw's ITR calculation is that the BCI systems are memory-less and stable discrete transmission channels. However, this is not always the case. In some memory BCI system, the output at any time is not just related to the input at that time, but also to prior inputs and outputs (e.g. Volosyak 2011). BCIs might also provide different selections based on prior selections, such as only presenting letters that can legally follow preceding letters in that language (Wills and Mackay 2006). The statistical properties of the transfer channels in these BCIs may change over time. All these types of BCIs (hereafter referred to as non-stable BCI) may achieve high performance, however it is not valid to use equation (1) for ITR calculation without appropriate modification (as they violate precondition (1)).

3. Guidelines for parameter estimation in online synchronous BCIs

In online synchronous BCIs' ITR calculation, the critical issue is to determine three parameters: the target identification accuracy (P), the time needed to output a symbol (T) and the total number of optional symbols (N). Normally, N is obvious in a system. The estimation of P requires an online test. T may be fixed, such as in classical P300 BCIs (e.g., Farwell and Donchin 1988), or may require testing, such as if the system continues monitoring the user's brain activity until reaching an adequate accuracy threshold (Gao *et al* 2003, Jin *et al* 2011).

ITR calculation is based on each selection that conveys meaning, such as a letter, symbol or wheelchair movement command. Figure 2 presents the general process of online BCI testing.

After online testing, the classification accuracy P can be estimated using the following formula

$$P = \frac{x}{n} \tag{3}$$

where n is the total number of test trials and x is the number of correct trials.

The estimated time T to output one symbol can be found by calculating the average time for each output symbol, as illustrated in equation (4).

$$T = \frac{t}{n} \tag{4}$$

However, during online tests, some uncertainty affects the estimation of these parameters (e.g. the number of test trials affects the estimation of P and the target shifting time affects the estimation of T).

In what follows, the issues affecting the estimation of the parameters and the principles of dealing with the issues will be discussed in detail.

3.1. Error analysis

3.1.1. The relationship between the error of the ITR and the error of P. The relationship between the ITR error ΔB_t and the error of classification accuracy ΔP is illustrated in equation (5)

$$\Delta B_t = \frac{60}{T} \cdot \log_2 \frac{P(N-1)}{1-P} \cdot \Delta P \tag{5}$$

As P, N and $1/T$ increase, the error of the ITR (ΔB_t) will become more and more sensitive to the error of ΔP. This means that the same error of P (e.g. $\Delta P = +0.05$) will have a greater impact on the ITR error as P, N and especially $1/T$ increase (see figure 3).

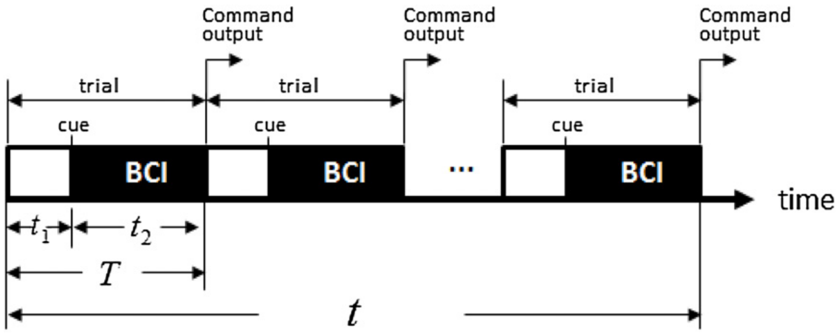

Figure 2. BCI ITR online communication components. t is the total time to send a complete message or command sequence and T is the period to output each symbol. t_1 is the pre-cue time, which is the time period from the end of the previous trial to the on-set of a new cue. During $t_{1,}$ subjects need to prepare for target identification and shift between target symbols. Typically, the brain activity during t_1, is ignored. t_2 is the time for BCI operation including brain signal analysis and command output.

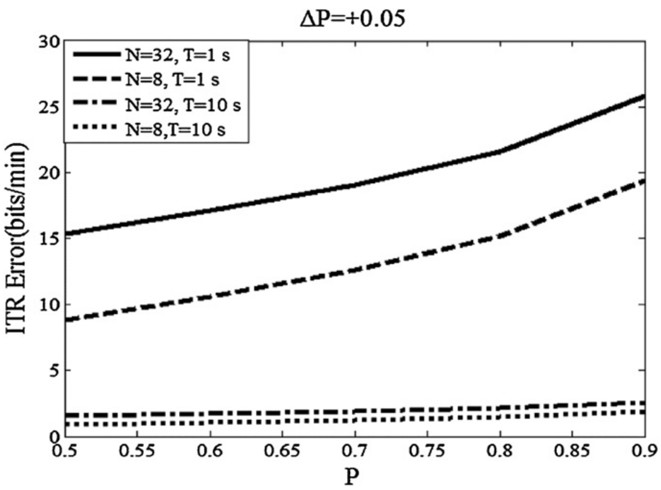

Figure 3. Error of the ITR across different P when the error of P is $+0.05$.

