EEG-Based Synchronized Brain-Computer Interfaces: A Model for Optimizing the Number of Mental Tasks

Julien Kronegg, Guillaume Chanel, Sviatoslav Voloshynovskiy, *Member, IEEE*, and Thierry Pun, *Member, IEEE*

***Abstract*—The information-transfer rate (ITR) is commonly used to assess the performance of brain–computer interfaces (BCIs). Various studies have shown that the optimal number of mental tasks to be used is fairly low, around 3 or 4. We propose an experimental validation as well as a formal approach to demonstrate and confirm that this optimum is user and BCI design dependent. Even if increasing the number of mental tasks to the optimum indeed leads to an increase of the ITR, the gain remains small. This might not justify the added complexity in terms of protocol design.**

***Index Terms*—Brain–computer interface (BCI), classification, electroencephalogram (EEG), information transfer rate, optimal number of mental tasks.**

I. Introduction

BRAIN–COMPUTER interfaces (BCIs) are input-devices that allow a user to communicate with a computer by way of thinking. Thoughts are inferred from the neuronal electrical activity; brain structure is however far too complex to allow users to think to whatever they want and still having a system able to infer what was thought of. The "thoughts" vocabulary must, therefore, be limited to a few mental tasks with well characterized and localized neuronal activity, such as imagination of finger movement or of a rotating object. If properly recognized from the analysis of the electroencephalogram (EEG) signals, each mental task can then be used as a command. Sample applications include wheelchair or virtual keyboard control in the context of rehabilitation for disabled people, and more recently entertainment.

Synchronized BCI systems are defined by Mason *et al.* [3] as BCIs that recognize unintentional control, and that are intermittently available for control. As opposed to self-paced BCIs, synchronized BCIs require the system to prompt the user for a response and, therefore, ignore unexpected user input. The performance of such systems is usually measured using the information-transfer rate (also called mutual information or bit-rate), as proposed by Wolpaw *et al.* [6]

$$B = V\left[\log_2 N + P\log_2 P + (1-P)\log_2\left(\frac{1-P}{N-1}\right)\right] \quad (1)$$

where the information-transfer rate B corresponds to the amount of information reliably received by the system, V is the application speed in trials/second (i.e., how many thoughts are recognized per second), P is the classifier accuracy (i.e., how well thoughts are recognized) and N is the number of mental tasks (or symbols) used in the "thoughts" vocabulary. This definition is well suited to compare keyboard-based BCI applications, but most likely not for pointing device BCIs where a Fitts' law [7] could be more appropriate. Moreover, it is not really suited to assess either the so-called "noncomparable" BCI devices [8] or the self-paced BCIs [9], and must be used with caution for evaluating experiments where the number of tasks N is greater than four [10]. Nevertheless, it has the advantage of being very simple; using this measure, commonly reported bit-rates range from 5 to 25 bits/min [11].

Experimental work has already been done aiming at optimizing the number of mental tasks used in BCIs. McFarland *et al.* [2] found that the optimal number of mental tasks is around 4, while Obermaier *et al.* [4] and Dornhege *et al.* [1] showed it to be 3 and later 4 [5]. These four studies confirmed that the optimal number of tasks depends on user skills. McFarland *et al.* also suggest that the optimal number of tasks could be improved for well trained users. Obermaier *et al.* hypothesized that the classification accuracy constantly decreases when the number of tasks N increases. Dornhege *et al.* also showed that the information-transfer rate (ITR) improvement for more than three tasks is tiny if the pairwise classification error is about 10%. They also suggest that increasing the ITR by increasing the number of tasks to more than 3 or 4 tasks is unlikely.

In this paper, we propose a model to compute the optimal number of tasks for synchronized BCIs. This model confirms the dependence of the optimal number of tasks on the user skills, as well as shows a dependence of the BCI design on this optimum. Further, we show that the achievable ITR improvement when increasing the number of tasks is low, an effect that comes from insufficient classifier accuracy.

The paper is organized as follow. Section II presents an experimental protocol with 2–4 tasks, which should allow to confirm that the optimal number of tasks depends on the user skills. Two BCIs with up to state of the art accuracies are proposed and evaluated to assess the impact of the BCI design on the optimal number of classes. Section III addresses the theoretical modeling of synchronized BCIs. In Section III-A, we introduce a model for computing the optimal number of tasks as a function of the accuracy. With this model, the potential accuracy gain when increasing the number of tasks can be determined. Section III-B proposes a model that explains why the ITR increase is low when increasing the number of tasks. Results are presented and discussed in Section IV.

Manuscript received December 21, 2005; revised August 29, 2006; accepted November 5, 2006. This work was supported in part by the Swiss National Center of Competence in Research (IM)2, Interactive Multimodal Information Management (http://www.im2.ch/) and in part by the European Network of Excellence SIMILAR.

The authors are with the Computer Vision and Multimedia Laboratory, University of Geneva, CH-1211 Geneva 4, Switzerland (e-mail: thierry.pun@cui.unige.ch).

Digital Object Identifier 10.1109/TNSRE.2007.891389

1534-4320/$25.00 © 2007 IEEE

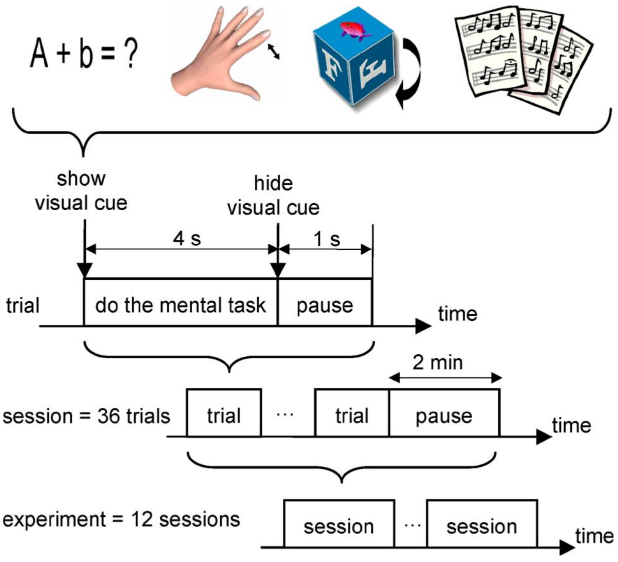


Fig. 1. Temporal structure of the protocol. Images are used to instruct the user. From left to right: mental calculation, finger movement, cube rotation, auditory evocation.

II. Experimental Validation

A. Participants

Four healthy right-handed male humans A, B, C, and D, 24–29 years old, participated in the study. They had no previous experience using BCI systems, except participant B (3 h). The participants were selected among the laboratory members for their availability and interest in the study. They were not remunerated. All participants filled a consent form before the experiment.

