Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms

Guido Dornhege, Benjamin Blankertz, Gabriel Curio, Klaus-Robert Müller

***Abstract*— Non-invasive EEG recordings provide for easy and safe access to human neocortical processes which can be exploited for a Brain-Computer Interface (BCI). At present, however, the use of BCIs is severely limited by low bit-transfer rates. Here, we systematically analyze and furthermore develop two recent concepts, both capable of enhancing the information gain from multichannel scalp EEG recordings: (1) the *combination of classifiers* each specifically tailored for different physiological phenomena, e.g. slow cortical potential shifts, such as the pre-movement Bereitschaftspotential, or differences in spatio-spectral distributions of brain activity (i.e. focal event-related desynchronizations), and (2) behavioral paradigms inducing the subjects to generate one out of several brain states (*multi-class approach*) which all bare a distinctive spatio-temporal signature well discriminable in the standard scalp EEG. We derive information-theoretic predictions and demonstrate their relevance in experimental data. We will show in particular that a suitably arranged interaction between these concepts can *significantly* boost BCI performances.**

***Index Terms*— EEG, Event-Related Desynchronization, Movement Related Potential, Brain-Computer Interface, Common Spatial Patterns, Multi-class, Feature Combination, Single-Trial-Analysis**

I. Introduction

The ultimate goal of brain-computer interfacing (BCI) is to translate intentions of a subject into a control signal for a device, say a computer application, a wheelchair or a neuroprosthesis (e.g. [1]). Recent years have seen continuous progress in both invasive (e.g. [2], [3], [4]) and non-invasive BCI technology (e.g. [1], [5], [6], [7]). In this paper we will focus on non-invasive electroencephalogram (EEG) based brain computer interfacing which have the appeal of both: easy application and absence of procedural risks, such as infection or cortical micro-lesions, but still suffer from low bit transfer rates.

It is known that EEG signals under appropriate well-designed experimental paradigms allow a subject to convey her/his intentions by, for example, motor imagery or execution of specific mental tasks. Once the intentions have manifested themselves in brain activity and have been measured by EEG, the scene is set for advanced signal processing and machine learning technology. First, appropriate feature vectors need to be extracted from the digitized EEG-signals. To produce a control signal for a device, say, left vs. right, these feature vectors are then translated either (1) by threshold criteria or simple equations (with only a few parameters to be estimated on some training data) or (2) by more complex decision functions that are learned on the training data by machine learning techniques like linear discriminant analysis (LDA), support vector machines (SVMs) or artificial neural networks (ANNs).

It is very helpful for classification if the EEG-feature vectors are extracted such that they hold the most discriminative information for a chosen paradigm. It is here where neurophysiological a-priori knowledge can be very beneficial (e.g. [5], [8]). For some behavorial paradigms it is well-known that several distinct – possibly independent – physiological processes play an important role. Several authors, for example in [9], [10] point out the potential gain in using all such features, however investigations of feature combinings were announced, but so far not covered in publications. In our recent work [8], we showed some initial highly promising steps in this direction that have the potential to increase BCI performance enormously.

Our present paper will thus provide an extensive investigation on methods for combining features and confirm their value in an experimental BCI context. We contribute by providing theoretical insights on the expected performance gain when using a combination method. The presented results hold under the assumption of a certain base performance on the single feature vector and the level of independence. A special focus is placed on the question of how to incorporate a-priori knowledge about feature independence. Furthermore, we discuss feature combination in the context of multi-class BCI systems. In this context some multi-class extensions of the well-known *Common Spatial Pattern (CSP)* algorithm (cf. [11]) are proposed and evaluated. Finally we will show that both, i.e. feature combination and multi-class extensions, *together* hugely increase the performance of the BCI system.

The present study considers solely an offline analysis of our BCI experiments. Note, however, that an offline scenario is the most stable and preferable when testing for substantial classification improvements when comparing the new combination and multi-class methods with single feature analysis, cf. discussion in [12]. Since all features are present in all movement intentions or imagined movements, we also expect high performance gains in online BCI experiments.

GD and BB are with Fraunhofer FIRST (IDA), Kekuléstr. 7, 12 489 Berlin, Germany, E-mail: guido.dornhege@first.fraunhofer.de.

GC is with the Dept. of Neurology, Campus Benjamin Franklin, Charité University Medicine Berlin, Hindenburgdamm 30, 12 203 Berlin, Germany.

KRM is with Fraunhofer FIRST (IDA), see above, and also with University of Potsdam, August-Bebel-Str. 89, 14 482 Potsdam, Germany.

The studies were supported by a grant of the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01IBB02A and FKZ 01IBB02B.

The rest of the paper proceeds as follows: we will first provide a short overview of the necessary neurophysiological background (Section II), then clarify some important theoretical aspects of feature combination (Section III) and multi-class learning (Section IV). Subsequently we focus on new combination methods in Section V, in particular multi-class algorithms for CSP, i.e. describe the classification and signal processing techniques used in the experiments. A description of results, comparisons to possible theoretical gain under assumption of feature independence and a discussion follow in sections VI and VII respectively.

II. Neurophysiological background

Combination. Significant gain can be expected from a combination of several single features if these single features provide complementary information for the classification task. In case of sensorimotor cortical processes accompanying finger movements Babiloni et al. [13] demonstrated that movement related potentials (MRPs) and event-related desynchronizations (ERD), i.e. an amplitude decrease of the pericentral μ- and β-rhythms, indeed show up with different spatio-temporal activation patterns across primary (sensori-)motor cortex (M-1), supplementary motor area (SMA) and the posterior parietal cortex (PP). This finding is backed by invasive (subdural) EEG recordings [14] during brisk, self-paced finger and foot movements: MRPs started over widely distributed areas of the sensorimotor cortices (*Bereitschaftspotential*) and focused at the contralateral M-1 hand cortex with a steep negative slope prior to finger movement onset, culminating in a negative peak approximately 100 ms after EMG onset. In contrast, a bilateral M-1 ERD preceding the movement appeared to reflect a more widespread cortical 'alerting' function. Most importantly, the ERD response magnitude did not correlate with the amplitude of the negative MRPs slope. Note that these studies analyze preparation and actual execution of real movements. We presume a similar existence and independence of MRP (cf. [15]) and ERD phenomena for imagined movements, as is confirmed in Section VI. We also use the term 'Movement Related Potentials' for imagined of movements here.

Apart from exploiting complementary information on cortical processes, combining features based on MRP and ERD can provide the additional benefit of better robustness against artifacts originating from outside the central nervous system (CNS), such as eye movements (measured with EOG) or muscular artifacts (EMG). While EOG activity mainly affects slow potentials, i.e. MRPs, EMG activity is detrimental to oscillatory features, i.e. ERD, cf. [1]. Accordingly, a classification method based on both features could be construed to appropriately handle trials that are contaminated by either one of these artifacts. Yet the risk of using non-CNS activity for classification, which would not be conform with the basic BCI criteria [1], must nonetheless be addressed explicitly (cf. Section V-C).