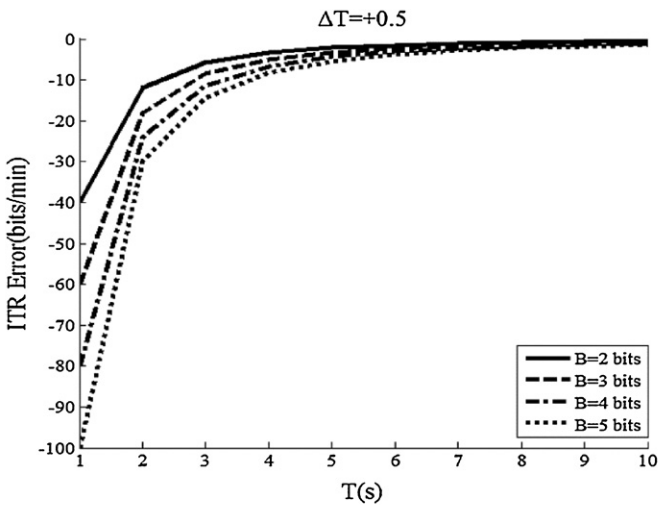

Figure 4. Error of the ITR across different T when the error of T is $+0.5$.

3.1.2. The relationship between the error of the ITR and the error of T. The relationship between the ITR error ΔB_t and the error of T (ΔT in equation (6)) is illustrated in equation (6)

$$\Delta B_t = \frac{-60}{T^2} \cdot B \cdot \Delta T \tag{6}$$

As B and $1/T$ increase, the error of the ITR(ΔB_t) will become more and more sensitive to ΔT. This means that the same error of T (e.g. $\Delta T = +0.5$) will have a greater impact on the error of the ITR as B and $1/T$ (especially when $1/T$ is above 1/5) increase (see figure 4).

Based on the analysis above, we have the following suggestion:

Suggestion 1. When reporting the ITR, N, P and T should be explicitly identified. As P, N and $1/T$ increase, the estimated accuracy of P and T should merit more attention to ensure accurate calculation.

Actually, during online tests, the number of test trials will affect the estimated accuracy of P and the time for switching between two target symbols will affect the estimation of T. Henceforth, these issues will be discussed.

3.2. P

3.2.1. The number of test trials. To effectively estimate chance performance, the number of input symbols must be adequate (Müller-Putz *et al* 2008); without enough input symbols, ITR estimation is not valid (Billinger *et al* 2013).

Accurate estimation of the classification accuracy (P) relies on a large number of test trials. However, it is impossible to input infinite samples. How many test trials are adequate? Answering this question requires assessing the relationship between the number of test trials and the estimated accuracy of classification accuracy (P).

This problem can be abstracted into an estimation of a parameter in a binomial distribution. Consider the Binomial distribution as follows:

$$x \sim B(n, P) \quad (n \geqslant 1, 0 < P < 1) \tag{7}$$

where x is the number of correct trials during tests, n is the total number of test trials and P is the real classification accuracy. Hence, $\frac{x}{n}$ represents the estimated classification accuracy.

From the confidence interval point of view, when estimating classification accuracy P, in order to ensure that the width of the confidence interval at the 1-α level is no more

Table 1. n_0 across different L and $\frac{x}{n}$.

	L		
$\frac{x}{n}$	0.2	0.1	0.02
0.5	93	381	9601
0.6	89	366	9217
0.7	78	320	8065
0.8	60	245	6145
0.9	37	141	3461

than L, the minimum number (n_0) of input symbols is as shown in equation (8) (appendix A).

$$n_0 = \frac{z_{\alpha/2}^2}{L^2} * \left[\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right] + \sqrt{\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right]^2 + L^2 \cdot (1 - L^2)}\right] \quad (8)$$

where $\frac{x}{n}$ represents the estimated classification accuracy.

If $\alpha = 0.05$, the confidence level is 0.95. For different L and $\frac{x}{n}$, the corresponding n_0 is as listed in table 1.

Based on the analysis above, a suggestion is given as follows.

Suggestion 2. To ensure an accurate estimation of classifier accuracy, enough test trials are needed. Hence, when the ITR is reported, the number of test trials should also be reported.

Fortunately, when P is above 0.5, the required number (n_0) of the test trials decreases monotonously with P (appendix B). However, to ensure an accurate estimation of P, the required number (n_0) still needs to be considered.

3.2.2. Error correction. According to the preconditions of using equations (1) and (3), the error symbols during the input process should not be corrected.

Similarly, a proper estimate of the ITR should not incorporate software tools that can increase effective throughput, such as error correction, word completion or goal-directed behavior (Allison *et al* 2007, Cincotti *et al* 2008, Allison 2010, Jin *et al* 2011). Any such tools should be described in adequate detail. If desired, the additional ITR estimate methods or other metrics could be generated that do account for such tools (Ferrez and Millán 2005, Bianchi *et al* 2007, Quitadamo *et al* 2012, Dal Seno *et al* 2010).

Suggestion 3. Authors should include an ITR estimation that does not include error correction or other methods to increase effective throughput. If a system does employ error correction, authors should adequately describe the methods and results and, if desired, include a modified ITR as well.

3.2.3. The occurrence probability of input symbols. According to precondition (2), during an online test, the occurrence probabilities of the input symbols should be the same ($p(w_i) = 1/N$). Therefore, to ensure that each input symbol is equally likely to be selected, BCIs should better be tested with randomly generated symbols from all N symbols. If the optional symbols do not share the same probability of being selected, a modified formula should be developed to calculate the ITR.