B. Data Acquisition

EEG data was recorded at 256 Hz using a 64 electrodes Biosemi Active Two [12] system with the ABC setup. The ABC setup is a derivation of the international 10/20 electrode placement system, and allows placing of up to 256 electrodes on the scalp. The mapping between the 64 electrodes and the 256-holders electrode cap is described in [13]. Each channel was filtered using a 4–45 Hz equiripple filter, which allows removing the power line noise (50 Hz) and the direct current and low-frequency components. Eye-blinks were not filtered out. Each electrode is referred to the local neighbourhood signal computed using a Surface Laplacian (SL) method [14]. Our results show that using the SL method improves the classification accuracy by 11% as opposed to using no reference. By comparison, common average reference only improves accuracy by about 4%. It has also been shown that Surface Laplacian reference give good results for focused brain activity [15], which is the case with the mental tasks that were selected.

C. Experimental Protocol

The participant was seated on a chair in front of an LCD computer screen, in a basement office offering reasonable immunity to electromagnetic noise. Following an offline synchronized protocol without feedback, he was instructed to execute the mental task corresponding to a trigger image displayed on the screen and depicting the required task. The term "synchronized" is used here in the sense of [3]: an intermittently available protocol and a transducer without unintentional control support. Four mental tasks were chosen (see Fig. 1): exact calculation of repetitive additions (T1), imagination of left finger movement (T2), mental rotation of a cube (T3), and evocation of a nonverbal audio signal (T4, e.g., a monophonic or polyphonic cell phone ring tone). According to physiological studies [16]–[19] and to other BCI designs [4], [20], these mental tasks generate activity in different brain regions, which ought to make classification easier.

Each mental task (trial) lasted for 4 s, followed by a 1-s pause. A session was composed of 36 trials of randomly selected tasks (each task thus appearing nine times), followed by a 2-min pause (totalling 5 min) (see Fig. 1). During pause periods, the user could think to whatever he wanted apart from thinking to the tasks themselves. Usually participants were relaxing, changing position, moving, etc., thus these periods were often characterized by significant muscle artifacts. During long pauses, participants were also filling in the questionnaire and discussing with the experimenter. The full experiment comprised 12 sessions (1 h), totalling 108 trials for each mental task. After each session, participants subjectively reported on the quality of their performance using questionnaires.

D. Features Extraction

We used two feature extraction methods: power of the short-term Fourier transform (STFT) with and without baseline subtraction. The baseline subtraction process is explained below. Both feature sets were selected because of the time-frequency spatial behavior of the brain when executing the selected mental tasks. The STFT is used instead of the discrete wavelet transform (DWT) because of its higher redundancy which should allow easier analysis (we, however, did not compare STFT with DWT). We also preferred STFT over the Continuous Wavelet Transform for its simpler interpretation. Similar features are also used by Coyle *et al.* [21] and produce accuracies of about 89% on 2-tasks BCIs. We made the hypothesis that the EEG signal is stationary for periods of about 230 ms, thus the STFT uses 60 samples windows with 50% overlap. The STFT power is computed by squaring the module of the complex STFT, over a subset of each 4-s trial only (see Fig. 2), which produces 20 timeframes. We used six 4.3-Hz frequency bands over the range 8.5–30 Hz. The number of features is consequently 7680 (64 electrodes × 20 timeframes × 6 frequency bands). Matlab's `specgram` function is used to compute the STFT.

When baseline subtraction is performed, the baseline is computed as the STFT mean power over the 1-s segment before the trial onset (see Fig. 2). The baseline is thus composed of six frequency bands over one timeframe. This baseline is subtracted to each of the 20 timeframes of the trial's features, for each electrode.

Using two feature sets (without and with baseline subtraction) allows for comparing BCIs with the same users set. This comparison supports our conclusions regarding the dependence of the optimal number of tasks on the BCI design.

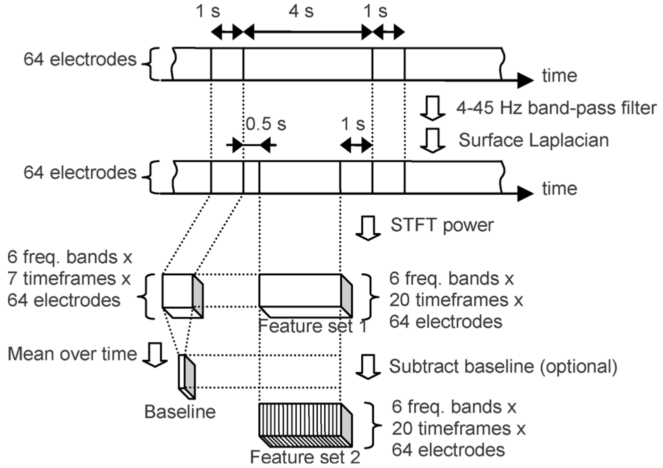


Fig. 2. Features extraction process. The first 0.5 s of the trial is not used because it contains an evoked potential due to the cue image perception. Using it could lead to classify the task of recognizing the cue image and not the associated mental task itself. The last second of the trial is not used in order to reduce the number of features. The baseline is computed over the 1-s segment occurring before the cue image display.

E. Classification

Three classifiers were experimented: a classification and regression tree (CART), a linear discriminant analysis (LDA) classifier, and a support vector machine (SVM) classifier. The CART Matlab classifier `treefit` is used with the Gini-index as a split criterion [22]. A five folds cross-validation was done with Matlab function `treetest` to obtain the best classification tree by pruning to the minimum tree size. The LDA Matlab classifier `classify` was used with pooled estimate of diagonal covariance matrices. The SVM classifier is a linear SVM from the OSU-SVM Matlab toolbox with default parameters [23]. All processing was done in Matlab. The LDA and SVM classifiers have proven their efficiency in the BCI 2005 classification competition [24], [25]. CART was selected for its known efficiency on power spectral density features [26].

We studied the existence of spatial, spectral, and temporal clustering for the four mental tasks. We did a Kruskal–Wallis statistical test with a 95% confidence level individually on each feature, to assess whether the samples originated from the same population. With baseline subtracted features 3100 out of 7680 have significantly different median values, which allows to differentiate between four mental tasks (this figure dropped to 1400 with features without baseline subtraction). This analysis however gives a partial view only because the LDA and SVM classifiers make decisions based on all the features. Thus a complete clustering cannot be determined without using feature selection techniques. We did not investigate further the issue of optimal features selection because of the exponential complexity that occurs when combining features as discussed in [27]. The feature selection process is an inherent problem of machine learning methods and still an open issue. Approaches aiming at finding the best features set in the BCI context do exist, such as in [28] and [29].

