Combined optimization of spatial and temporal filters for improving Brain-Computer Interfacing

Guido Dornhege, Benjamin Blankertz, Matthias Krauledat, Florian Losch, Gabriel Curio, Klaus-Robert Müller

Abstract— **Brain-Computer Interface (BCI) systems create a novel communication channel from the brain to an output device by bypassing conventional motor output pathways of nerves and muscles. Therefore they could provide a new communication and control option for paralyzed patients. Modern BCI technology is essentially based on techniques for the classification of single-trial brain signals. Here we present a novel technique that allows the simultaneous optimization of a spatial and a spectral filter enhancing discriminability rates of multi-channel EEG single-trials. The evaluation of 60 experiments involving 22 different subjects demonstrates the significant superiority of the proposed algorithm over to its classical counterpart: the median classification error rate was decreased by 11%. Apart from the enhanced classification, the spatial and/or the spectral filter that are determined by the algorithm can also be used for further analysis of the data, e.g., for source localization of the respective brain rhythms.**

Index Terms— **EEG, Event-Related Desynchronization, Brain-Computer Interface, Common Spatial Patterns, Single-Trial-Analysis**

I. Introduction

Brain-Computer Interface (BCI) research aims at the development of a system that allows direct control of, e.g., a computer application or a neuroprosthesis, solely by human intentions as reflected by suitable brain signals, cf. [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. We will be focussing on noninvasive, electroencephalogram (EEG) based BCI systems. Such devices can be used as tools of communication for the disabled or for healthy subjects that might be interested in exploring a new path of man-machine interfacing, say when playing BCI operated computer games. Furthermore BCI research helps to explain how different mental states are reflected in the brain and how the respective EEG patterns can be characterized. Therefore it contributes also to more general neuroscientific issues.

A classical approach to establish EEG-based control is to set up a system that is controlled by a specific EEG feature which is known to be susceptible to conditioning and to let the subjects learn the voluntary control of that feature, cf. [4]. In contrast, the Berlin Brain-Computer Interface (BBCI) uses well established motor competences in control paradigms and a machine learning approach to extract subject-specific discriminability patterns from high-dimensional features. This approach has the advantage that the long subject training needed in the operant conditioning approach is replaced by a short calibration measurement (20 minutes) and machine training (1 minute). The machine adapts to the specific characteristics of the brain signals of each subject, accounting for the high inter-subject variability. With respect to the topographic patterns of brain rhythm modulations the Common Spatial Patterns (CSP) (see [14]) algorithm has proven to be very useful to extract subject-specific, discriminative spatial filters. So far the frequency band on which the CSP algorithm operates is either selected manually or unspecifically set to a broad band filter, cf. [14], [6]. Thus a simultanenous optimization of a frequency filter with the spatial filter is highly desirable given the individual variability across different subjects. Recently, in [15] the CSSP algorithm was presented, in which very simple frequency filters (with one delay tap) for each channel are optimized together with the spatial filters. Although the results showed an improvement of the CSSP algorithm over CSP, the flexibility of the frequency filters is still very limited. Here we present a method that allows to simultaneously optimize an arbitrary FIR filter within the CSP analysis on BCI data. The proposed algorithm outperforms CSP and CSSP on average, and for certain data sets (where a separation of the discriminative rhythm from dominating non-discriminative rhythms is of importance) a considerable increase of classification accuracy can be achieved. We would like to stress however that the new CSSSP (Common Sparse Spectral Spatial Pattern) method is *by no means* limited to BCI applications. On the contrary it is a completely generic new signal processing technique that is applicable for all general single trial EEG settings that require discrimination between EEG states based on modulations of brain rhythms.

II. Experimental Setup

In this paper we investigate data from 60 EEG experiments with 22 different subjects. All experiments included so called calibration sessions without feedback which are used to train subject-specific classifiers. Many experiments also included feedback sessions in which the subject could steer a cursor or play a computer game like *brain-pong* by BCI control. Data from feedback sessions are not used in this a-posteriori study since they depend on an intricate interaction of the subject with the original classification algorithm.