Multi-class. It is an open question how many brain states could and should be used to implement a BCI. Using more than the two classes, as are involved in the usual binary decision tasks, requires that a suitable number of well discriminable brain-states can be identified. Obermaier et al. ([16]) report initial results showing the use of three classes yields improved BCI results. Before we analyze new algorithms to increase the information transfer rate (ITR) by extending BCI to multi-class paradigms, two psychological aspects shall be addressed: In principle, multi-class decisions should be derived from a decision space natural to human subjects. In a BCI context such decisions will be performed more 'intuitively', i.e. without a need for prolonged training, if the differential brain states are naturally related to a set of intended actions. This is the case, for example, for movements of different body parts which have a somatotopically ordered lay-out in the primary motor cortex resulting in spatially discriminable patterns of EEG signals, such as MRPs and ERDs specific for finger, elbow or shoulder movement commands. Notably, also non-motor spatially discriminable patterns of EEG signals can be induced by either auditory or visual imagery, for example when imagining a tune to move a cursor upwards vs. imaging a visual scene to induce a downward movement. However, this kind of contingency between 'brain action' and 'world effect' would be counter-intuitive. While humans are able to adapt and to learn such complex tasks, this could take *weeks* of training before it could be performed fast, quickly and 'automatically'. Another critical aspect of multi-class paradigms would arise if these classes could be identified only at the expense of lower accuracy which is likely to confuse the user in an online feedback setting.

III. Combining classifiers: theoretical aspects

Although the following calculations are easy and similar results (in other context) can be found in the literature, we demonstrate them here to highlight the theoretical reasons for using feature combination. We start with a set of N feature vectors described by random variables X_j with binary labels $Y \in \{\pm 1\}$ for $j \in \{1,...,N\}$. Let us further assume that for each feature an optimal classifier, i.e. with minimal misclassification risk, $f_j : \mathbb{R}^{d_j} \to \mathbb{R}$ can be found, where $P(f_j(X_j)|Y = \pm 1) = \mathcal{N}(\mu_{j,\pm 1}, \sigma^2_{j,\pm 1})$ and d_j is the dimension of the feature vector X_j. This is always the case if the feature vectors are Gaussian distributed with equal covariances under use of the optimal Bayes classifier LDA. In this case we additionally get $\sigma_j^2 = \sigma^2_{j,\pm 1}$ and furthermore for equal class priors[1] $\mu_j = \pm\mu_{j,\pm 1}$. The expected misclassification risk c_j of the feature vectors X_j is then one-to-one related to the quotient μ_j/σ_j. In more detail if we define $g(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp(-\frac{(x-z)^2}{2}) dx^2$, we get $c_j = g(\mu_j/\sigma_j)$. Although the following simple strategy is not necessarily optimal, it is sufficient to get a lower bound for the increase in performance by using more feature vectors. We define the combined classifier for a collection of feature vectors $x = (x_1, \ldots, x_N)$ by the sign of the function $\sum_{j=1}^{N} f_j(x_j)/\sigma_j$, i.e. we normalize each single classifier to have variance 1 on each class and sum them up. If the feature vectors are independent, we know that the sum of Gaussian distributions are again Gaussian distributed with mean (or variances) equal to the sum of the means (or variance) of all feature vectors. Consequently our constructed classifier is Gaussian distributed

[1] i.e. $P(Y = 1) = P(Y = -1) = 0.5$

[2] ':=' means 'defined as'

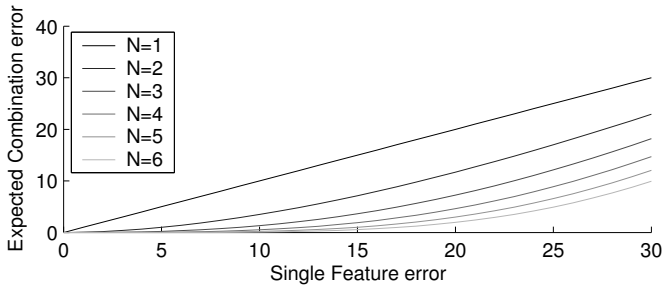

Fig. 1. In the figure the expected misclassification risk in % for combination of N different feature vectors are confronted to the single expected misclassification risk (assumption: all single feature vectors have the same performance and are mutually independent and Gaussian distributed).

with mean $\mu := \sum_{j=1}^{N} \mu_j/\sigma_j$ and $\sigma^2 := N$, if we assume independence. Then the expected misclassification risk c of this combined problem is

$$c = g(\mu/\sqrt{N}) = g\left(\frac{\sum_{j=1}^{N} g^{-1}(c_j)}{\sqrt{N}}\right). \qquad (1)$$

If all feature vectors have the same misclassification risk $\hat{c}$ we find the well-known rule for the signal-to-noise-ratio increase of the mean of N independent identical distributed Gaussian distribution of $\sqrt{N}$ (cf. [17]), namely that the combined problem has a misclassification risk $c = g(\sqrt{N} g^{-1}(\hat{c})) < \hat{c}$, since g is strictly monotonically decreasing. This result is shown in Fig. 1 for different numbers of feature vectors N and reveals a large gain of a combination strategy under this assumptions, e.g. combining 5 independent feature vectors with an error rate of 20 % each leads to an overall error of 3 %! Equation (1) in its general form, i.e. with different c_j's, is used in Section VI to compare the actual performance of the proposed feature combiners with performance (1) that could theoretically be obtained when features are perfectly independent.

IV. Multi-class paradigms: theoretical aspects

Our aim is to find a subset out of many possible brain states (classes) which is most profitable for the use as control paradigm in a BCI system. Here we investigate this issue in general from a pure information theoretic perspective. Using more classes has the potential to increase ITR, although the classification performance decreases. For subsequent theoretical considerations we assume Gaussian distributions with equal covariance matrices for all classes, which is a reasonable assumption for a wide range of EEG features. Furthermore we assume equal class priors, this means that all classes are expected to be used the same number of times, which is typical for many BCI applications. For three classes and pairwise equal classifications errors err, bounds for the expected classification error can be calculated in the following way: Let $(X,Y) \in \mathbb{R}^n \times \mathscr{Y}$ ($\mathscr{Y} = \{1,2,3\}$) be random variables with $P(Y=i) = 1/3$ (equal class priors) and $P(X|Y=i) = \mathscr{N}(\mu_i, \Sigma)$ for $i = 1,2,3$. Scaling appropriately, we can assume $\Sigma = I$. We define the *optimal* classifier by $f^* : \mathbb{R}^n \to \mathscr{Y}$ with $f^* = argmin_{f \in F} P(f(X) \neq Y)$, where F is some class of functions[3]. Similarly $f^*_{i,j}$ describes the optimal classifier between classes i and j. We directly get $err := P(f^*_{i,j}(X) \neq Y) = g(\|(\mu_i - \mu_j)/2\|_2)$ for $i \neq j$ with g as defined in the last section. Therefore we get $\|\mu_j - \mu_i\|_2 = \Phi(err)$ for all $i \neq j$ with some $\Phi(err) > 0$ and finally due to symmetry and equal class priors $P(f^*(X) \neq Y) = Q(\|X\|_2 \geq \min_{j=2,3}(\|X - \mu_j + \mu_1\|_2/2))$ where $Q = \mathscr{N}(0,I)$. Since evaluation of probabilities for polyhedrons in the Gaussian space is difficult, we only estimate lower and upper bounds. We can directly reduce the problem to a 2 dimensional space with $\mu_1 = 0$ by shifting, rotating and by Fubini's theorem. Since $\|\mu_j - \mu_i\|_2 = \Phi$ for all $i \neq j$ the means lie at the corners of an equilateral triangle. With $arg : \mathbb{R}^2 \to [-\pi, \pi)$, $arg(x) = \phi$, if $x = re^{i\phi}$ in the unique polar coordinates representation, we define the sets (see Fig. 2):