F. Validation

We used a sampled version of the leave-v-out stratified cross-validation (see Algorithm 1) with $\mathrm{M} = 1000$ sampling steps for the LDA and SVM classifiers and only $\mathrm{M} = 200$ sampling steps for the CART classifier because CART is already cross-validated and takes more CPU power. At each pass, 80% of the instance set D is randomly taken for training $(\mathrm{TRN_i})$ while the remaining 20% is left for testing $(\mathrm{TST_i})$. This method allows computing the mean and variance of the classifier accuracy. We preferred it over the leave-one-out cross-validation which allows computing the mean accuracy only, and over the k-fold cross-validation which allows computing only the mean accuracy and a coarse estimation of the variance. The drawback of the leave-v-out stratified cross-validation is its high computational cost. Our tests also show that leave-v-out cross-validation leads to 2%–2.5% lower classification accuracies than the leave-one-out cross-validation, and that the mean accuracy is very stable and does not depend on the training set size.

Algorithm 1. Sampled version of the leave-v-out stratified cross-validation. The model $\mathrm{h_i}$ is trained on $\mathrm{TRN_i}$ and tested on $\mathrm{TST_i}$.

```
for i = 1 to M
    divide D in TRNi and TSTi by random
    sampling
        with same class probability
    $h_i$ = learn on TRNi
    $accuracy_i = h_i$ (TSTi)
end for
$m$ = mean $(accuracy_i)$
$v$ = var $(accuracy_i)$
```

For $N < 4$, all tasks combinations (e.g., [T1, T2], [T1, T3, T4]) have been tested. For each task combination, the accuracy of each feature/classifier combination is compared using a 1-tailed t-test with 99.5% confidence level. As the t-test only compares two feature/classifiers (H0 hypothesis is "feature/classifier A performs as well as feature/classifier B"), the best overall feature/classifier for each task combination is determined by majority voting over the number of rejected H0 hypothesis. Then, another 1-tailed t-test with 99.5% confidence level was conducted to determine the best task combination in term of accuracy (see Table I). Our tests show a 11% difference between the best and the worst mental tasks combination, for $N = 2$ and $N = 3$.

III. Theoretical Modeling

A. Optimal Number of Tasks

We propose to describe the classification accuracy P as a function of the number of tasks N using a linear model

$$P(N) = aN + b, \text{ with } a < 0 \text{ and } 1/N \leq P(N) \leq 1 \quad (2)$$

where the slope a depends on the user skill as well as on the BCI design, and the intercept b is such that $0.5 \leq P(2) \leq 1$ since in the case of $N = 2$ classes one wants to have an accuracy

TABLE I

MEAN ACCURACIES AND STANDARD DEVIATIONS (SQUARE BRACKETED) FOR FOUR PARTICIPANTS USING FEATURES WITH AND WITHOUT BASELINE SUBTRACTION. FEATURES COMPUTED WITH BASELINE SUBTRACTION ALWAYS LEAD TO HIGHER ACCURACIES (SHOWN IN BOLD FONT). FOR ALL PARTICIPANTS, THE 2-TASKS COMBINATION [T1 T4] LEADS TO THE HIGHEST ACCURACY, WHILE THE 3-TASKS COMBINATION [T1 T2 T4] LEADS TO THE HIGHEST ACCURACY FOR ALL PARTICIPANTS, EXCEPT FOR PARTICIPANT B ([T1 T3 T4]). FOR PARTICIPANT A/FEATURES WITH BASELINE SUBTRACTION/$N = 4$, LDA AND SVM ACCURACIES ARE NOT SIGNIFICANTLY DIFFERENT, THUS BOTH ARE HIGHLIGHTED IN BOLD

	Number of tasks N	Features without baseline subtraction				Features with baseline subtraction			
		User A	User B	User C	User D	User A	User B	User C	User D
SVM	2	59.5 [6.9]	79.3 [5.5]	**68.3 [5.5]**	**77.2 [6.0]**	**89.7 [4.0]**	**91.7 [3.8]**	**96.3 [2.7]**	**91.7 [3.8]**
	3	39.0 [5.3]	65.6 [5.3]	**53.3 [5.5]**	**48.2 [5.7]**	78.1 [4.4]	**81.7 [4.3]**	**88.0 [3.5]**	**87.1 [3.8]**
	4	29.6 [4.5]	54.2 [4.9]	**42.4 [4.8]**	**35.0 [4.6]**	**67.4 [4.6]**	**71.2 [4.4]**	**73.9 [4.1]**	**77.1 [4.0]**
LDA	2	**66.0 [7.4]**	**85.4 [5.8]**	62.2 [7.2]	57.3 [9.3]	88.4 [5.0]	87.2 [4.9]	90.0 [4.3]	87.4 [5.0]
	3	**47.7 [5.8]**	**72.1 [5.5]**	45.3 [5.7]	36.0 [6.3]	**80.1 [4.8]**	74.4 [5.3]	79.0 [4.8]	79.5 [4.8]
	4	**37.4 [5.0]**	**58.0 [5.3]**	35.1 [5.0]	25.8 [4.7]	**67.2 [4.7]**	62.7 [4.9]	62.5 [4.6]	69.1 [4.6]
CART	2	55.6 [7.1]	61.1 [7.1]	49.3 [6.6]	57.9 [7.1]	75.7 [6.3]	74.5 [6.7]	80.0 [6.3]	82.5 [6.0]
	3	37.3 [6.3]	44.8 [5.4]	32.7 [5.8]	41.2 [5.7]	56.5 [6.4]	50.6 [6.0]	62.2 [5.9]	69.2 [5.7]
	4	29.2 [4.3]	33.7 [4.7]	26.0 [4.4]	29.3 [4.6]	43.3 [5.4]	41.3 [5.6]	46.3 [5.4]	55.9 [5.3]

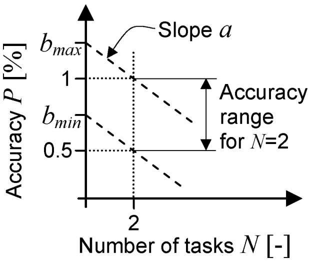


Fig. 3. Computation of the minimum and maximum values of intercept b.

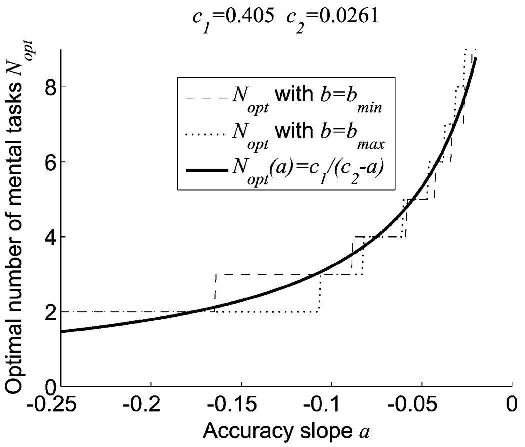


Fig. 4. Optimal number of tasks ((4)) as a function of the accuracy slope a and minimum/maximum values of the intercept b. The continuous line is the N_{opt} model described by (5).

higher than or equal to the random guess; thus, b varies between $b_{\min} = 0.5 - 2 \cdot a$ and $b_{\max} = 1 - 2 \cdot a$ (see Fig. 3).