Suggestion 4. To ensure that each input symbol is equally likely to be selected, BCIs should ideally be tested with randomly generated symbols from all N symbols.

3.3. T

The timing of a BCI operation is determined by the system in synchronous BCIs. However, as shown in figure 3, a pre-cue time (t_1) is always needed so that subjects can prepare for target identification and shift between targets. In practice, t_1 could be thought of as either a part of a BCI operation or not. The inclusion of t_1 can substantially influence T, especially when t_2 is short (e.g. $t_2 < 5$ s); estimating the ITR becomes quite different between these two cases (see figure 4). The ITR calculation with t_1 reflects the comprehensive performance of the BCI, including the subject's effectiveness for the system, while the ITR calculation without t_1 reflects a hypothetical BCI performance. Some articles estimate the ITR both with and without t_1 (Townsend *et al* 2010, Jin *et al* 2011).

Suggestion 5. When reporting ITRs, authors should explain all of the factors in the ITR calculation, such as whether t_1 is included. Reporting different values of ITR is acceptable if this principle is maintained, which would allow readers to compare ITRs more effectively across different groups and calculation methods.

3.4. N

According to Wolpaw's definition, N is the number of users' possible selections, which is the same as the number of possible outputs in synchronous BCIs. In order to meet the requirement of precondition (1), N should remain constant during BCI operation.

In addition, some BCIs (such as menu-based BCIs or multiple-step decision BCIs) face one decision: whether N is determined as the number of possible selections in each decision-making step or as the total number of users' possible selections (the total number of possible outputs by BCIs).

We posit that BCIs can be thought of as a black box. From the whole system angle, for such BCIs, N can be determined as the number of users' total possible selections instead of the number of possible selections in each decision-making step, as long as they meet the requirement of the preconditions of equation (1). This view is consistent with our comment at the beginning of section 3 and supported by others (McFarland and Krusienski 2012): N should be based on the number of end selections, or meaningful outputs, rather than the stages necessary to get there.

Suggestion 6. N should remain constant throughout the whole test.

3.5. Subjects

BCI performance (including the ITR) varies across subjects. However, in some papers the ITR was based on an elite subset of subjects who performed well (Gao *et al* 2003, Billinger *et al*

2013), which obviously could not reflect the BCI performance across a large population.

To objectively reflect BCI performance across different subjects, the following suggestion is given.

Suggestion 7. Results should be presented from each subject tested, including individual ITRs and statistical results. If any data were rejected from further analyses, the amount of data and the reason(s) for rejection should be described. If results are presented from subject(s) who were exceptional, this fact should be noted.

4. A platform for online BCIs performance testing

4.1. Necessity of a test platform

It is well believed that the ultimate test of any BCI is how it performs in actual online operation (McFarland and Krusienski 2012). As we discussed above, there are many issues affecting the ITR estimation of online BCIs. Different papers report different ways to calculate it. A general platform which is effective for the online implementation of different BCI paradigms may help to reduce the uncertainty and artifacts and provide a relatively objective way to compare different BCIs' online performances. In addition, imposing different tasks which simulate the applications in everyday life should be valuable for evaluating real-world BCI applications. Based on the above considerations, we developed a task-oriented test platform which is intended for the public to benefit the BCI community.

4.2. Overview of the test platform

This platform was successfully implemented during the Beijing BCI Competition 2010 hosted by our lab in Tsinghua University (Supported by the National Nature Science foundation of China). Thirty-five teams from 17 universities participated in this competition. This was an online competition of different BCI systems with the proposed general test platform.

To evaluate the performance of online BCIs in real applications, this BCI test platform was designed to be task-oriented. In the competition, the three tasks included: (i) switching control; (ii) character input (typing); and (iii) virtual automobile control. All three tasks were chosen to simulate the real-world BCI applications. The tasks (i) and (iii) will be briefly introduced hereinafter. Task (ii), which is convenient for ITR evaluation, will be discussed in subsection 4.3.

The switching control task is designed in a home environment, as shown in figure 5. There are six switches related to different appliances (TV, DVD, lamp, curtain, door and air conditioner labeled from '1' to '6'). The system accepts six commands from '1' to '6' for the corresponding switch control. The individual switch will change the ON/OFF state once the system receives the corresponding commands. The participants have to turn on all the devices. The winner is the participant who takes the least time to complete the task.

The virtual automobile control task is designed to test the device control abilities with BCIs. Participants are asked to control the virtual automobile by adjusting its speed and direction to pass the stations from '1' to '6' sequentially. Figure 6 shows an example of the route (the sequential destinations). When the automobile passes a destination with a proper speed, the participant gets 50 points. The task is limited to 5 min. If the participants finish the test within 5 min, they earn one point for each remaining second. The winner is the one who gets the highest score.

The software for this general platform can be downloaded for free from the website: http://166.111.152.146/bci/Default.aspx.

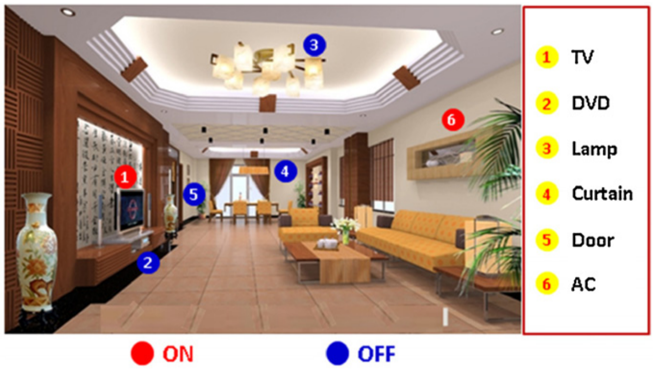

Figure 5. The switching control task. In the home environment, there are 6 switches related to different appliances (TV, DVD, lamp, curtain, door and air conditioner labeled from '1' to '6'). The participants are asked to turn on all the devices.