$$\begin{aligned}
A &:= \{x \in \mathbb{R}^2 \,|\, \mu_3^\top x > \Phi^2/2 \wedge arg(x) > \pi/3\} \\
B &:= \{x \in \mathbb{R}^2 \,|\, \mu_2^\top x > \Phi^2/2 \wedge arg(x) < 0\} \\
C_l &:= \{x \in \mathbb{R}^2 \,|\, \|x\|_2 > \Phi/\sqrt{3} \wedge arg(x) \in [0, \pi/3]\} \\
C_u &:= \{x \in \mathbb{R}^2 \,|\, \|x\|_2 > \Phi/2 \wedge arg(x) \in [0, \pi/3]\} \\
R &:= \{x \in \mathbb{R}^2 \,|\, \|x\|_2 \geq \|x - \mu_j\|_2,\ j = 2,3\}
\end{aligned}$$

We directly see that $A \cup B \cup C_l \subset R \subset A \cup B \cup C_u$. Due to symmetry, the equilateral triangle, polar coordinates transformation, some integral calculations and $P(R) = P(f^*(X) \neq Y)$ we finally get

$$\frac{\exp(-\Phi(err)^2/6)}{6} \leq P(f^*(X) \neq Y) - err \leq \frac{\exp(-\Phi(err)^2/8)}{6}. \qquad (2)$$

To compare classification performances involving different numbers of classes, we use the ITR quantified as bit rate per decision I as defined due to Shannon's theorem: $I := \log_2 N + p \log_2 p + (1-p)\log_2((1-p)/(N-1))$ with number of classes N and classification accuracy p (cf. [18]). Fig. 2 (right) shows these bounds ("3 range") for the ITR as a function of the expected pairwise misclassification errors. Note that less strict assumptions for the problem, like having more classes, make calculation of such bounds much harder. Here the results of such situations were obtained by simulation. We therefore visualize values on simulated data (100000 data points for each class) in the same figure, under the assumptions described above for $N = 2, ..., 6$ classes. While, the figure confirms our estimated bounds, it also shows that under these strict assumptions multi-class BCI yields significant gain in ITR. However, the biggest insight of this figure is that the gain of using more than 4 classes is small if the pairwise classification error is about 10 % or more. Under more realistic assumptions, i.e. more classes have increasing pairwise classification error compared to a sensibly chosen subset, it is improbable that increasing the number of classes to more than four will increase the bit rate. However, this depends strongly on the respective pairwise errors. If a suitable number of different brain states can be discriminated well, then extensions to more classes are indeed useful.

[3]For the moment we pay no attention to whether such a function exists. In the current setup F is usually the space of all linear classifiers, and under the probability assumptions mentioned above such a minimum exist.

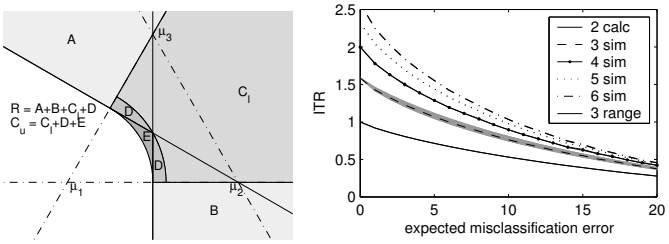

Fig. 2. The figure on the left visualizes a method to estimate bounds for the ITR depending on the expected pairwise misclassification risk ($P(R)$) for three classes. The figure on the right shows the ITR depending on the expected classification error [%] for simulated data for different number of classes (3-6 sim) and for 2 classes the real values (2 calc). Additionally the expected range (see equation (2)) (3 range) for three classes is visualized.

V. Data acquisition and Feature Extraction

A. Experiments

We recorded brain activity from 10 healthy subjects (codes *aa*, *ac af*, *ah*, *ak*, *ar*, *as*, *av*, *aw* and *ay*) in 15 different experiments with multi-channel EEG amplifiers using 64 or 128 channels band-pass filtered between 0.05 and 200 Hz and sampled at 1000 Hz. For offline analysis all signals were down-sampled to 100 Hz. Surface EMG at both forearms and one leg, as well as horizontal and vertical EOG signals, were recorded to check for muscle activation and eye movements, but no trial was rejected.

The subjects in this experiment sat in a comfortable chair with their arms relaxed on arm-rests. All 4.5 seconds one of 2, 3 or 6 different letters appeared on a computer screen for 3 seconds. During this period the subject had to do one of 6 different things according to the displayed letter: imagined movement of ***l***eft or ***r***ight hand or ***f***oot, or imagination of a ***v***isual (with eyes open), ***a***uditory or ***t***actile sensation. Only subjects *af*, *ak* and *as* used all 6 classes. One experiment was done with the 4 classes *l*,*r*, *f* and *v* with subject *aw*. Six experiments were conducted with the 3 classes *l*, *r* and *f* and subjects *aa*, *af*, *ar*, *av*, *aw* and *ay* and two with the 3 classes *f*, *a* and *v* with subjects *ac* and *ah*. Finally, subjects *aa*, *ac* and *ah* also took part in an experiment with the two classes *l* and *r*. For each class 160–200 trials were recorded.

The aim of classification in these experiments is to discriminate trials of different classes using the whole period of imagination. A further reasonable objective, i.e. to detect a new brain state as early as possible, was not an object of this particular study. Note that the classes *v* and *a* were included only to study single-trial EEG classification while those mental tasks are not intended for the use in our BCI system. The tasks were chosen because their cortical activation patterns can be well differentiated at a macroscopic scale of several centimeters so that both slow cortical potentials and oscillatory effects should be expected to be discriminable in principle.