This work is an extended version of [1]

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

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

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

The studies were supported by a grant of the *Bundesministerium für Bildung und Forschung* (BMBF), FKZ 01IBE01A/B. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778 and by the *Deutsche Forschungsgemeinschaft* (DFG), FOR 375/B1. This publication only reflects the authors' views.

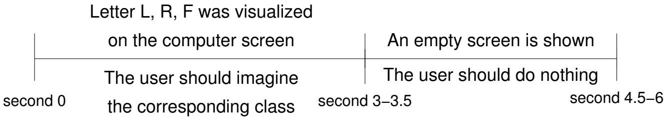

Fig. 1: The figure shows the time structure of a sequence in an experiment. Approximately 100 sequences per session and 3-4 session were recorded. At time point zero a letter is shown to the user on a computer screen. The user should start immediately with imagination of the class corresponding to this letter and stop doing so when the letter disappears after 3–3.5 seconds. After further 1.5–2.5 seconds a new letter is shown.

In the experimental sessions used for the present study, labeled trials of brain signals were recorded in the following way: The subjects were sitting in a comfortable chair with arms lying relaxed on the armrests. All 4.5–6 seconds one of 3 different visual stimuli indicated for 3–3.5 seconds which mental task the subject should accomplish during that period (cf. Fig. 1). The investigated mental tasks were imagined movements of the left hand (*l*), the right hand (*r*), and one foot (*f*). Note that in a few experiments only two mental tasks were used. Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 32, 64 resp. 128 channels. Besides EEG channels, we recorded the electromyogram (EMG) from both forearms and the leg as well as horizontal and vertical electrooculogram (EOG) from the eyes. The EMG and EOG channels were used exclusively to make sure that the subjects performed no real limb or eye movements correlated with the mental tasks that could directly (artifacts) or indirectly (afferent signals from muscles and joint receptors) be reflected in the EEG channels and thus be detected by the classifier, which operates on the EEG signals only. Between 120 and 200 trials for each class were recorded. In this study we investigate only binary classifications, but the results can be expected to safely transfer to the multi-class case.

III. Neurophysiological Background

According to the 'homunculus' model, first described in [16], for each part of the human body exists a corresponding region in the primary motor and primary somatosensory area of the neocortex. The 'mapping' from the body part to the respective brain areas approximately preserves topography, i.e., neighboring parts of the body are represented in neighboring parts of the cortex. For example, while the feet are located close to the vertex, the left hand is represented lateralized (by about 6 cm from the midline) on the right hemisphere and the right hand almost symmetrically on the left hemisphere.

Macroscopic brain activity during resting wakefulness contains distinct 'idle' rhythms located over various brain areas, e.g., the μ-rhythm can be measured over the pericentral sensorimotor cortices in the scalp EEG, usually with a frequency of about 10 Hz ([17]). Furthermore, in electrocorticographic recordings Jasper and Penfield ([16]) described a strictly local β-rhythm at about 20 Hz over the human motor cortex. In non-invasive scalp EEG recordings the 10 Hz μ-rhythm is commonly mixed with the 20 Hz-activity. Basically, these rhythms are cortically generated; while the involvement of a thalamo-cortical pacemaker has been discussed since the first description of EEG by Berger ([18]), Lopes da Silva ([19]) showed that cortico-cortical coherence is larger than thalamo-cortical pointing to a convergence of subcortical and cortical inputs.

The moment-to-moment amplitude fluctuations of these local rhythms reflect variable functional states of the underlying neuronal cortical networks and can be used for brain-computer interfacing. Specifically, the pericentral μ- and β-rhythms are diminished, or even almost completely blocked, by movements of the somatotopically corresponding body part, independent of their active, passive or reflexive origin. Blocking effects are visible bilateral but with a clear predominance contralateral to the moved limb. This attenuation of brain rhythms is termed event-related desynchronization (ERD), see [20].

Since a focal ERD can be observed over the motor and/or sensory cortex even when a subject is only imagining a movement or sensation in the specific limb, this feature can well be used for BCI control: The discrimination of the imagination of movements of left hand vs. right hand vs. foot can be based on the somatotopic arrangement of the attenuation of the μ and/or β-rhythms. To this end, different ways to improve the classification performance of the CSP algorithm were suggested, e.g., [15].