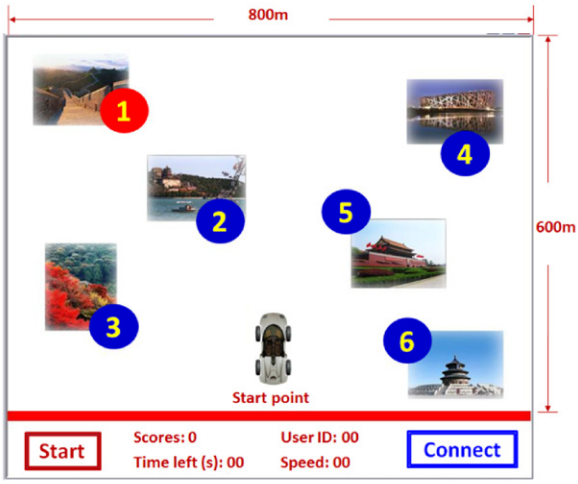

Figure 6. The virtual automobile control task. Participants are asked to control the virtual automobile by adjusting its speed and direction to pass the stations from '1' to '6', sequentially assigned by the test platform, within 5 min.

4.3. The typing task in the test platform

In the character input (typing) task, a long enough random sequence of target symbols, chosen from a vocabulary of total 40 different kinds of symbols (26 letters, 10 digits and 4 punctuation marks), were presented to the subjects, as shown in figure 7. Subjects were asked to input symbols sequentially. Hence, this was a classic copy-spelling task (Farwell and Donchin 1988, Birbaumer *et al* 1999, Bin *et al* 2011, Jin *et al* 2011). The test time duration was six minutes. Subjects

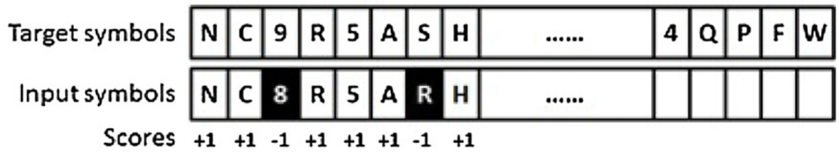

Figure 7. The task (ii): character input. The top row contained a long enough random sequence of target symbols. The symbols in the second row were the input symbols (based on the BCI classifier results). If the input symbol matched the target symbol, the subject earned one point. Otherwise, the subject lost one point.

Table 2. Test results of the typing task in BCI Competition 2010.

Team	Amplifier	Type	Paradigm	P (%)	T (sec/sym)	Score	ITR (bits/min)
1	Neurosan-40	synchronous	P300	98.61	5	70	61.7
2	BrainProducts	synchronous	P300	95.92	7.34	45	39.7
3	Biosemi	synchronous	Motion	82	7.2	32	30.8
4	Neurosan-40	synchronous	P300	85.71	8.57	30	27.8
5	TsinghuaMiPower	synchronous	SSVEP	80.49	8.78	25	24.5
6	TsinghuaMiPower	synchronous	SSVEP	87.88	10.9	25	23.8
7	G-Tec	synchronous	SSVEP	55.32	7.66	5	15.4
8	SYMTOP	synchronous	P300	56.67	12	4	10.2

could use any type of BCI to complete the task. Actually, the competition encouraged a variety of different approaches (such as P300 or SSVEP), feature selection methods, classification techniques, etc. The subject's BCI system sent the code reflecting the chosen symbol to the server of a test platform through TCP/IP.

A number of metrics can be used to evaluate the performance of online BCIs, including P, T, ITR, scores (e.g. subjects were awarded one point if the symbol selected by the BCI matched the target symbol and lost one point if they did not match) and so on.

This test platform has considered the details in online parameters estimation discussed above. In summary, it has the following advantages:

(1) After a 6 min test, the number of test trials is determined, so the confidence level of P can be calculated. For fast BCIs, the test time is long enough to ensure that the task involves a relatively large number of trials, which allows a relatively accurate estimate of P and the ITR.
(2) The time t_1 for switching between two target symbols is explicitly included in T.
(3) The target symbols are randomly generated so that $p(w_i) = 1/N$ is valid.
(4) No correction is allowed during the online test, which is critical for a proper estimation of classification accuracy.
(5) It is a task-oriented test platform supporting tests for different online BCI paradigms. For synchronous BCIs, the ITR can be calculated using equation (1). For other online BCIs (such as asynchronous BCI) whose ITR cannot be calculated using equation (1), the scores they get in the task can be seen as a way to evaluate their performance from a practical perspective. Further, it can be proved that the scores positively correlate with the speed and classification accuracy of BCIs, and hence with the ITR (appendix C).

In the Beijing BCI Competition 2010, eight teams from different institutions participated in the typing task competition. Subjects tried the typing task with various types of BCI approaches, including different P300, SSVEP and motion-VEP BCIs. The results are illustrated in table 2.