The choice of a linear model seems the most natural after examination of various reported studies whose findings are graphically depicted in Fig. 7. As discussed in Section IV, the linear model is preferred over either a second-order polynomial or logarithmic models based on the correlation coefficient. This linear model is also justified because the number of possible values of N is small, typically limited to 7 ± 2 for protocols implying the memorization of mental tasks [30]. Consequently, using more complex models does not seem necessary.

For a given BCI design, users with differing skills will not perform in the same way. The user skills are measured here using the classifier accuracy P. The parameters a and b and thus $P(N)$ will therefore differ from one user to another. Similarly, for a given user, all BCI designs will not perform identically; different BCIs will exhibit differing a, b and thus $P(N)$. It is however difficult to differentiate the impact of the user skills from the impact of the BCI design: users are not usually shared by BCI researchers and BCI designs are tuned for a given set of test users. In this work, we will use the average participant performance on our own users to determine to which extent the BCI design influences the parameters a and b.

While (1)'s underestimation of Shannon ITR increases with the number of tasks used [10], this underestimation remains small, e.g., about only 0.2 bits for $N = 10$. Using the accuracy model from (2), we can thus compute the ITR from (1)

$$B(N,a,b,V) = V\left[\log_2 N + (aN+b)\log_2(aN+b) + (1-aN-b)\log_2\left(\frac{1-aN-b}{N-1}\right)\right]. \quad (3)$$

The corresponding optimal number of tasks for a given constant protocol speed V is

$$N_{\text{opt}} = \arg\max_N B(N,a,b,V = \text{constant}). \quad (4)$$

Fig. 4 presents (4) as a function of slope values a within the useful range, for both intercept values $b_{\min}$ and $b_{\max}$. The optimal number of tasks N_{opt} mostly depends on a and to a lesser extent on b: the optimal number of tasks N_{opt} differs from at most one mental task for minimum and maximum values b.

The model described by (2) allows determining the feasibility of a specific BCI design. For example, if one wants to design a BCI with $N_{\text{opt}} = 8$ mental tasks, the linear model of Fig. 4 leads to a value of $a = -0.04$. Considering the best case $b =$

$b_{\max} = 1.08$, the classifier accuracy must be $P(8) = -0.04 \cdot 8 + 1.08 = 0.76 = 76\%$. As such accuracies cannot be reached for eight classes using current BCI technologies, it does not seem worthwhile attempting to build an 8 tasks BCI.

As the optimal number of tasks N_{opt} from (4) only depends to a lesser extent on the intercept b, we propose to model it as a function of the accuracy slope a only

$$N_{\text{opt}}(a) = \frac{c_1}{c_2 - a}. \tag{5}$$

This function is a good approximation of (4), and yields the minimum mean-squares error compared with other models (linear, second-order and third-order polynomial, logarithmic, exponential). The parameter values c_1 and c_2 have been estimated using the mean-squares error method for $b = b_{\min}$, and for $b = b_{\max}$. Given the low N_{opt} dependency on b, we chose to use the mean between these two extremes, hence $c_1 = 0.405$ and $c_2 = 0.0261$. This allows to predict the optimal number of mental tasks as well as to explain why some users can have $N_{\text{opt}} > 2$. For example, a user with an accuracy slope of $a = -0.05$ should get an optimal number of tasks of 5, while a user with an accuracy slope of $a = -0.2$ will get an optimal number of tasks of 2.

B. ITR Improvement

We model the BCI as a channel with encoder and decoder, corrupted by additive white Gaussian noise (AWGN) [31] as in our previous work [10], [32]. A mental task W (e.g., "mental calculation") among N possible tasks is generated by the brain, encoded into a feature vector X by a feature extraction process, and then transmitted to a system that decodes the task $\hat{W}$. The feature vector X is contaminated by an independent AWGN $Z \sim \mathcal{N}(0, \sigma_z^2)$ induced by the background activity of the brain, yielding the noisy feature $Y = X + Z$ (see Fig. 6). We denote by $p(y|x_i)$ the probability that a feature y is correctly recognized when x_i is emitted; since Z is Gaussian, we have

$$p(y|x_i) = 1/(\sqrt{2\pi}\sigma_Z)e^{-(y-x_i)^2/2\sigma_Z^2}.$$

Whereas not all features are Gaussian distributed in practice, the Gaussian assumption is commonly accepted in the BCI community [1], [5], [10], [32]–[38].

We model the feature vector X as a pulse amplitude modulated (PAM) signal of N equiprobable states with constrained energy $\mathrm{E}[X^2] \leq \sigma_x$. The equiprobable assumption is commonly accepted by the BCI community (e.g., [1], [5], [6], [10], [32], [35]–[37]). The PAM assumption is reasonable, considering that several features used in the BCI community can be reduced to such PAM signal (e.g., mu/beta rhythm modulation [2], power spectral density [1], [4]). If a Bayes classifier is used, the probability that a mental task w_i is recognized as the mental task $\hat{w}_j$ becomes

$$p(\hat{w}_j|w_i) = \int_{R_j} p(y|x_i)dy$$

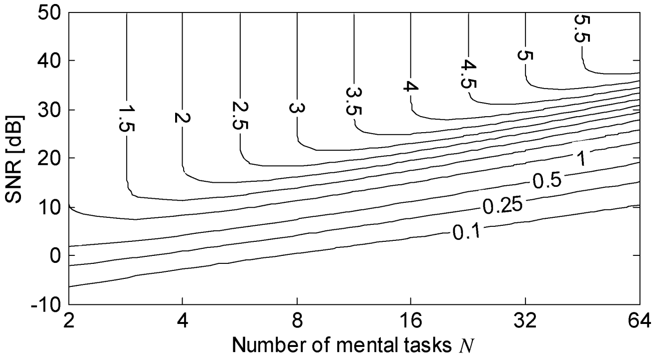


Fig. 5. Contour lines plot of constants Wolpaw's ITR in bits/trial computed from (7) with $V = 1$; the ITR here varies between 0.1 and 5.5 bits/trial.

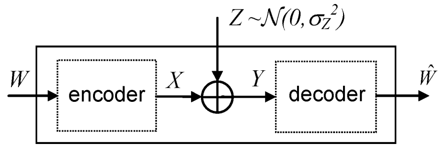


Fig. 6. BCI channel model. The user's brain executes the mental task W, producing a feature vector X. The noisy feature vector Y is produced by adding the background activity Z, and decoded as $\hat{W}$ by the classifier.