B. Classification and Validation

Although a wide range of classifiers are available, we typically use Linear Discriminant Analysis (LDA) in the context of the BCI feature vectors to be presented in Section V-C. The reason for this is the concept to using 'simple methods first' and the fact that in our BCI studies linear classification methods were rarely found to perform worse than non-linear classifiers (cf. also [19], [20]). Furthermore the assumption of Gaussian distributions with equal covariance matrices holds well for the SUB feature vectors described later (cf. [21]). It was an interesting outcome of the BCI Competition 2003 ([12]) that on all 5 different kinds of BCI data sets linear methods either achieved the minimum test error among the competing algorithms or were at least not significantly worse than the best non-linear method, cf. [22]. In typical BCI scenarios high dimensional feature vectors, but only a small number of training samples are available. In these 'weak' feature vectors discriminative information is spread across many dimensions. A problematic effect of these high-dimensional small sample training sets is the well-known curse of dimensionality and overfitting problems. One possible alternative to avoid this is to perform a *strong preprocessing* in order to extract low dimensional feature vectors which are more tractable for most classifiers. In most situations such a 'strong' preprocessing is difficult to find since rather strong assumptions about the data distributions have to be made, which can be problematic in a BCI context. Therefore a different strategy which is well-established in machine learning, called *regularization*, is used where the idea is to appropriately limit the complexity of the classifiers function class. Typically a so-called regularization parameter has to be adapted to the data, that trades off the incurred training errors versus the stiffness of the function (see e.g. [20], [23]). For LDA regularization is done by modifying the covariance matrix

$$\Sigma \mapsto \lambda\Sigma + (1-\lambda)I, \tag{3}$$

i.e. by shrinking high eigenvalues and attenuating low eigenvalues of Σ ([24]).

To assess classification performance, the generalization error was estimated by a 10×10-fold cross-validation. Strictly speaking, the search for good regularization coefficients has to be performed on the training set in cross-validation. So in this offline analysis one would have to do a cross-validation (for model selection, MS) on each of the 100 randomly chosen training sets within a cross-validation procedure (for estimating the generalization error, GE), which is very time consuming. Alternatively, doing model selection by cross-validation on all trials could lead to overfitting and underestimation of the generalization error. As a practical intermediate way MS-cross-validation was performed beforehand on a 3×10-fold cross-validation on randomly chosen subsets of trials, which have the same size as the training sets in the GE-cross-validation, i.e. here 90 % of the whole set. This procedure was tested in several settings without any significant bias on the estimation of the GE, cf. [25].

For the purpose of this paper we analyze BCI experiments pursuing three directions: First, out of all binary subsets of classes in the presented experiments we compare the best performance of the single modality feature vectors (as presented in Section V-C) to the combination results when using all three feature vectors on these binary subsets. Second, we compute under some simplifications (PROBdiff is omitted due to its complexity and only a fixed preprocessing was

chosen) a feature combination on all suitable (see below) subsets of classes ($m = 3, \ldots, 6$) with multi-class extensions of the CSP algorithm presented in Section V-E. Finally, to be able to conclude that the combination of feature combination and multi-class extensions yield an improvement in the BCI context, these results are used to find the best setup of classes measured by the bit rate per decision I (cf. Section IV). For the choice of the best multi-class setup we follow [16] [4] here.

Note that we omit setups where the subjects did not generate discriminable brain signals. The criterion for rejection was that all three kinds of single feature vectors presented in Section V-C resulted in a classification error of more than 20 %. As result, 49 binary subsets and 95 multi-class settings remained, allowing to draw meaningful conclusions concerning feature combination and multi-class paradigms. More specifically subjects *ac* and *ah* were completely omitted; for the multi-class combination only two 6 class experiments could be used completely, the four class experiment with subject *aw*, and three class experiments with subjects *aa*, *ac*, *ar*, *av*, *aw* and *ay*.

C. Feature Extraction

The present behavioral paradigms allow to study the two prominent brain signals accompanying motor and sensory imagery: (1) the MRP, focussed over the corresponding motor cortex contralateral to the involved hand, or slow negative EEG-shifts over sensory cortices, and (2) the ERD appearing as a regional attenuation of the μ- and/or β-rhythms. Fig. 3 shows these effects calculated from subject *aa* on the classes *l* and *r*.

In the following we describe methods to compute feature vectors that can capture slow EEG shifts (such as MRP) or ERD effects. Note that all filtering techniques that will be used are causal. Thus all presented methods are applicable in online systems.

For the binary classifications some free parameters were chosen from appropriately fixed parameter sets by cross-validation for all experiments; each classification setting is separately described in Section V-B. This selection was done to obtain the most appropriate setting for each single-feature analysis. These values were used for both classifying trials based on single-feature vectors and the combined classification. In the multi-class settings a fixed setup of parameters was chosen which works well for all subjects and subsets of classes. Note that we expect a further increase of the multi-class performance if we carefully and extensively choose parameters for every individual and depending on the number of classes.

Slow non-oscillatory EEG potential shifts.

To quantify the slow potential shifts, such as the lateralized MRP, we proceeded in a similar fashion to our approach in [5] (Berlin Brain-Computer Interface, BBCI). Small modifications were made to take account of the different experimental setups. Signals were baseline corrected over the interval 0–300 ms and down-sampled by calculating five jumping means in several consecutive intervals beginning at 300 ms and ending between 1500 and 3500 ms (in multi-class 2500 ms). Optional a causal elliptic IIR low-pass filter at 2.5 Hz was applied to the signals beforehand for the binary classification. We call the whole algorithm which extracts feature vectors from slow potential shift features 'SUB', due to the fact that it mainly serves to subsample the signals.

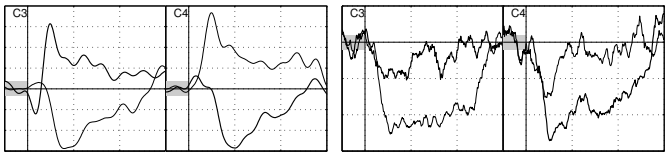

Fig. 3. MRP (left) and on 7–30 Hz bandpass-filtered ERD (right) curves (both spatial Laplace filtered) for subject *aa* in the time interval (*x*-axis) -500 ms to 3000 ms relative to stimulus on the channels C3 and C4. Thin and thick lines are averages over right or left hand trials respectively. The contralateral negativation resp. desynchronization is clearly observable.

To derive feature vectors for the ERD effects we use two different methods which reflect different aspects of brain rhythm modulations. The first (AR) considers the spectral distribution of the most prominent brain rhythms whereas the second (CSP) reflects spatial pattern distribution of the most prominent power modulation in specified frequency bands. A combination of both by using autoregressive models after calculating common spatial patterns is conceivable as a further strategy.

Autoregressive models (AR).

In an autoregressive model of order p each time point of a time series is represented as a fixed linear combination (AR coefficients) of the last p data points. The model order p is considered as a free parameter which was in our setting selected between 5 and 12 (in multi-class fixed to 8). The AR coefficients reflect oscillatory properties of the EEG signal, but do not contain the overall amplitude information. Accounting for this by adding the variance to the feature vector improves classification accuracy. To prevent the AR models from being distorted by EEG-baseline drifts, the signals were high-pass filtered at 4 Hz. In order to sharpen the spectral information to focal brain sources (spatial) Laplacian filters were applied. The interval for estimating the AR parameters started at 500 ms and the end points were chosen between 2000 ms and 3500 ms (in multi-class it was fixed to 2500 ms).

Common spatial patterns (CSP).