There is another feature independent from the ERD reflecting imagined or intended movements, the movement related potentials (MRP), denoting a negative DC shift of the EEG signals in the respective cortical regions. See [21], [22] for an investigation of how this feature can be exploited for BCI use and combined with the ERD feature. This combination strategy was able to greatly enhance classification performance in offline studies. In this paper we focus only on improving the ERD-based classification, but all the improvements presented here can also be used in the combined algorithm.

There are two problems when using ERD features for BCI control:

(1) The strength of the sensorimotor idle rhythms as measured by scalp EEG is known to vary strongly between subjects. This introduces a high intersubject variability on the accuracy with which an ERD-based BCI system works.

(2) The precentral μ-rhythm is often superimposed by the much stronger posterior α-rhythm, which is the idle rhythm of the visual system. It is best articulated with eyes closed, but also present in awake and attentive subjects, see Fig. 2 at channel Pz. Due to volume conduction the posterior α-rhythm interferes with the precentral μ-rhythm in the EEG channels over motor cortex. Hence a μ-power based classifier is susceptible to modulations of the posterior α-rhythm that occur due to fatigue, change in attentional focus while performing tasks, or changing demands of visual processing. When the two rhythms have different spectral peaks as in Fig. 2, channels Cz and C4, a suitable frequency filter can help algorithms that optimize spatial filters to find the more discriminative spectral peak. The subject specific optimization of such a filter integrated in the CSP algorithm is addressed in this paper.

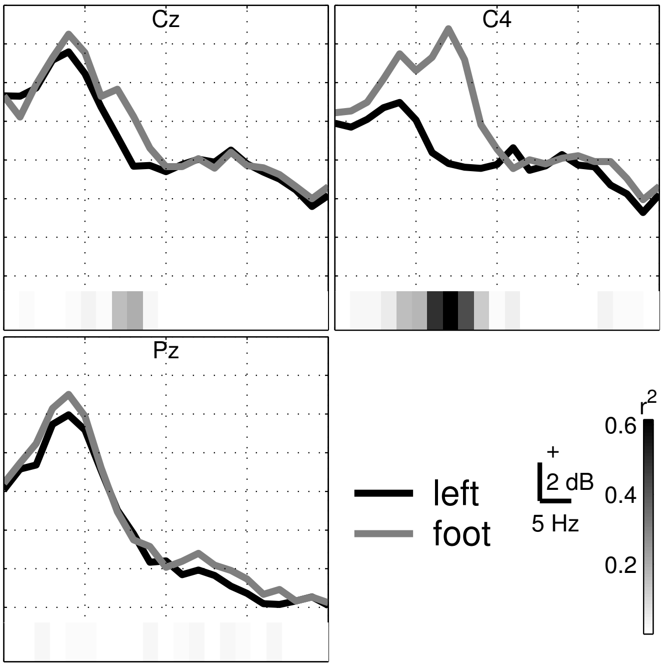

Fig. 2: The plot shows exemplarily the spectral energy ([dB] on the y-axis) for one subject during left hand (dark line) and foot (light line) motor imagery between 5 and 25 Hz (x-axis) at scalp positions Pz, Cz and C4. In both central channels two peaks, one at 8 Hz and one at 12 Hz are visible during foot imagery. The first peak seems to reflect mainly the idle rhythm of the visual system (α-rhythm) whereas the latter peak reflects the idle rhythm of the sensory motor area of the left hand which is present during foot imagery (not involving the left hand) and blocked during left hand imagery. Below each channel the r^2-value which measures discriminability is shown. It clearly indicates that the peak around 12 Hz contains more discriminative information.

IV. Spatial Filter - the CSP Algorithm

The common spatial pattern (CSP) algorithm ([23]) is very useful when calculating spatial filters for detecting ERD effects ([24]) and for ERD-based BCIs, see [14], and has been extended to multi-class problems in [25]. Given two distributions in a high-dimensional space, the (supervised) CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and at the same time minimize variance for the other class. After having band-pass filtered the EEG signals to the rhythms of interest, high variance reflects a strong rhythm and low variance a weak (or attenuated) rhythm. Let us take the example of discriminating left hand vs. right hand imagery. According to Section III, the spatial filter that focusses on the area of the left hand is characterized by a strong motor rhythm during imagination of right hand movements (left hand is in idle state), and by an attenuated motor rhythm during left hand imagination.