R_j being the Bayes decision region for $\hat{w}_j$ [39]. The terms $p(\hat{w}_j|w_i)$ constitute the transition matrix. The noise variance is computed from $\text{SNR} = 10 \cdot \log_{10}(\sigma_X^2/\sigma_Z^2)$, with the feature energy arbitrarily fixed at $\sigma_X^2 = 1$, so that $\sigma_Z^2 = 10^{-\text{SNR}/10}$. The system accuracy estimate $\hat{P}$ [(6)] is computed as the mean of the transition matrix diagonal

$$\hat{P} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi 10^{-SNR/10}}} \int_{R_i} e^{-\frac{(y-x_i)^2}{2\cdot 10^{-SNR/10}}} dy \tag{6}$$

Thus, from (1) and (6), we can compute Wolpaw's ITR as a function of the SNR and N (see Fig. 5)

$$\hat{B} = V\left[\log_2 N + \hat{P}\log_2 \hat{P} + (1-\hat{P})\log_2\left(\frac{1-\hat{P}}{N-1}\right)\right]. \tag{7}$$

The assumptions underlying Wolpaw's ITR (equiprobable classes, same accuracy for all classes, error distributed equally on the remaining classes) are not valid in all practical cases. However, our previous work [10] has shown that Wolpaw's ITR is very close to Shannon's ITR when the number of tasks does not exceed about five. When the number of tasks is greater than five, Wolpaw's definition underestimates Shannon's ITR, and this underestimates increases with N. The use of Wolpaw's ITR is thus justified in this paper, as the maximum number of tasks currently used in BCIs is 6: the underestimation is thus small.

Using our channel model, we can expect that an increase in the number of mental tasks N will lead to an increase of the ITR, if and only if the SNR is sufficiently high. For example, if the SNR is around 40 dB for every N, the ITR will greatly increase as N increases (see Fig. 5). Conversely, if the SNR is around 0 dB for every N, the ITR will even decrease. Consequently, only BCIs with good accuracy (high SNR) will significantly benefit from an increase of the number of mental tasks.

IV. Results and Discussion

Participants' brain signals are classified using the two feature sets and the three classifiers described in the previous section. Participants reported a mislabelled trials rate lower than 1%. The mislabelled rate is the proportion of trials during which participant made the wrong mental task. Such low rate should not significantly affect the classification. Participants also reported that fatigue increases after 1.5-2 h of experiment; only the first hour of data was thus used (108 trials per task). The questionnaires seemed to indicate that well trained users are less subject to fatigue. However, fatigue was not our main concern in this paper, thus we did not investigate further its consequences. Table I shows the mean and standard deviation of the accuracy for the best task combination for participants A to D respectively, selected by 1-tailed t-test. Bold font is used to highlight the best feature/classifier selected by t-test for each participant and for $N = 2$ to 4. The t-test values are not shown in this paper.

Depending on the randomly chosen training and test sets as selected by Algorithm 1, the difference between the minimum and maximum accuracies can be up to 30% for $N = 2$. This appears as high standard deviations in Table I, and can be explained by the fact that the two classifiers used are known to have low bias but high variance. The best feature/classifier pair is user dependant: for three participants, the features with baseline subtraction and SVM classifier give the highest accuracy, while for one participant the features with baseline subtraction and LDA perform better. This confirms Wolpert's "no free lunch theorems" [40], stating that "for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class." The significant dependence of the classification results on the measurement session and on the user allows considering distinct measurements as separate classification problems. Accuracies for CART are on the average lower by 15% than SVM accuracies when $N = 2$ and by 10% when $N = 4$, either with or without baseline subtraction. This seems to indicate that, at least in our setting, a CART classifier is not well adapted to this type of features. We nevertheless included CART accuracies in Table I because a CART classifier could be the optimal classifier for another user outside our set of participants, according to Wolpert's theorem.

Following the taxonomy from Mason *et al.* [41], we selected several BCI systems ([1], [2], [4], [5]) that are comparable, using the ITR, in the sense of [8]: all these BCIs use endogenous transducers and EEG signal, and have discrete classes without idle support, except [2] which uses discretized continuous classes. The average accuracy from Table I is in the upper state-of-the-art accuracy range (see Fig. 7). Moreover, the accuracy's standard deviation is much lower than [5]. It should also be emphasized that our participants are much less trained than in other studies and that a simpler setup without feature selection is used. By comparison, participants in [2] and [4] were highly skilled, while the participants' previous experience was not specified in [1] and [5]. As our participant set is rather limited in size, our reported performance may not be representative of the whole population. As Güger *et al.* pointed out, not all people can use a BCI at the same accuracy level [42].

The information-transfer rates are computed using (1) in which the accuracies are those produced by the best classifier for each participant. They are reported in Table III.

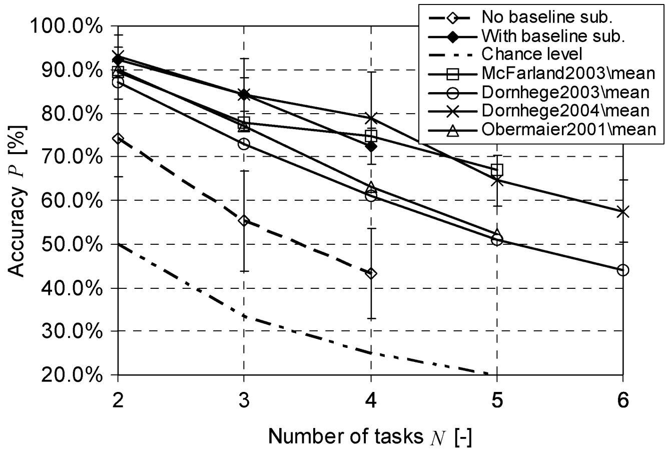


Fig. 7. Mean accuracies for the best task combination and the best feature/classifier for participants A–D and comparison with results from other studies. Error bars show the accuracy's standard deviation between users. For studies where only the ITR is given, the accuracy is computed from (1).

A. Optimal Number of Tasks

The results obtained in the present as well as in other studies (see Fig. 7 and Table I) confirm that N and P are interdependent: P decreases when N increases. The very existence of an optimal number of mental tasks can thus be inferred from the structure of (1): as N increases the first term $\log_2 N$ increases the ITR while the other two terms decrease the ITR. The experimental accuracy (Fig. 7) is significantly correlated with the one given by the linear, second-order polynomial and logarithmic models ($R^2 = 0.981$, $R^2 = 0.971$, and $R^2 = 0.982$, respectively). Linear and logarithmic models are significantly better than the second-order polynomial model (t-test, 10% significance). There is however no significant difference between the linear and logarithmic models. Given the low number of points on which these models are fitted, the simplest model should be preferred. Thus, the linear accuracy model (2) is the best suited to fit the experimental accuracy curve (Fig. 7). If the number of mental tasks was to increase in the future with the progress in analysis techniques, a logarithmic or second-order polynomial model could be more appropriate for describing accuracy; the linear model would represent the worst-case scenario.

As an example, Fig. 8 compares the proposed ITR model from (3) with real data, for several accuracy slopes a computed by linear regression, leading to varying optimal numbers of mental tasks (from 2 to 4). This figure shows that the optimal number of mental tasks is not necessarily 3–4 as stated in [1], [2], [4], and [5], but varies according to the accuracy slope a.