This method was originally suggested for binary classification of EEG trials in [11]. Projections with the most differing power-ratios in feature space are computed. These can be calculated by a simultaneous diagonalization of the covariance matrices of both classes. Only a few orientations with the highest ratio between their eigenvalues (in both directions) are selected. Note that this CSP approach can also be used for slow cortical potentials after some appropriate modifications for determining the covariance matrices, cf. [26]. First of all, to focus on effects in the α- and/or β-band (in multi-class only α) the signals were filtered between 8 and 13 Hz (for α),

[4]The choice of the optimal set-up is chosen as the best multi-class combination *without* reiterating the EEG experiment with this chosen set-up. Although, in principle, this could induce a bias, [16] used this pragmatic strategy in order to avoid repeating experiments without being able to use exactly the same conditions. It can also be seen as a pre-experiment to an online BCI multi-class session.

15 and 25 Hz (for β) or 7 and 30 Hz (for α and β) and the band with best classification performance was chosen. The number of CSPs used per class was a free parameter to be chosen between 2 and 4 in the binary case (in multi-class see V-E). The intervals of interest were chosen as described above for the AR model. Feature vectors consisted of the variances of the CSP projected trials, cf. [11]. A further performance gain can be achieved by calculating the logarithm of these variances. Note that the CSP algorithm depends on the label. Consequently this could result in overfitting, i.e. underestimating the classification error, if we did this algorithm beforehand on all trials – similar to training a classifier on all trials and calculating the training error as test error. To avoid this overfitting one has to calculate the common spatial patterns only on the training set in cross-validation, to use these patterns to project the test set on a lower number of channels and to determine the test error with the further processing (calculation of variance) and application of the classifier. Finally, the CSP algorithm allows to neglect regularization, since only very low dimensional feature vectors are left with comparatively high numbers of samples.

D. Combination algorithms

Feature combination or sensor fusion strategies are rather common in speech recognition (e.g. [27]) or vision (e.g. [28]) or robotics (e.g. [29]) where either signals on different time-scales or from distinct modalities need to be combined. Typical approaches use a concatenation of the single feature vectors (discussed as CONCAT below), or a winner-takes-all strategy, which however cannot increase performance above the best single feature vector analysis. Furthermore, combinations that use a joint probabilistic modeling [28] appear promising, but were not tested in the framework of this paper. We propose two further methods that incorporate independence assumptions (PROB and to a smaller extent META) and allow individual decision boundary fitting to single feature vectors[5] (META). In this Sectionwe will only outline the algorithms for binary classification on labels ± 1. Extension of these strategies to multi-class is straightforward.

(CONCAT). Here classification is applied to the concatenation of all single feature vectors. Note that careful regularization is necessary to account for the increased dimensionality [23], [20].

(PROB). We start with a set of N feature vectors described by random variables X_j for $j = 1,\ldots,N$ with binary class labels $Y \in \{\pm 1\}$. Furthermore, for each feature vector X_j an optimal classifier f_j on the single feature vector space D_j, i.e. which minimizes the misclassification risk, is given. Denoting $g_{j,y}$ as the densities of $P(f_j(X_j)|Y = y)$ for each feature vector X_j and class label $y = \pm 1$, f as the optimal classifier on the combined feature vector space $D = (D_1,\ldots,D_N)$, X as the combined random variable $X = (X_1,\ldots,X_N)$ and g_y as densities of $P(f(X)|Y = y)$, this means under the assumption of equal class priors that for $x = (x_1,\ldots,x_N) \in D$

$$f_j(x_j) = 1 \quad \Leftrightarrow \quad \hat{f}_j(x_j) := \log \frac{g_{j,1}(x_j)}{g_{j,-1}(x_j)} > 0,$$
$$f(x) = 1 \quad \Leftrightarrow \quad \hat{f}(x) := \log \frac{g_1(x)}{g_{-1}(x)} > 0.$$

The assumption of independence between the feature vectors allows us to factorize the combined density, i.e. to compute $g_y(x) = \prod_{j=1}^{N} g_{j,y}(x_j)$ for the class labels $y = \pm 1$. This leads to the optimal decision function

$$f(x) = 1 \Leftrightarrow \hat{f}(x) = \sum_{j=1}^{N} \hat{f}_j(x_j) > 0.$$

In our application, where we assume additionally that all feature vectors X_j are Gaussian distributed with equal covariance matrices, i.e. $P(X_j \,|\, Y = y) = \mathcal{N}(\mu_{j,y}, \Sigma_j)$, and $w_j := \Sigma_j^{-1}(\mu_{j,1} - \mu_{j,-1})$, we get the following classifier

$$f(x) = 1 \Leftrightarrow \sum_{j=1}^{N} [w_j^\top x_j - \frac{1}{2}(\mu_{j,1} + \mu_{j,-1})^\top w_j] > 0.$$

In terms of LDA this corresponds to forcing the elements of the estimated covariance matrix that belong to different feature vectors to zero. Consequently, less parameters have to be estimated and distortions by accidental correlations of independent variables are avoided. It should be noted that a non-linear version of PROB with a gaussian assumption for each feature vector can be formulated analogously to quadratic discriminant analysis (QDA), cf. [24]. To avoid overfitting we have to regularize PROB, and there are two ways feasible ways of doing so: Regularization of the covariance matrices with one global parameter (PROBsame) or with separately selected parameters corresponding to the single-type features (PROBdiff) as described in equation (3).

Note that PROB differs from the combination algorithm of section III by the fact that it does not contain the normalization.

(META). After training the individual classifiers on each single feature vector beforehand a meta level classifier is applied to their continuous output. Although this allows a custom-made choice of classifiers for each feature vector which can be useful, e.g. if the decision boundary is linear for one feature vector and non-linear for another, we simply use regularized LDA for each of the feature vectors, and select the regularization coefficients for each single feature vector individually, i.e. each classifier is individually regularized. For the meta level classifier that combines the single classifier results we find that regularization is not needed anymore in practice, since the meta classifier acts on very low dimensional feature vectors.

When we use LDA as the classifier or in general the logarithm of the quotient of both class densities the difference between PROB and META consists of the fact that PROB simply sums up all individual single feature vector classifiers, whereas META additionally learns a weighting between all classifiers and uses this for decision making. Moreover, META allows learning of a bias which can usually be neglected.

[5] i.e. to SUB, AR, CSP as described in Section V-C

Consequently, META extracts discriminative information from single feature vectors independently and the Meta classification may exploit inter-relations (also, for example, hidden dependencies) based on the output of the individual decision functions. Hence possible high level relations are taken into account while independence is assumed on a low level.

E. CSP multi-class extensions

An extension of the CSP algorithm to multi-class paradigms has been previously considered only in [30]. After a short description of this algorithm, which we subsequently refer to as IN, we will in the following suggest two new methods for multi-class CSP beyond it:

Using CSP within the classifier (IN):. This algorithm reduces a multi-class to several binary problems (cf. [31]) and was suggested in [30] for CSP in a BCI context. Therefore CSP is only used in its binary version such that the variances of the projections to the CSPs are employed as inputs for an LDA-classifier for each 2-class combination. New trials are projected on these CSPs and are assigned to the class for which most classifiers are voting.