This criterion is exactly what the CSP algorithm optimizes: maximizing variance for the class of right hand trials and at the same time minimizing variance for left hand trials. Furthermore the CSP algorithm calculates the dual filter that will focus on the area of the right hand in sensor space. Moreover a series of orthogonal filters of both types can be determined.

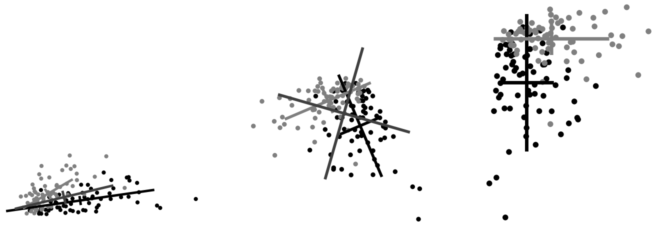

Fig. 3: The figures illustrates the calculation of CSP patterns. The plot on the left shows the original data distributions for two classes (light gray and black points). The respective means and covariance matrices are visualized by the principle axes. A mid gray cross indicates the distribution of the whole data set (both classes taken together). The latter distribution is whitened (linear projection such that the variance in evary direction is 1), see plot in the middle. Here the principle axes of the class distributions are perpendicular, cf. eqn 3. A suitable rotation makes these axes coincide with the coordinate axes.

The CSP algorithm is trained on labeled data, i.e., we have a set of trials s_i, $i = 1, 2, ...$, where each trial is represented as a real-valued matrix of several channels (as rows) and time points (as columns). A spatial filter $w \in \mathbb{R}^{\#channels}$ projects these trials to the signal $w^\top s_i$ with only one channel. The idea of CSP is to find a spatial filter w such that the projected signal has high *power* for one class and low *power* for the other. Here the power for a trial is calculated by the variance in the time domain. Obviously simultaneous maximization of one term and minimization of another term require instructions how to do so. An important observation helps here, namely that only the direction of the spatial solution w and not the length is important for further calculation and interpretation. Consequently this independence of a solution for w from scaling allows to fix the length of w to some arbitrary value, say that the sum of the variances of all projected trials and both classes is fixed to 1. In doing so maximization of the sum of the variances of all projected trials of one class directly leads to the minimization of the sum of the variances of all projected trials of the other class, since the sum of both is constant. Formally this is expressed by the following optimization problem:

$$\max_w \sum_{i:\text{Trial in Class 1}} \mathrm{var}(w^\top s_i), \quad \text{s.t.} \quad \sum_i \mathrm{var}(w^\top s_i) = 1, \qquad (1)$$

where $\mathrm{var}(\cdot)$ is the variance of the vector. An analogous formulation can be given for the second class.

Using the definition of the variance we simplify the problem to

$$\max_w \; w^\top \Sigma_1 w, \quad s.t. \quad w^\top(\Sigma_1 + \Sigma_2)w = 1, \qquad (2)$$

where Σ_y is the covariance matrix of the trial-concatenated matrix of dimension [#channels × #concatenated time-points] belonging to the respective class $y \in \{1, 2\}$.

Formulating the dual optimization problem we see that the problem can be solved by calculating a matrix Q and diagonal matrix D with elements in $[0, 1]$ such that

$$Q\Sigma_1 Q^\top = D \qquad \text{and} \qquad Q\Sigma_2 Q^\top = I - D \qquad (3)$$
$$(\Rightarrow Q(\Sigma_1 + \Sigma_2)Q^\top = I)$$

and by choosing the highest diagonal element of D and the corresponding row vector of Q. This row vector is the solution w for equation (2).