The platform is flexible in several aspects. The parameters of the platform can be adjusted to test the performance of BCIs according to a variety of metrics. First, by adjusting the length of the test time, we can test the online performance of BCIs across time. Second, the number of optional input symbols can be adjusted according to the demands of different tasks. In the competition, from a practical perspective, N is set at 40 for all the BCIs in the character input task, so BCIs have to adopt measures to complete the task. However, in this situation, the BCIs may not achieve their best performance. Hence, strictly speaking, to evaluate the best performance (or the ITR) of BCIs, it is better that N should be adjusted according to each BCI's demand. Third, according to different tasks, the ratio of the awarded points (when the input is correct) and the lost points (when the input is wrong) can also be adjusted to evaluate the performance of BCIs in special situations. For example, if errors are very problematic, the number of points lost for each error could be increased.

5. Discussion

5.1. Theoretical calculation of the ITR

As discussed above, asynchronous BCIs and non-stable BCIs cannot use equation (1) for ITR calculation directly. Among them, the *ITR* for asynchronous BCIs can be calculated using general equations based on mutual information (Nykopp 2001, Kronegg *et al* 2005, Fatourechi *et al* 2006). However, in practice, it is difficult to know the prior probability $(p(w_i))$ and the element of the information transfer matrix $p(y_j/x_i)$ exactly.

For non-stable BCIs, the property of the transfer channel may be more complex. Hence, ITR calculation will be more difficult in practice.

Actually, as we discussed in section 2, equation (1) is based on some preconditions. However, some preconditions (e.g. preconditions (3) and (4)) cannot be strictly accorded to the fact. Hence, whatever the parameter's estimation accuracy a is, the calculated BCI ITR using equation (1) is an approximation of the truth.

5.2. Comprehensive evaluation of BCI

Accuracy versus ITR: some users may prefer a system that is highly accurate over one that maximizes the ITR (Wolpaw *et al* 2002; Billinger *et al* 2013). Therefore, an 'improvement' to the system that allows more selections and a higher overall ITR at the expense of reduced accuracy may annoy the user. On the other hand, if new BCI systems feature improved tools for error correction, people might not mind a lower accuracy because many errors are corrected later. In most BCI systems, it is relatively easy for trained experts to modify one or more of the parameters that influence the ITR. Ideally, however, BCI systems should be flexible to allow the user and any caretakers to easily modify relevant parameters without expert help.

The efficiency and the utility: in 2007 and 2010, two new metrics were proposed, the efficiency (Bianchi *et al* 2007, Quitadamo *et al* 2012) and the utility (Dal Seno *et al* 2010) that emphasize the contribution of the control interface (Mason and Birch 2003) and the final application of the system. These task-oriented metrics are suitable for evaluating the overall performance of a BCI system with error correction strategies and identify optimal parameters as well as operating settings.

Hence, the ITR is only one of many factors relevant to BCI evaluation. There are dozens of factors that could influence a decision about which BCI system is better overall (Allison 2010). These factors may vary substantially across different users, BCIs and situations. Any comprehensive evaluation of BCIs should assess many other aspects such as cost, the need for outside support, invasiveness, training time, ease of use, comfort, etc.

However, the BCI community seriously lacks a common way for defining the performance of BCI systems and, even within the same metric, different papers report different ways to calculate it. To overcome this problem, first a common language for communication is needed. A clear set of definitions that define each entity of a BCI may be very helpful in this regard (Mason and Birch 2003). Second, an open data set (such as the BCI competition data set) is needed, which can be extremely useful in comparing different models and different feature selection methods as offline evaluations (McFarland and Krusienski 2012). Third, a general online test platform for online BCIs performance evaluation is needed, which would be very helpful to reduce the uncertainty and artifacts; this would provide a common and relatively objective way to compare BCI performance across different real-world applications.

5.3. Practical value of the platform

As discussed in 5.1 and 5.2, two problems exist: (i) a theoretical calculation of the *ITR* without error is almost impossible in practice; and (ii) a comprehensive evaluation of BCIs involves a lot of factors and lacks standardizations.

The test platform we developed aims to help solve the above two problems. First, besides the ITR, the task-oriented test platform emphasizes the practical value of BCIs, which is consisted with other papers' views (Bianchi *et al* 2007, Quitadamo *et al* 2012, Dal Seno *et al* 2010). It allows an evaluation of different online BCIs from the practical perspective, which is especially useful for BCIs where the ITR cannot be calculated using equation (1). Second, the platform is flexible in several aspects. The parameters of the platform can be further adjusted to test BCI performance according to a variety of metrics, which is helpful for a comprehensive evaluation of a BCI. In addition, this test platform can be used as a research platform to study the problems in online BCIs during practical application (e.g. the trend of online BCIs' performance and the change of the user's brain state across time can be studied by adjusting the length of the test time).

Certainly, the current platform may need further improvement. We encourage the researchers in the BCI community to use this platform and would appreciate any suggestions to improve the platform.