One versus the rest CSP (OVR):. We suggest a subtle modification of the approach above which permits us to compute the CSP approach before the classification. We compute spatial patterns for each class against all others[6]. An LDA multi-class classification is then performed on the variances of the projections of the EEG signals on all these CSP patterns. The OVR approach appears rather similar to the approach IN, but there is in fact a large practical difference. OVR does multi-class classification on all projected signals whereas IN does binary classification on the CSP patterns according to the binary choice.

Simultaneous diagonalization (SIM):. In the binary case, the CSP algorithm finds a simultaneous diagonalization of both covariance matrices whose eigenvalues sum to one. Thus a possible extension to many classes, i.e. many covariances $(\Sigma_i)_{i=1,...,N}$ is to find a matrix R and diagonal matrices $(D_i)_{i=1,...N}$ with elements in $[0,1]$ and satisfying $R\Sigma_i R^T = D_i$ for all $i = 1,...,N$ and $\sum_{i=1}^{N} D_i = I$. Such a decomposition can be done exactly for $N = 2$; for $N > 2$, in general, only approximative solutions can be obtained. Several algorithms exist for approximate simultaneous diagonalization (cf. [32], [33], [34]), we use the recent algorithm described in [34] due to its speed and reliability. As opposed to the two class problem, there is no canonical way to choose the relevant CSP patterns for multi-class CSP. We explored several options such as using the highest or lowest eigenvalues. We discovered that the best strategy was based on the assumption that two different eigenvalues for the same pattern have the same effect if their ratios to the mean of the eigenvalues of the other classes are multiplicatively inverse to each other, i.e. their product is 1. Thus all eigenvalues λ are mapped to[7] $\text{score}(\lambda) := \max(\lambda, 1/(1+(N-1)^2\lambda/(1-\lambda)))$ and a specified number m of highest scores for each class are used as CSP patterns. It should be mentioned that each pattern is only used once, namely for the class which has the highest score. If a second class chooses the same pattern it is left out for this class and the next one, i.e. with the next highest score for this class, is chosen. Finally conventional LDA multi-class classification is performed on the variances of the projected trials as before.

For the purpose of this paper we will use these combination methods in two directions: First of all, we will show that a performance gain can be observed in a BCI context when using SIM and OVR against IN. Second, we will use these algorithms together with Feature Combination in the multi-class experiments to increase the ITR further.

Note that in order to evaluate the performance of a BCI system one should regard the Information Transfer Rate per minute and not per decision for ITR. In this case, where the trials have a fixed (mean) duration of 4.5 s all values have to be multiplied by $60/4.5 \sim 13$ to get the ITR per minute. However, this is not done due to the following two reasons: 1. It does not have any influence on the results and comparisons in this paper since the rate for all experiments is constant. 2. The decision rate is chosen to be very small to be sure that we can train suitable classifiers. Due to the fact that decisions can be made faster in feedback experiments, once a classifier of suitable quality is trained, real performance is only meaningful there. Note that it is not enough to consider only the time interval used for classification since the intermediate period (e.g. relaxing) could also be important to get a good performance.

VI. RESULTS

To confirm our hypothesis that the chosen feature vectors are independent to a sufficient degree we investigate the following issues. Firstly, we calculate the correlation matrix of the concatenated feature vectors exemplarily for subject *aa* (cf. Fig. 4 left). Here correlations within each feature vector are observable on the block diagonal, whilst weak inter-feature correlation is visible in the non-block-diagonal fields. Secondly, we classify each trial of each suitable 2-class subset experiment in a leave-one-out cross-validation (e.g. [35]). We are hereby able to calculate for all correctly classified (or misclassified) trials of one feature vector what the distribution of correctly classified (or misclassified) trials are for another feature vector. This is is visualized in the center of Fig. 4. We do this for all 2-class subsets of all experiments and calculate the mean of these distributions. If all feature vectors are independent, the bars should be of the same size. Thirdly, we calculate the 3×3-correlation matrix of both the continuous output of the leave-one-out approach and the error vector resulting from this procedure for the different feature vectors (cf. Fig. 4 right). All these pictures together reveal independence between the feature vectors SUB and AR or CSP and only a weak dependence between AR and CSP. So a high gain by a suitable feature combination can be expected.

In Fig. 5 the suggested combination algorithms applied to all three feature vectors are compared to the best single feature vector result for all 2 and multi-class combinations out of the

[6]Note that this can be done similarly with pairwise patterns, but in our studies no advantage was observed and therefore one-versus-the-rest is favorable, since it chooses less patterns.

[7]For $N = 2$ this results in $\max(\lambda, 1-\lambda)$.

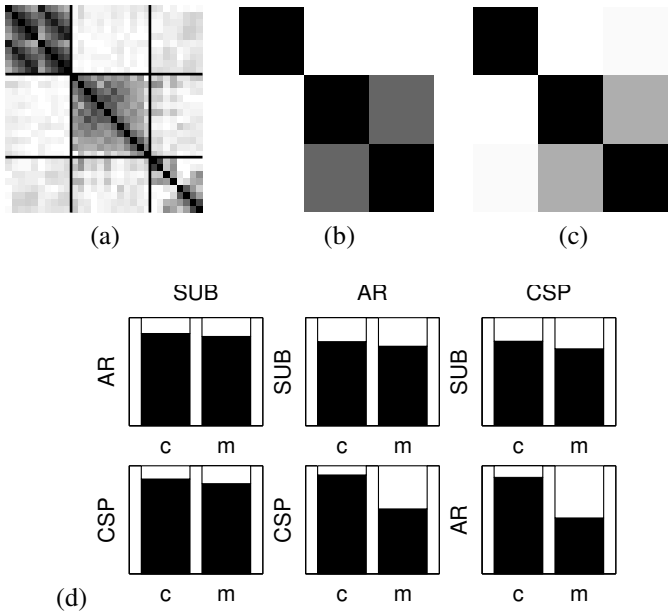

Fig. 4. In subplot (a) the absolute values of the correlation matrix for subject *aa* for (selected dimensions of) the feature-concatenated vectors are plotted as an image, where white points are values close to zero. The block order from top to bottom and left to right is SUB, AR, CSP. The two 3×3 matrices in (b) and (c) visualize the correlation matrix between the feature vectors (b) with respect to the real-valued leave-one-out output of each trial and (c) with respect to the leave-one-out error vector for each trial. In both cases the order is given by (SUB, AR, CSP) (from top to bottom and left to right). The bars in (d) show for each feature (indicated on the top of the bars) for the correct classified ('c') and misclassified ('m') feature vectors in a leave-one-out approach the portion of correct classified (black) and misclassified (white) in another feature vector (written to the left of each bar) as mean of all binary subsets of classes for all experiments.