A solution of equation (3) can be revealed in the following way (see Fig. 3 for a visualization). First *whiten* the matrix $\Sigma_1+\Sigma_2$, i.e., determine a matrix P such that $P(\Sigma_1+\Sigma_2)P^\top = I$ which is possible due to positive definiteness of $\Sigma_1+\Sigma_2$. Then define $\hat{\Sigma}_y = P\Sigma_y P^\top$ and calculate an orthogonal matrix R and a diagonal maxtrix D by eigenvalue theory such that $\hat{\Sigma}_1 = RDR^\top$. Therefore $\hat{\Sigma}_2 = R(I-D)R^\top$ since $\hat{\Sigma}_1+\hat{\Sigma}_2 = I$ and $Q := R^\top P$ satisfies (3). The projection that is given by the j-th row of matrix R has a relative variance of d_j (j-th element of D) for trials of class 1 and relative variance $1-d_j$ for trials of class 2. If d_j is near 1 the filter given by the j-th row of R maximizes variance for class 1, and since $1-d_j$ is near 0, minimizes variance for class 2. Typically one would retain some projections corresponding to the highest eigenvalues d_j, i.e., CSPs for maximizing variance for trials of class 1 and minimizing variance for class 2, and some corresponding to the lowest eigenvalues, i.e., CSPs with the opposite property.

V. Combined Spectral and Spatial Filter

As discussed in Section III the content of discriminative information in different frequency bands is highly subject-dependent. For example the subject whose spectra are visualized in Fig. 2 shows a highly discriminative peak at 12 Hz whereas the peak at 8 Hz does not show good discrimination. Since the lower frequency peak has high band energy a better performance in classification can be expected, if we reduce the influence of the lower frequency peak for this subject. However, for other subjects the situation looks different, i.e., the classification might fail if we exclude this information. Thus it is desirable to optimize a spectral filter for better discriminability. Here are two approaches to this task.

CSSP. In [15] the following was suggested: Given s_i the signal s_i^τ is defined to be the signal s_i delayed by τ timepoints with respect to the sampling rate. In CSSP the usual CSP approach is applied to the concatenation of s_i and s_i^τ in the channel dimension, i.e., the delayed signals are treated as new channels. By this concatenation step the algorithm is able to neglect or emphasize specific frequency bands. Of course, this strongly depends on the choice of τ which can be accomplished by some validation approach on the training set. More complex frequency filters can be found by concatenating more delayed EEG-signals with several delays. In [15] it was concluded that in typical BCI situations where only small training sets are available, the choice of only one delay tap is most effective in the CSSP approach. The increased flexibility of a frequency filter with more delay taps does not trade off the increased complexity of the optimization problem.

CSSSP. The idea of our new CSSSP algorithm is to learn a complete global spatial-temporal filter in the spirit of CSP and CSSP.

A digital frequency filter consists of two sequences a and b with length n_a and n_b such that the signal x is filtered to y by

$$\begin{aligned}a(1)y(t) &= b(1)x(t)+b(2)x(t-1)+\ldots+b(n_b)x(t-n_b-1)\\ &- a(2)y(t-1)-\ldots-a(n_a)y(t-n_a-1)\end{aligned}$$

where t denotes the index in the time series. Here we restrict ourselves to FIR (finite impulse response) filters by defining $n_a = 1$ and $a = 1$. Furthermore we define $b(1) = 1$ and fix the length of b to some T with $T > 1$. By this restriction we resign some flexibility of the frequency filter but it allows us to find a suitable solution in the following way: We are looking for a real-valued sequence $b_{1,\ldots,T}$ with $b(1) = 1$ such that the trials

$$s_{i,b} = s_i + \sum_{\tau=2,\ldots,T} b_\tau s_i^\tau \tag{4}$$

can be classified better in some way.

Using equation (1) we have to solve the problem

$$\max_{w,b,b(1)=1} \sum_{i:\text{Trial in Class 1}} \operatorname{var}(w^\top s_{i,b}), \quad \text{s.t.} \quad \sum_i \operatorname{var}(w^\top s_{i,b}) = 1. \tag{5}$$