6. Conclusion

In summary, this paper addresses the issues critical to objectively understanding the ITR and describes objective methods for its estimation in online BCIs. Many issues affect ITR calculation, which are often disregarded, and many groups use different methods. Hence, when calculating the ITR, we urge authors to make a thorough and informative estimation, further they should describe all the conditions under which the ITR is calculated. Authors may wish to provide different measures of the ITR to facilitate comparisons across studies and groups. In addition, by introducing a task-oriented test platform that is effective for the online implementation of different BCI paradigms, this paper provided a relatively objective way to compare different BCIs' online performance to reduce the uncertainty and artifacts and emphasized the importance of evaluating performance (including the ITR) of online BCIs from the practical perspective.

More generally, we encourage our colleagues in the BCI community to work together to agree on standards for reporting BCI performance and other facets of BCIs. These standards could include terms, definitions, guidelines, methods and models to describe the BCI systems and comparison metrics. Such standards could facilitate effective reporting and a comparison across groups, which would help newcomers in BCI research who may be confused by the myriad of different reporting approaches across groups. Developing such standards may require significant discussion and compromise, perhaps mediated through a BCI Society and/or workshops or other events (Allison 2011; Allison *et al* 2013; see future-bnci.org).

Acknowledgments

This work is supported by the National Natural Science Foundation of China under (no. 90820304) and the National

Basic Research Program (973) of China (no. 2011 CB933204). We would like to thank Mr Zhikai Chen for his help to create the Beijing BCI competition 2010 website.

Appendix A. The relationship between the estimated accuracy of *P* and the number of test trials

Consider the Binomial distribution as follows:

$$x \sim B(n, P) \quad (n \geqslant 1, 0 < P < 1)$$

where x is the number of correct trials during test, n is the total number of test trials and P is the real classification accuracy of BCIs. Hence, $\frac{x}{n}$ represents the estimated classification accuracy.

From the confidence interval point of view, If the width of confidence interval of $\frac{x}{n}$ at the level $1–\alpha$ is W, it follows that

$$W = \frac{\sqrt{\left(2 \cdot n \cdot \frac{x}{n} + z_{\alpha/2}^2\right)^2 - 4 \cdot \left(n + z_{\alpha/2}^2\right) \cdot n \cdot \left(\frac{x}{n}\right)^2}}{\left(n + z_{\alpha/2}^2\right)}$$

To ensure that W is no more than L, i.e.

$$W = \frac{\sqrt{\left(2 \cdot n \cdot \frac{x}{n} + z_{\alpha/2}^2\right)^2 - 4 \cdot \left(n + z_{\alpha/2}^2\right) \cdot n \cdot \left(\frac{x}{n}\right)^2}}{\left(n + z_{\alpha/2}^2\right)} \leqslant L$$

It follows that

$$n \geqslant \frac{z_{\alpha/2}^2}{L^2} * \left[\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right] + \sqrt{\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right]^2 + L^2 \cdot (1 - L^2)}\right]$$

Let

$$n_0 = \frac{z_{\alpha/2}^2}{L^2} * \left[\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right] + \sqrt{\left[\left[2 \cdot \frac{x}{n} \cdot \left(1 - \frac{x}{n}\right)\right] - L^2\right]^2 + L^2 \cdot (1 - L^2)}\right]$$

So, we get $n \geqslant n_0, n_0 \in Z, n \in Z$. For example, e.g. when $\alpha = 0.05, z_{\alpha/2} = 1.96, L = 0.2, \frac{x}{n} = 0.8$, then, $n_0 = 60$

Appendix B. The relationship between n_0 and *P*

From confidence interval point of view, let

$$t = \frac{x}{n}$$

$$y = \frac{z_{\alpha/2}^2}{L^2} * [[[2 \cdot t \cdot (1 - t)] - L^2] + \sqrt{[[2 \cdot t \cdot (1 - t)] - L^2]^2 + L^2 \cdot (1 - L^2)}]$$

And

$$n_0 = \lceil y \rceil$$

It follows that

$$\frac{\partial y}{\partial t} = \frac{z_{\alpha/2}^2}{L^2} \cdot (-4t + 2) \cdot \left[1 + \frac{[[2t(1 - t)] - L^2]}{\sqrt{[[2t(1 - t)] - L^2]^2 + L^2(1 - L^2)}}\right]$$

When

$$t \in (0.5, 1)$$

then

$$\frac{\partial y}{\partial t} < 0$$

y and n_0 decrease monotonously with t.

When

$$t \in (0, 0.5)$$

then

$$\frac{\partial y}{\partial t} > 0$$

y and n_0 increase monotonously with t.

Appendix C. The relationship between score and BCI's accuracy and speed

The score achieved by the team can be illustrated as follows:

$$s = x \cdot m + y \cdot n \quad (m \geqslant 0, n \geqslant 0, x \geqslant 0, x > y)$$

where x is the points earned when the input is correct, while y is the points earned when the input is wrong. m is the total number of correct inputs during the whole test (6 min), while n is the total number of wrong inputs. Finally, s is the score.

m plus n is the total number of inputs during the whole test (6 min),which is positively correlated with the speed of BCIs. And the accuracy can be illustrated as follows:

$$P = \frac{m}{m + n}(0 \leqslant P \leqslant 1)$$

Then it follows that

$$\begin{aligned} s &= (m + n) \cdot [P \cdot x + (1 - P) \cdot y] \\ &= (m + n) \cdot [(x - y) \cdot P + y] \end{aligned}$$

As the sum of m and n is above zero, if x is greater than y, we can ensure that s is positively correlated with the accuracy P.