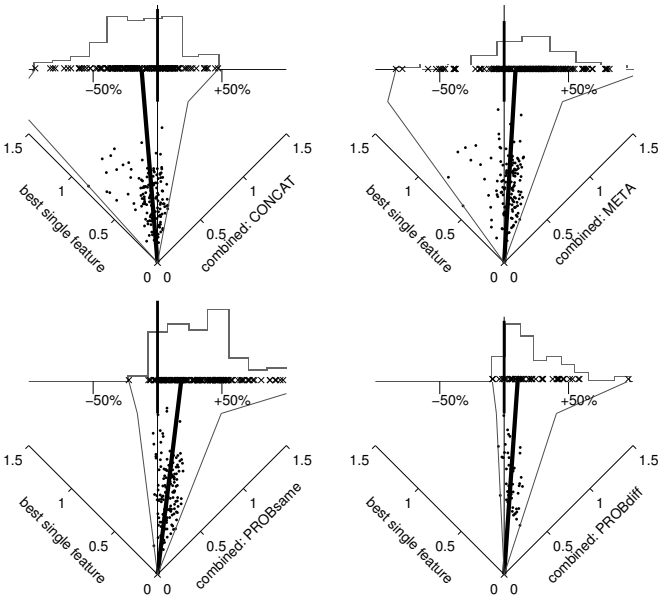

Fig. 5. The visualized scatter plots show the ITR on the best single feature vector based classifier against the presented combination methods for all 2 and multi-class combinations of all experiments except the ones described above. Above each scatter plot a histogram of the increase in percent in ITR compared to the best single feature vector is plotted. For points right of the vertical line through 0 in each scatter plot the combination algorithm outperforms the best single feature vector. The fat line shows the regression line of the points through the zero point calculated by minimizing the squared loss.

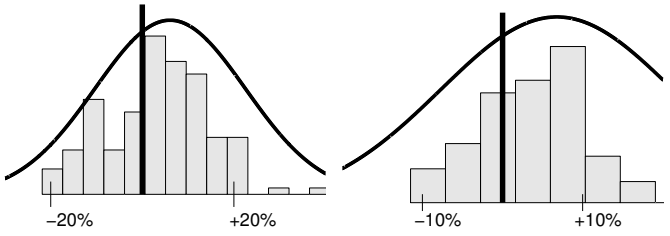

Fig. 6. The visualized histograms show the decrease in % in classification error of algorithms SIM (left) resp. OVR (right) compared to IN over all 95 multi-class combinations. For values higher than 0 SIM resp. OVR outperform IN. Furthermore an approximation of the histogram by a gaussian distribution is plotted.

presented experiments, except for the excluded cases described above. Typically the standard algorithm CONCAT does not increase performance, due to the fact that the dimensionality of the problem increases enormously and therefore estimation of the huge covariance matrices is rather error-prone. Furthermore, a small gain for the algorithm META against the best single feature vector result is observable. PROBsame and PROBdiff usually perform best and reveal a high performance gain compared to the best single feature vector result. Only a small advantage of PROBdiff is detectable (where only the binary results were taken into account). However, PROBsame might still be preferred due to the fact that the time to train the classifier is much shorter. Note that in the multi-class case PROBdiff was not calculated due to the computational complexity, therefore in this part of the figure less points are plotted. The scatter plots clearly exhibit visible superiorities (in 144 results[8]). Applying significance analyses here is somewhat problematic. Since all possible subsets of given set of mental states are considered, classification is done, e.g., for classes $\{l,r,f\}$ and $\{l,a,t\}$ of the same experiment. These observations are clearly dependent since the trials of class l are involved in both. This means that the assumption of independent observations is violated. With this caveats in mind we employ a test of significance in analogy to [36].

[8]In Fig. 5 every point corresponds to the ITR for one chosen subset of classes in the multi-class paradigm, i.e., for one possible BCI setting.

The Wilcoxon Rank test yields that META, PROBsame and PROBdiff significantly exceed the best single feature with $p < 0.1$, $p < 0.0001$, and $p < 0.01$ respectively. In contrast CONCAT performs worse than the best single feature with $p < 0.005$.

Fig. 6 reveals the predominance of the algorithms SIM and OVR over IN, although in some cases IN still performs better. Modification of IN to one-vs-the-rest classification do not change this result. With the Wilcoxon Rank Test and $p < 0.05$ we see that SIM and OVR outperform IN significantly whilst a significant difference between SIM and OVR is not visible. We should conclude that SIM and OVR are to be preferred to IN.

Finally, in Fig. 7 the ITRs for different numbers of classes for each subject are visualized. For each case we choose the best configuration out of all tested ones and compare this to the best result without combination. Furthermore, we add

the theoretical gain given by equation (1) for the two class subsets. We see that in each case the combination algorithm shows enhanced performance, but that the theoretical gain can not be achieved, presumably due to the weak dependencies between AR and CSP feature vectors. If we now compare the results for a different number of classes we should take into consideration that an extensive model selection was performed for the binary case. For all configurations with more than two classes a fixed set of hyper-parameters was chosen and the setup could therefore be occasionally suboptimal. Thus, further improvements are perceivable, particularly if we do a similarly extensive hyperparameter search.

Nevertheless, if we take the results of the combination methods into consideration, then in all cases where we have enough classes of suitable pairwise discrimination a gain is observable when using more than two classes, except for subject *ar*. However, the full setting of all classes in the two 6 class experiments should not be chosen, since there are big differences in the pairwise discrimination results and therefore a suitably chosen subset results in a higher ITR. Note that in both 6 class experiments the highest gain is achieved with four classes. Interestingly, this is in contrast to the results without using our combination methods where the highest gain is found in a three class setup, a results that was also found in [16], or in some cases with a 2-class setup.

For subjects *aw* and *ay* a high discrepancy in discriminability between SUB feature vectors (>20 % in the binary case) on one hand and the AR or CSP feature vectors (<10 % in the binary case) on the other is observable. The gain here by combination is small which can be expected under these circumstances. Nevertheless, a gain or at least similar results were achieved with the feature combination. Consequently, combination also works in settings where some single classifiers are very powerful, and others are not. But is not recommend, due to the small possible increase on one hand and higher complexity on the other.

The same is true for subject *ar*, but here the slow potentials were disturbed by high drifts due to measurement problems. A repeated experiment should show if feature combination can also help to increase performance for this subject.

VII. Discussion

We pursued the path towards enhanced bit transfer rates in EEG-based BCI technology by using: (a) feature combination and (b) multi-class paradigms. This paper includes new algorithmic aspects such as the development of new feature combination strategies and a new algorithm that fully generalizes previous work on CSP ([11]) to multi-class problems as well as a theoretical contribution of how much can be gained when using more than two decision alternatives for BCI. We show that across a number of subjects both strategies are successful in boosting ITRs.

The use of a combination of feature vectors of independent physiological nature has already been suggested several times before [9], [10]. Nevertheless our work (see also [8]) is the first to pursue this interesting aspect in detail, showing that bit rate gains as high as 50% can be achieved when *learning* the appropriate combination (PROB) between Bereitschaftspotential, desynchronization dynamics, and spatial maps. Although the physiological processes might appear independent from the medical point of view, it was not initially clear that we could confirm this independence assumption on the raw data level and thus gain from this fact when classifying, even if not all feature vectors are fully independent, e.g. AR and CSP.