Let us define by $\Sigma_y^\tau := E(\langle s_i(s_i^\tau)^\top + s_i^\tau s_i^\top | i : \text{Trial in Class } y\rangle)$ for $\tau > 0$ and $\Sigma_y^0 := E(\langle s_i s_i^\top | i : \text{Trial in Class } y\rangle)$, namely the correlation between the signal and the by τ delayed signal. Since we can assume that $E(\langle s_i^\tau s_i^\top, | i : \text{Trial in Class } y\rangle) \approx E(\langle s_i^{\tau+j}(s_i^j)^\top, | i : \text{Trial in Class } y\rangle)$ for small $j > 0$, equation (5) can be approximately simplified to

$$\begin{aligned}\max_{b,b(1)=1} \max_{w} \quad & w^\top \left(\sum_{\tau=0}^{T-1}\left(\sum_{j=1}^{T-\tau} b(j)b(j+\tau)\right)\Sigma_1^\tau\right)w,\\ s.t. \quad & w^\top \left(\sum_{\tau=0}^{T-1}\left(\sum_{j=1}^{T-\tau} b(j)b(j+\tau)\right)\left(\Sigma_1^\tau+\Sigma_2^\tau\right)\right)w = 1.\end{aligned} \tag{6}$$

With the usual CSP techniques we can calculate the optimal w for each b (see equation (2) and (3)). Since $b(1) = 1$, a $(T-1)$-dimensional problem remains which can be solved using optimization techniques like gradient or line-search methods if T is not too large.

Consequently we get for each class a frequency band filter and a pattern (or similar to CSP more than one pattern by choosing the next eigenvectors).

However, with increasing T the complexity of the frequency filter has to be controlled in order to avoid overfitting. One way to restrict the complexity of a solution is to enforce a sparse solution for b, i.e. a solution for b with only a few non-zero entries. Sparsity of b is achieved by introducing a regularization term in the following way:

$$\begin{aligned}\max_{b,b(1)=1} \max_{w} \; & w^\top \left(\sum_{\tau=0}^{T-1}\left(\sum_{j=1}^{T-\tau} b(j)b(j+\tau)\right)\Sigma_1^\tau\right)w - C/T\|b\|_1,\\ s.t. \; & w^\top \left(\sum_{\tau=0}^{T-1}\left(\sum_{j=1}^{T-\tau} b(j)b(j+\tau)\right)\left(\Sigma_1^\tau+\Sigma_2^\tau\right)\right)w = 1.\end{aligned} \tag{7}$$

Here C is a non-negative regularization constant, which has to be chosen, e.g., by cross-validation. Since the 1-norm is used in this formulation sparse solutions are achieved. (see e.g. [26], [27] for a discussion of sparsification approaches). With higher C we get sparser solutions for b until at one point the usual CSP approach remains, i.e., $b(1) = 1, b(m) = 0$ for $m > 1$. We call this approach *Common Sparse Spectral Spatial Pattern* (CSSSP) algorithm. Note that a usual personal

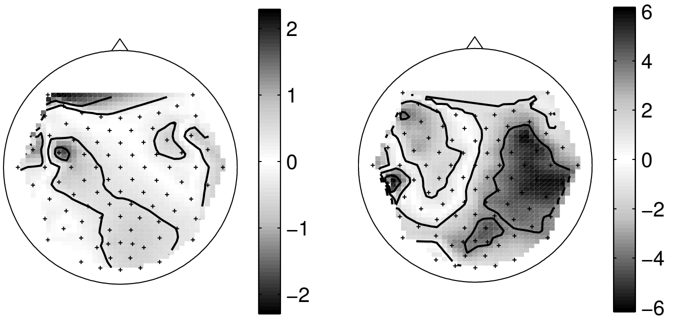

Fig. 4: The figure on the left shows the most discriminative pattern obtained by the classical CSP algorithm applied to broad band filtered data for one experiment and left hand vs. foot imagery. By the use of the proposed combined spatial and spectral optimization the CSSSP algorithm extracts the pattern on the right which shows a much clearer and more plausible topography.

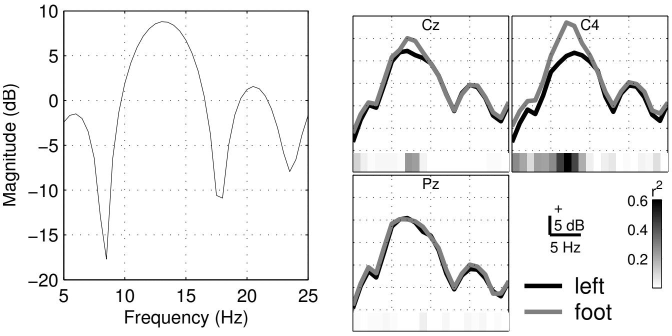

Fig. 5: The plot on the left shows the frequency response for one learned frequency filter for the subject whose spectra is shown in Fig. 2. In the plot on the right the resulting spectra are visualized after applying the frequency filter. By this technique the classification error could be reduced from 17.3 % to 1.4 %.

computer calculates the CSSSP solution on this data in about one minute.