If x is greater than y, and if y is above zero, we can ensure that s is positively correlated with the sum of m and n, which is positively correlated with the speed of BCIs.

If x is greater than y, and if y is not above zero, to ensure that s is positively correlated with the speed of BCIs, it follows that

$$P > \frac{y}{y - x} \quad (x \geqslant 0, y < 0)$$

In our platform x is 1, while y is –1. So, when P is above 0.5 (all the teams met this requirement) we can ensure that s is positively correlated with the speed of BCIs.

References

Allison B Z 2010 Toward ubiquitous BCIs *Brain-Computer Interfaces (The Frontiers Collection)* eds B Graimann, G Pfurtscheller and B Allison (Berlin, Heidelberg: Springer) pp 357–87

Allison B Z 2011 Trends in BCI research: progress today, backlash tomorrow *ACM XRDS* **18** 18–22

Allison B Z, Dunne S, Leeb R, Millán Josédel R and Nijholt A 2013 Recent and upcoming BCI progress: overview, analysis, and recommendations towards practical brain-computer interfaces *Biol. Med. Phys. Biomed. Eng.* 1–13

Allison B Z and Neuper C 2010 Could anyone use a BCI? (B+H)CI: The human in brain–computer interfaces and the brain in human–computer interaction *Brain-Computer Interfaces* ed D S Tan and A Nijholt *(Human-Computer Interaction Series)* (London: Springer) pp 35–54

Allison B Z, Wolpaw E W and Wolpaw J R 2007 Brain computer interface systems: progress and prospects *Expert Rev. Med. Devices* **4** 463–74

Bianchi L, Quitadamo L R, Garreffa G, Cardarilli G C and Marciani M G 2007 Performances evaluation and optimization of brain computer interface systems in a copy spelling task *IEEE Trans. Neural Syst. Rehabil. Eng.* **15** 207–16

Billinger M, Daly I, Kaiser V, Jin J, Allison B Z, Müller-Putz G R and Brunner R 2013 Is it significant? Guidelines for reporting BCI Performance *Toward Practical Brain-Computer Interfaces: Biological and Medical Physics, Biomedical Engineering* (Berlin: Springer) chapter 17 pp 333–54

Bin G, Gao X, Wang Y, Li Y, Hong B and Gao S 2011 A high-speed BCI based on code modulation VEP *J. Neural Eng.* **8** 025015

Birbaumer N, Ghanayim N, Hinterberger T, Iversen I, Kotchoubey B, Kübler A, Perelmouter J, Taub E and Flor H 1999 A spelling device for the paralyzed *Nature* **398** 297–8

Birch G E, Bozorgzadeh Z and Mason S G 2002 Initial on-line evaluations of the LF-ASD brain–computer interface with ablebodied and spinal-cord subjects using imagined voluntary motor potentials *IEEE Trans. Neural Syst. Rehabil. Eng.* **10** 219–24

Blain-Moraes S, Schaff R, Gruis K L, Huggins J E and Wren P A 2012 Barriers to and mediators of brain–computer interface user acceptance: focus group findings *Ergonomics* **55** 516–25

Boostani R, Graimann B, Moradi M H and Pfurtscheller G 2007 A comparison approach toward finding the best feature and classifier in cue-based BCI *Med. Biol. Eng. Comput.* **45** 403–12

Cincotti F, Mattia D, Aloise F, Bufalari S, Schalk G, Oriolo G, Cherubini A, Marciani M G and Babiloni F 2008 Non-invasive brain–computer interface system: towards its application as assistive technology *Brain Res. Bull.* **75** 796–803

Dal Seno B, Matteucci M and Mainardi L T 2010 The utility metric: a novel method to assess the overall performance of discrete brain–computer interfaces *IEEE Trans. Neural Syst. Rehabil. Eng.* **18** 20–8

Farwell L A and Donchin E 1988 Talking off the top of your head: toward a mental prothesis utilizing event-related brain potentials *Electroencephalogr. Clin. Neurophysiol.* **70** 510–23

Fatourechi M, Mason S G, Birch G E and Ward R K 2006 Is information transfer rate a suitable performance measure for self-paced brain interface systems in *Proc. IEEE Int. Symp. on Signal Processing and Information Technology* pp 212–6

Fatourechi M, Ward R K and Birch G E 2008 A self-paced brain–computer interface system with a low false positive rate *J. Neural Eng.* **5** 9–23

Ferrez P W and Millán J d R 2005 You are wrong!—automatic detection of interaction errors from brain waves *Proc. 19th Int. Joint Conf. on Artificial Intelligence (July 30–August 5 2005, Edinburgh, Scotland)* pp 1413–8

Gao X, Xu D, Cheng M and Gao S 2003 A BCI-based environmental controller for the motion-disabled *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 137–40

Huggins J E, Wren P A and Gruis K L 2011 What would brain–computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis *Amyotrophic Lateral Scler.* **12** 318–24

Jin J, Allison B Z, Sellers E W, Brunner C, Horki P, Wang X and Neuper C 2011 An adaptive P300-based control system *J. Neural Eng.* **8** 036006

Krauledat M, Dornhege G, Blankertz B, Curio G and Müller K-R 2004 The Berlin brain–computer interface for rapid response *Biomed. Tech.* **49** 61–62