The modeled $N_{\mathrm{opt}}(a)$ is strongly correlated with the measured N_{opt}, see Table II ($R^2 = 0.83$, or $R^2 = 0.82$ when the modeled N_{opt} is rounded towards the next integer value). The proposed model can thus be considered as valid. This also explains why the optimal number of tasks is user dependent. The average across participants over the accuracy slope a computed from our data for features with and without baseline subtraction is significantly different. This confirms that the optimal number of mental tasks is also BCI design dependent. Moreover, averaged accuracy slopes a from data reported in [1], [2], [4], and [5] are different from our own averaged values; to a lesser extent, this is another confirmation that the optimal number of tasks is

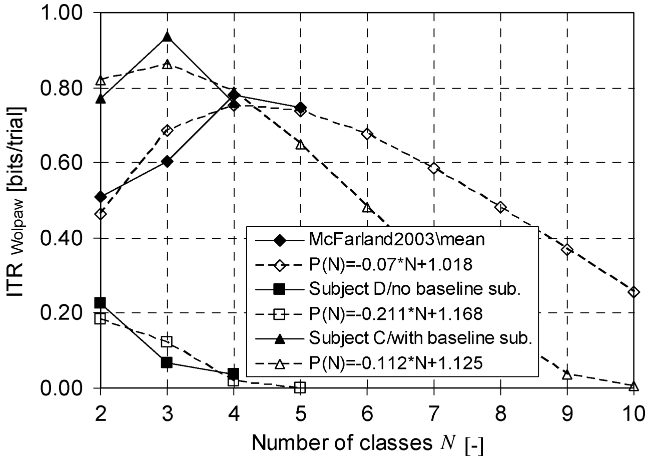


Fig. 8. Comparison between the proposed model from (3) (dashed lines) and ITR on real data computed using (1) (plain lines), for different accuracy slopes a showing optimal number of tasks of $N = 2, 3, 4$. As the number of fitted points is limited, the simplest model was chosen, see (2).

TABLE II
MEASURED AND MODELED OPTIMAL NUMBER OF SYMBOLS FOR PARTICPANTS A TO D AS WELL AS FOR OTHER PUBLISHED STUDIES. ERRORS ARE IN BOLD

Participant	Accuracy slope a	Measured N_{opt}	Modelled $N_{opt}(a)$
A (no baseline sub.)	-0.143	2	2.4
B (no baseline sub.)	-0.137	3	**2.5**
C (no baseline sub.)	-0.129	3	2.6
D (no baseline sub.)	-0.211	2	1.7
A (with baseline sub.)	-0.112	3	2.9
B (with baseline sub.)	-0.103	3	3.1
C (with baseline sub.)	-0.112	3	2.9
D (with baseline sub.)	-0.073	3	**4.1**
Mean of [1]	-0.108	3	3.0
Mean of [2]	-0.070	4	4.2
Mean of [4]	-0.127	3	2.6
Mean of [5]	-0.071	4	4.2

TABLE III
ITRS FOR USERS A TO D. THE HIGHEST ITR IS HIGHLIGHTED IN BOLD FONT. THE AVERAGE HIGHEST ITR AND SQUARE BRACHETED STANDARD DEVIATION ARE 0.22 [0.17] AND 0.80 [0.13] FOR FEATURES WITHOUT/WITH BASELINE SUBTRACTION, RESPECTIVELY. THE AVERAGE ITR IMPROVEMENT ARE 0.02 [0.02] AND 0.19 [0.09] FOR FEATURES WITHOUT/WITH BASELINE SUBTRACTION, RESPECTIVELY

Number of tasks N	ITR [bits/trial] (no baseline subtraction)				ITR [bits/trial] (with baseline subtraction)			
	A	B	C	D	A	B	C	D
2	**0.08**	0.40	0.10	**0.23**	0.52	0.59	0.77	0.59
3	0.06	**0.45**	**0.12**	0.07	**0.67**	**0.72**	**0.94**	**0.90**
4	0.05	0.35	0.10	0.04	0.57	0.68	0.76	0.86

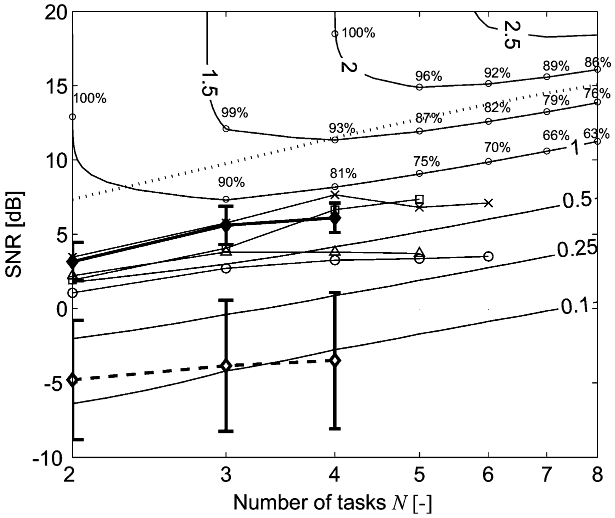


Fig. 9. ITR curves as a function of the number of tasks N and of the signal-tonoise ratio. The average ITR over participants A-D with and without baseline subtraction are depicted in plain black and dashed black respectively, error bars showing the standard deviation. ITR curves from other studies are displayed in plain gray (same symbols as in Fig. 7). The plain gray contour lines with indication of the required percentages of recognition rate show an ITR of 0.1 to 2.5 bits/trial. The dotted line shows a hypothetical BCI with accuracy modelled by (2) with $a = -0.03$ and $b = 1.05$; in this case, the ITR improvement would be significant when increasing the number of mental tasks. For a given BCI, the SNR is computed from the accuracy by gradient descent search on (6).

design dependent. However, as previously stated, this is only indicative since participants are not shared between researchers.

If future efforts in terms of BCI design were to permit significant increases of the optimal number of mental tasks, Wolpaw's ITR definition would lead to a significant underestimation of the real ITR. The optimal number of tasks as given by (5) would then loose accuracy.

B. ITR Improvement

Not all participants exhibited an increase in ITR when the number of mental tasks increased. In our case, the optimal number of mental tasks is $N = 3$ when baseline subtraction is performed (0.67 to 0.94 bits/trial, see Table III), or between 2 and 3 when baseline subtraction is not used (0.08 to 0.45 bits/trial). The amount of ITR improvement, if any, seems to be user dependent, as shown by large variations on ITR for features without baseline subtraction, and to a lesser extent, for features with baseline subtraction, with a standard deviation over mean ratio of $0.02/0.02 = 100\%$ and $0.09/0.19 = 47\%$, respectively. The reduced dependency of the ITR on the user's skills when baseline subtraction is performed is logical as one of the goals of this subtraction is indeed to decrease this dependency. However, the amount of ITR improvement is also BCI design dependent, as shown by the large variation between the average ITR with and without baseline subtraction, see Table III. Refining the BCI design should lead to an increase in the mean ITR over users and to a decrease of the ITR standard deviation among users (0.22 [0.17] to 0.80 [0.13] bits/trial, see Table III), but this could not be proved as the participant set is somewhat limited.