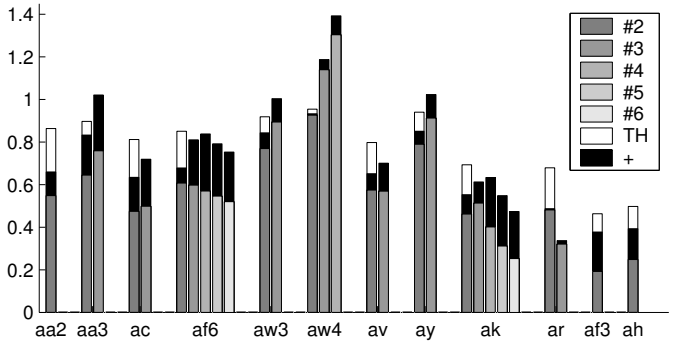

Fig. 7. The bar plot visualizes the highest ITRs for all algorithms without combination (all colors except black) presented here for all subjects for different numbers of classes from 2 (dark gray, #2) to 6 (light gray, #6). As a prolongation of each color bar we show the performance gain achieved with a combination method in black (+) and for the two class subsets the gain in white (TH) which theoretically can be achieved by formula (1) if feature vectors are perfectly independent. The number behind the subject code specifies the number of classes used for the specific experiment for subjects who took part in more than one experiment.

In this paper we have been using an offline set-up for our evaluations. The next step would be to provide an online feedback based on these methods. In our first experiments with this approach we train a classifier offline after a short training phase of about 30 minutes as described above, and then present continuous feedback under ongoing classification on the measured EEG. First approaches based only on CSP feature vectors (both binary and multi-class) show good performance and furthermore the opportunity to achieve much higher rates than 4.5 seconds for each decision. The ultimate challenge is now to fully transfer the proposed combination techniques to a real online feedback scenario.

Acknowledgments

We thank S. Harmeling, M. Kawanabe, A. Ziehe, G. Rätsch, S. Mika, P. Laskov, C. Schäfer, C. von Wrede and M. Krauledat for helpful discussions.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, pp. 767–791, 2002.

[2] M. Laubach, J. Wessberg, and M. Nicolelis, "Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task," *Nature*, vol. 405, no. 6786, pp. 523–525, 2000.

[3] P. Kennedy, R. Bakay, M. Moore, K. Adams, and J. Goldwaithe, "Direct control of a computer from the human central nervous system," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 198–202, 2000.

[4] G. A. Reina, D. W. Moran, and A. B. Schwartz, "On the relationship between joint angular velocity and motor discharge during reaching," *J. Neurophysiol.*, vol. 85, no. 6, pp. 2576–2589, 2001.

[5] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying single trial EEG: Towards brain computer interfacing," in *Advances in Neural Information Processing Systems (NIPS 01)*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Eds., vol. 14, 2002, pp. 157–164.

[6] B. O. Peters, G. Pfurtscheller, and H. Flyvbjerg, "Automatic differentiation of multichannel EEG signals," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 1, pp. 111–116, 2001.

[7] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, pp. 297–298, 1999.

[8] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Combining features for BCI," in *Advances in Neural Inf. Proc. Systems (NIPS 02)*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15, 2003, pp. 1115–1122.

[9] J. R. Wolpaw, D. J. McFarland, and T. M. Vaughan, "Brain-computer interface research at the Wadsworth Center," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 222–226, 2000.

[10] J. A. Pineda, B. Z. Allison, and A. Vankov, "The effects of self-movement, observation, and imagination on μ–rhythms and readiness potential (RP's): Toward a brain-computer interface (BCI)," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 219–222, June 2000.

[11] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 4, pp. 441–446, 2000.

[12] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, 2004, to appear.

[13] C. Babiloni, F. Carducci, F. Cincotti, P. M. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni, "Human movement-related potentials vs desynchronization of EEG alpha rhythm: A high-resolution EEG study," *NeuroImage*, vol. 10, pp. 658–665, 1999.

[14] C. Toro, G. Deuschl, R. Thather, S. Sato, C. Kufta, and M. Hallett, "Event-related desynchronization and movement-related cortical potentials on the ECoG and EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 93, pp. 380–389, 1994.

[15] R. Beisteiner, P. Hollinger, G. Lindinger, W. Lang, and A. Berthoz, "Mental representations of movements. Brain potentials associated with imagination of hand movements," *Electroencephalogr. Clin. Neurophysiol.*, vol. 96, no. 2, pp. 183–193, 1995.

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

[17] A. Mood, F. Graybill, and D. Boes, *Introduction to the theory of statistics*. McGraw-Hill Book Company, 1974.

[18] 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. Rehab. Eng.*, vol. 8, no. 2, pp. 164–173, 2000.

[19] L. Parra, C. Alvino, A. C. Tang, B. A. Pearlmutter, N. Yeung, A. Osman, and P. Sajda, "Linear spatial integration for single trial detection in encephalography," *NeuroImage*, vol. 7, no. 1, pp. 223–230, 2002.

[20] K.-R. Müller, C. W. Anderson, and G. E. Birch, "Linear and non-linear methods for brain-computer interfaces," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 2, 2003, 165–169.

[21] B. Blankertz, G. Dornhege, C. Schäfer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio, "Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 2, pp. 127–131, 2003.

[22] B. Blankertz, "BCI competition 2003 results (web page)." [Online]. Available: http://ida.first.fhg.de/projects/bci/competition/results/

[23] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, no. 2, pp. 181–201, May 2001. [Online]. Available: http://www.first.gmd.de/persons/Mueller.Klaus-Robert/review.ps.gz

[24] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[25] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for AdaBoost," *Machine Learning*, vol. 42, no. 3, pp. 287–320, 2001.

[26] G. Dornhege, B. Blankertz, and G. Curio, "Speeding up classification of multi-channel brain-computer interfaces: Common spatial patterns for slow cortical potentials," in *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering. Capri 2003*, 2003, pp. 591–594.

[27] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid hmm/connectionist approach," *Signal Processing Magazine*, pp. 25–42, 1995.

[28] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," 1996.

[29] S. Thrun, A. Bücken, W. Burgard, D. Fox, T. Fröhlinghaus, D. Henning, T. Hofmann, M. Krell, and T. Schmidt, "Map learning and high-speed navigation in RHINO," in *AI-based Mobile Robots*, D. Kortenkamp, R. Bonasso, and R. Murphy, Eds. MIT Press, 1998.

[30] J. Müller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, pp. 787–798, 1999.

[31] E. Allwein, R. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[32] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J.Mat.Anal.Appl.*, vol. 17, no. 1, p. 161 ff., 1996.

[33] D.-T. Pham, "Joint approximate diagonalization of positive definite matrices," *SIAM J. on Matrix Anal. and Appl.*, vol. 22, no. 4, pp. 1136–1152, 2001.

[34] A. Ziehe, P. Laskov, K.-R. Müller, and G. Nolte, "A linear least-squares algorithm for joint diagonalization," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 469–474.

[35] G. Wahba, *Splines Models for Observational Data*. Philadelphia: Series in Applied Mathematics, Vol. 59, SIAM, 1990.

[36] E. Curran, P. Sykacek, S. Roberts, W. Penny, M. Stokes, I. Jonsrude, and A. Owen, "Cognitive tasks for driving a brain computer interfacing system: a pilot study," *IEEE Trans. Rehab. Eng.*, 2003, in press.