Kronegg J, Alecu T and Pun T 2003 Information theoretic bit-rate optimization for average trial protocol brain-computer interfaces *Proc. HCI Int. 2003 (Crete, Greece)* pp 1437–40

Kronegg J, Voloshynovskiy S and Pun T 2005 Analysis of bit rate definitions for brain-computer interfaces *Proc. Int. Conf. on Humancomputer Interaction (HCI '05) (June 2005, Las Vegas, Nevada, USA)* pp 40–6

Mason S G and Birch G E 2000 A brain-controlled switch for asynchronous control applications *IEEE Trans. Biomed. Eng.* **47** 1297–307

Mason S G and Birch G E 2003 A general framework for brain-computer interface design *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 70–85

McFarland D J and Krusienski D J 2012 BCI signal processing: feature translation *BCI Principles and Practice* (Oxford, New York: Oxford University Press) chapter 8

McFarland D J, Sarnacki W A and Wolpaw J R 2003 Brain–computer interface (BCI) operation: optimizing information transfer rates *Biol. Psychol.* **63** 237–51

Meinicke P, Kaper M, Hoppe F, Heumann M and Ritter H 2003 Improving transfer rates in brain computer interfacing: a case study *Advances in Neural Information Processing Systems* ed S Becker, S Thrun and K Obermayer (Cambridge, MA: MIT Press) pp 1107–14

Millán J R and Mouriño J 2003 Asynchronous BCI and local neural classifiers: an overview of the adaptive brain interface project *IEEE Trans. Neural Syst. Rehabil. Eng.* **11** 159–61

Müller-Putz G, Scherer R, Brunner C, Leeb R and Pfurtscheller G 2008 Better than random: a closer look on BCI results *Int. J. Bioelectromagnetism* **10** 52–55

Nykopp T 2001 *Statistical modelling issues for the adaptive brain interface MSc Thesis* Department of Electrical and Communication Engineering, Helsinki University of Technology

Ortner R, Allison B Z, Korisek G, Gaggl G and Pfurtscheller G 2011 An SSVEP BCI to control a hand orthosis for persons with tetraplegia *IEEE Trans. Neural Syst. Rehabil. Eng.* **19** 1–5

Pfurtscheller G and Neuper C 2001 Motor imagery and direct brain–computer communication *Proc. IEEE* **89** 1123–34

Pierce J R 1980 *An introduction to information theory* (New York: Dover)

Quitadamo L R, Abbafati M, Cardarilli G C, Mattia D, Cincotti F, Babiloni F, Marciani M G and Bianchi L 2012 Evaluation of the performances of different P300 based brain-computer interfaces by means of the efficiency metric *J. Neurosci. Methods* **203** 361–8

Roberts S and Penny W 2000 Real-time brain computer interfacing: a preliminary study using Bayesian learning *Med. Biol. Eng. Comput.* **38** 56–61

Scherer R, Lee F Y, Schlögl A, Leeb R, Bischof H and Pfurtscheller G 2008 Towards self-paced brain–computer communication: navigation through virtual worlds *IEEE Trans. Biomed. Eng.* **55** 675–82

Scherer R, Schloegl A, Lee F, Bischof H, Janša J and Pfurtscheller G 2007 The self-paced Graz brain–computer interface: methods and applications *Comput. Intell. Neurosci.* **2007** 79826

Schlögl A, Kronegg J, Huggins J E and Mason S G 2007 Evaluation criteria for BCI Research *Toward Brain–Computer Interfacing* eds G Domhege, J d R Millán, T Hinterberger, D McFarland and K-R Müller (Cambridge, MA: MIT Press) pp 297–312

Sellers E W and Donchin E 2006 A P300-based brain-computer interface: initial tests by ALS patients *Clin. Neurophysiol.* **117** 538–48

Shannon C E and Weaver W 1964 *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press)

Townsend G, Graimann B and Pfurtscheller G 2004 Continuous EEG classification during motor imagery—simulation of an asynchronous BCI *IEEE Trans. Neural Syst. Rehabil. Eng.* **12** 258–65

Townsend G, LaPallo B, Boulay C, Krusienski D, Frye G, Hauser C, Schwartz N, Vaughan T, Wolpaw J and Sellers E 2010 A novel P300-based brain–computer interface stimulus presentation paradigm: moving beyond rows and columns *Clin. Neurophysiol.* **121** 1109–20

Tsui C S L, Gan J Q and Roberts S J 2009 A self-paced brain–computer interface for controlling a robot simulator: an online event labelling paradigm and an extended Kalman filter based algorithm for online training *Med. Biol. Eng. Comput.* **47** 257–65

Volosyak I 2011 SSVEP-based Bremen–BCI interface—boosting information transfer rates *J. Neural Eng.* **8** 036020

Wills S A and MacKay D J C 2006 Dasher—an efficient writing system for brain-computer interfaces? *IEEE Trans. Neural Syst. Rehabil. Eng.* **14** 244–6

Wolpaw J R, Birbaumer N, McFarland D J, Pfurtscheller G and Vaughan T M 2002 Brain–computer interfaces for communication and control *Clin. Neurophysiol.* **113** 767–91

Wolpaw J R, Ramoser H, McFarland D J and Pfurtscheller G 1998 EEG-Based communication: improved accuracy by response verification *IEEE Trans. Rehabil. Eng.* **6** 326–33