The average gain when increasing the number of tasks is relatively small (to the optimal N_{opt}: 0.19 bits/trial in our case, 0.02 bits/trial for [1], 0.04 bits/trial for [4], and 0.25 bits/trial for [2], and 0.31 bits/trial for [5]), because the ITR curves tend to "follow" the constant ITR curves from Fig. 9. Therefore, BCIs with "low" classification accuracy will not exhibit significant ITR improvements, if any, when increasing the number of mental tasks to any number of mental tasks. This extends Dornhege *et al.* conclusion [1] which predicts that increasing the number of mental tasks to more than 3 will lead to small

ITR improvements for typical BCI accuracies (i.e., more than 10% of errors for 2-class problems). This applies to all current state-of-the-art EEG-based BCIs (see Fig. 9). For BCIs relying on other types of electrodes, e.g., elecrocortico-graphic or intracortical electrodes, it is likely that the signal-to-noise ratio (SNR) will be higher than with scalp EEG signals, leading to a higher optimal number of mental tasks and to a higher ITR improvement.

V. Conclusion

Commonly reported results state that the optimal number $N_{\rm opt}$ of mental tasks for BCI applications is 3 to 4. It is further experimentally demonstrated by some researchers that this optimal number could be user-dependant. In this paper, we proposed and validated a model which allows to compute the optimal number of mental tasks and which explains the dependency of this number on both user skills and BCI design. We also showed that when increasing the number of mental tasks from $N = 2$ to $N = N_{\rm opt}$, the ITR and the ITR improvement both depend on the user skills and on the BCI design. Moreover, we observed that a good BCI design could improve the mean ITR and reduce the ITR standard deviation among users.

Our model also predicted that increasing the number of mental tasks from $N = 2$ to $N = N_{\rm opt}$ only produces a very limited improvement in terms of information-transfer rate. This is due to the fact that current BCI accuracies are too low to produce significant ITR increases. This observation is in accordance with Dornhege *et al.* theoretical model [1].

Even if increasing the number of tasks would lead to an increase in ITR, the small gain might not justify the added complexity in terms of protocol design. We can thus conclude that it is currently premature to aim at significantly increasing the information-transfer rate by increasing the number of tasks; improving classification accuracy should be the first target. This might prove to be difficult, given the accuracies that must be attained to obtain a given ITR (see Fig. 9). For instance, at $N = 4$, the accuracy must be 81%, 93%, and almost 100% to obtain an ITR of 1, 1.5, and 2 bits/trial, respectively. We feel that current promising methods to improve the classification accuracy include optimal features selection approaches, the development of new machine learning techniques, as shown by the third BCI competition results [24], and last but not least better mental tasks selection process, as shown in our task combination tests and in [43].

Acknowledgment

The authors would like to thank Dr. T. Iulian Alecu and Dr. O. Koval for fruitful discussions. The authors would also would like to acknowledge the anonymous reviewers for their detailed comments.

References

[1] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Increase information transfer rates in BCI by CSP extension to multi-class," presented at the Adv. Neural Inf. Proc. Syst. Conf. (NIPS 03) Vancouver, BC, Canada, Dec. 11–13, 2003.

[2] D. J. McFarland, W. A. Samacki, and J. R. Wolpaw, "Brain-computer interface operation: Optimizing information transfer rates," *Biol. Psychol.*, vol. 63, pp. 237–251, 2003.

[3] S. G. Mason and G. E. Birch, "Temporal control paradigms for direct brain interfaces—Rethinking the definition of asynchronous and synchronous," presented at the HCI Int. Conf., Las Vegas, NV, 2005.

[4] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, no. 3, pp. 283–288, Sep. 2001.

[5] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, Jun. 2004.

[6] J. R. Wolpaw, H. Ramoser, D. J. McFarland, and G. Pfurtscheller, "EEG-based communication: Improved accuracy by response verification," *IEEE Trans. Rehabil. Eng.*, vol. 6, no. 3, pp. 326–333, Sep. 1998.

[7] R. M. Baecker and W. A. S. Buxton*, Readings in Human-Computer Interaction: A Multidisciplinary Approach*. Los Altos, CA: Morgan Kaufmann, 1987.

[8] M. M. Moore-Jackson, S. G. Mason, and G. E. Birch, "Analyzing trends in brain interface technology: A method to compare studies," *Ann. Biomed. Eng.*, vol. 34, pp. 859–878, 2006.

[9] S. G. Mason, J. Kronegg, J. E. Huggins, A. Schlögl, M. Fatourechi, R. Kaidar, R. Scherer, and A. Buttfield, "Asynchronous BCI performance evaluation," *BCI-Info.Org. Res. Papers* May 2006 [Online]. Available: http://bci-info.tugraz.at/Research_Info/documents/articles/self_paced_tech_report-2006-05-19.pdf/view

[10] J. Kronegg, S. Voloshynovskiy, and T. Pun, "Analysis of bit-rate definitions for brain-computer interfaces," presented at the Int. Conf. Human-Computer Interaction (HCI'05) Las Vegas, NV, 2005.

[11] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, "Brain-computer interface technology: A review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, Jun. 2000.

[12] Biosemi [Online]. Available: http://www.biosemi.com. Amsterdam, The Netherlands, 2005

[13] J. Kronegg, T. I. Alecu, and G. Chanel, Electrode placement at the computer vision and multimedia lab Univ. Geneva, Geneva, Switzerland, 2005.

[14] T. I. Alecu, S. Voloshynovskiy, and T. Pun, "EEG cortical imaging: A vector field approach for Laplacian denoising and missing data estimation," presented at the IEEE Int. Symp. Biomed. Imag.: From Nano to Macro (ISBI'04), Arlington, VA, 2004.

[15] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, pp. 1842–1857, 1999.

[16] S. Dehaene, E. Spelke, P. Pinel, R. Stanescu, and T. S. , "Sources of mathematical thinking: Behavioral and brain-imaging evidence," *Science*, pp. 970–974, 1999.

[17] A. R. Halpern and R. J. Zatorre, "When that tune runs through your head: A PET investigation of auditory imagery for familiar melodies," *Cerebral Cortex*, vol. 9, pp. 697–704, 1999.

[18] S. M. Kosslyn, G. J. Digirolamo, W. L. Thompson, and N. M. Alpert, "Mental rotation of objects versus hands: Neural mechanisms revealed by positron emission tomography," *Psychophysiology*, vol. 35, pp. 151–161, 1998.

[19] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw, "Mu and beta rhythm topographies during motor imagery and actual movements," *Brain Topogr.*, vol. 12, pp. 177–186, 2000.

[20] E. Curran, P. Sykacek, M. Stokes, S. Roberts, W. Penny, I. Johnsrude, and A. M. Owen, "Cognitive tasks for driving a brain computer interfacing system: A pilot study," *IEEE Trans. Rehabil. Neural Rehabil. Eng.*, vol. 12, no. 1, pp. 48–54, Mar. 2001.

[21] D. Coyle, G. Prasad, and T. M. McGinnity, "A time-frequency approach to feature extraction for a brain-computer interface with a comparative analysis of performance measures," *EURASIP J. Appl. Signal Process.*, vol. 19, pp. 3141–3151, 2005.

[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone*, Classification and Regression Trees*. Boca Raton, FL: Chapman and Hall, 1993.

[23] J. Ma and Y. Zhao, OSU-SVM Matlab Toolbox v3.0 Ohio State Univ., Columbus, 2002.

[24] B. Blankertz, "The BCI competition III: Validating alternative approaches to actual BCI problems," *Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, Jun. 2006.

[25] J. Kronegg and D. A. Rofes-Gonzalez, BCI Competition III—Dataset V—TR05-02 Univ. Geneva, Switzerland, Comput. Sci. Dept. Tech. Rep., 2005.

[26] J. D. R. Millán, M. Franzé, J. Mourino, F. Cincotti, and F. Babiloni, "Relevant EEG features for the classification of spontaneous motor-related tasks," *Biol. Cybern.*, vol. 86, pp. 89–95, 2002.

[27] A. Date, "An information theoretic analysis of 256-channel EEG recordings: Mutual information and measurement selection problem," presented at the Int. Conf. Independent Component Anal. Blind Signal Separation (ICA2001), San Diego, CA, 2001.

[28] E. Gysels, P. Renevey, and P. Celka, P. Celka, Ed., "SVM-based recursive feature elimination to compare phase synchronization computed from broadband and narrowband EEG signals in brain-computer interfaces," *Signal Process. Special Issue: Neuronal Coordination Brain: A Signal Process. Perspective*, vol. 85, no. 11, pp. 2178–2189, 2005.

[29] . : D. A. Peterson, J. N. Knight, M. J. Kirby, C. W. Anderson, and M. H. Thaut, "Feature selection and blind source separation in an EEG-based brain-computer interface," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 19, pp. 3128–3140, 2005.

[30] G. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Rev.*, vol. 63, pp. 81–97, 1956.

[31] T. M. Cover and J. A. Thomas*, Elements of Information Theory*. New York: Wiley, 1991.

[32] J. Kronegg, T. I. Alecu, and T. Pun, "Information theoretic bit-rate optimization for average trial protocol brain-computer interfaces," presented at the HCI Int., Crete, Greece, 2003.

[33] P. Celka, B. Boashash, and P. Colditz, "Preprocessing and time-frequency analysis of newborn EEG seizures," *IEEE Eng. Med. Biol.*, vol. 20, no. 5, pp. 30–39, Sep./Oct. 2001.

[34] J. Kohlmorgen and B. Blankertz, "Bayesian classification of single-trial event-related potentials in EEG," presented at the Artifical Neural Networks (ICANN 2002), Madrid, Spain, Aug. 27–30, 2002.

[35] M. Sahin, "Noise tolerance as a measure of channel discrimination for multi-channel neural interfaces," presented at the Annual Int. Conf. IEEE Eng. in Med. and Biol. Soc., Istanbul, Turkey, Oct. 25–28, 2001.

[36] A. Schlögl, C. Keinrath, R. Scherer, and G. Pfurtscheller, "Information transfer of an EEG-based brain-computer interface," presented at the Int. IEEE EMBS Conf. Neural Engineering, Capri, Italy, 2003.

[37] A. Schlögl, C. Neuper, and G. Pfurtscheller, "Estimating the mutual information of an EEG-based brain-computer interface," *Biomedizinische Technik*, vol. 47, pp. 3–8, 2002.

[38] R. C. Smith, Electroencephalograph based brain computer interfaces Faculty Electrical Electronic Eng., Univ. College Dublin, Dublin, U.K., 2004.

[39] S. D. G. Duda, P. E. Hart, and D. G. Stork*, Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[40] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[41] S. G. Mason, A. Bashashati, M. Fatourechi, K. F. Navarro, and G. E. Birch, "A comprehensive survey of brain interface technology designs," *Ann. Biomed. Eng.*, vol. 35, no. 2, pp. 137–169, Feb. 2007.

[42] C. Güger, G. Edlinger, W. Harkam, I. Niedermayer, and G. Pfurtscheller, "How many people are able to operate an EEG-based brain-computer interface (BCI)?," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 145–147, Jun. 2003.

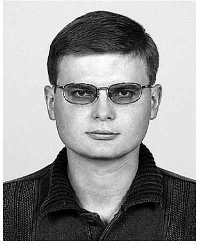
[43] K. Tavakolian, F. Vasefi, K. Naziripour, and S. Rezaei, "Mental task classification for brain computer interface applications," presented at the Canadian Student Conference on Biomedical Computing, Kingston, ON, Canada, 2006.

Julien Kronegg received the engineer degree in electronics and computer science from the Engineering School of Geneva, Geneva, Switzerland, in 1997, and the M.S. degree in computer science from the University of Geneva, Geneva, Switzerland in 2001 and the Ph.D. degree from the University of Geneva, following his work within the Computer Vision and Multimedia Laboratory (CVML).

His research interests include brain-computer interaction, image processing, and web-mining.

Guillaume Chanel received the engineer degree in computing and robotics and the M.S. degree in automatics from Montpellier University, France, in 2002. He is currently working toward the Ph.D. degree at the Computer Vision and Multimedia Laboratory (CVML) of the University of Geneva, Geneva, Switzerland.

His research interests concern the detection of emotional states based on the analysis of EEGs and other physiological signals in order to improve human–computer interaction.



Sviatoslav Voloshynovskiy (M'02) received the Ph.D. degree in electrical engineering from Lvivska Polytechnika University, Lviv, Ukraine in 1996. In 1998/1999 he was a visiting scholar at the University of Illinois at Urbana-Champaign.

He joined the Computer Science Department, University of Geneva, Switzerland in 1999 where he currently is associate Professor and Head of the Stochastic Image Processing group. His current research interests are in information-theoretic aspects of digital data hiding, visual communications with side information and stochastic image modeling for denoising, compression and restoration. He has coauthored over 100 journal and conference papers in these areas as well as nine patents.

Thierry Pun (M'93) received the Ph.D. in image processing from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 1982.

He worked several years at the National Institutes of Health, Bethesda, MD, then joined the University of Geneva, Geneva, Switzerland in 1986, where he currently is full Professor at the Computer Science Department and Head of the Computer Vision and Multimedia Laboratory. He has authored or coauthored seven patents and over 250 journal and conference papers. His current research interests, related to multimodal interaction and multimedia information systems, concern: physiological signals analysis for brain-computer Interaction and emotion assessment, multimodal interfaces for blind users, data hiding, information retrieval systems.