In Fig. 4 the influence of the temporal filter on the choice of the spatial filter is shown.

VI. Feature Extraction, Classification and Validation

A. Feature Extraction

After choosing all channels except the EOG and EMG and a few of the outermost channels of the cap we apply a causal band-pass filter from 7–30 Hz to the data, which encompasses both the μ- and the β-rhythm. For classification we extract the interval 500–3500 ms after the presented visual stimulus. To these trials we apply the original CSP ([14]) algorithm (see Section IV), the extended CSSP ([15]), and the proposed CSSSP algorithm (see Section V). For CSSP we choose the best τ by leave-one-out cross validation on the training set. For CSSSP we present the results for different regularization constants C with fixed $T = 16$. With this T frequency filters with suitable characteristics are available. However, a modification of this parameter was not tested. Furthermore we use 3 patterns per class which leads to 6-dimensional output signal. As a measure of the amplitude in the specified frequency band we calculate the logarithm of the variances on the spatially and temporally filtered output signals as feature vectors.

B. Classification and Validation

The presented preprocessing reduces the dimensionality of the feature vectors to six. Since we have 120 up to 200 samples per class for each data set, there is no additional need for regularization beyond the one of the CSSSP procedure when using linear classifiers according to our experience. When testing non-linear classification methods on these features, we could not observe any statistically significant gain for the given experimental setup when compared to Linear Discriminant Analysis (LDA) (see also [28], [7], [29]). Therefore we choose LDA for classification.

For validation purposes the (chronologically) first half of the data are used to train a classifier which is then applied to the second half of the data to estimate the performance of the classifier. For a first analysis the regularization constant C was chosen fixed to $C = 0.1, 0.5, 1, 5$ to be able to estimate the influence of this constant. The validation procedure included an automatic selection of the hyperparameter C, an optimal C was chosen out of $\{0, 0.01, 0.1, 0.2, 0.5, 1, 2, 5\}$ for each dataset individually by a 2×5-fold cross validation on the trainset only. In this 2×5-fold cross validation the training data is split randomly into 5 disjoint subsets of nearly equal size. Now a classifier is trained on 4 subsets and is applied to the excluded subset. This is repeated for all 5 subsets for 2 different splittings such that one gets 10 errors. The means of these errors for each parameter C were used to select the optimal C. For CSSP the optimal parameter τ was also chosen by cross validation on the training set.

VII. Results

Fig. 5 shows one chosen frequency filter for the subject whose spectra are shown in Fig. 2 and the remaining spectrum after using this filter. As expected the filter detects that there is a high discriminability in frequencies at 12 Hz, but only a low discrimination in the frequency band at 8 Hz. Since the lower

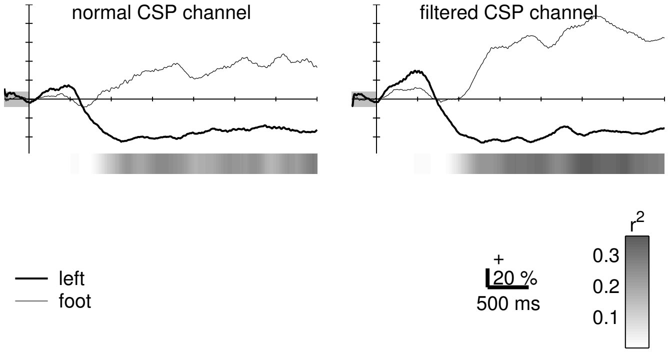

Fig. 6: The plot on the left shows the ERD curve on one projected CSP channel for the subject whose spectra is shown in Fig. 2. After applying the temporal filter calculated by the CSSSP approach the ERD curve on the right is obtained. Below each plot the r^2-values as a measure of discriminability are visualized. They show that the ERD curves on the right can be discriminated better.

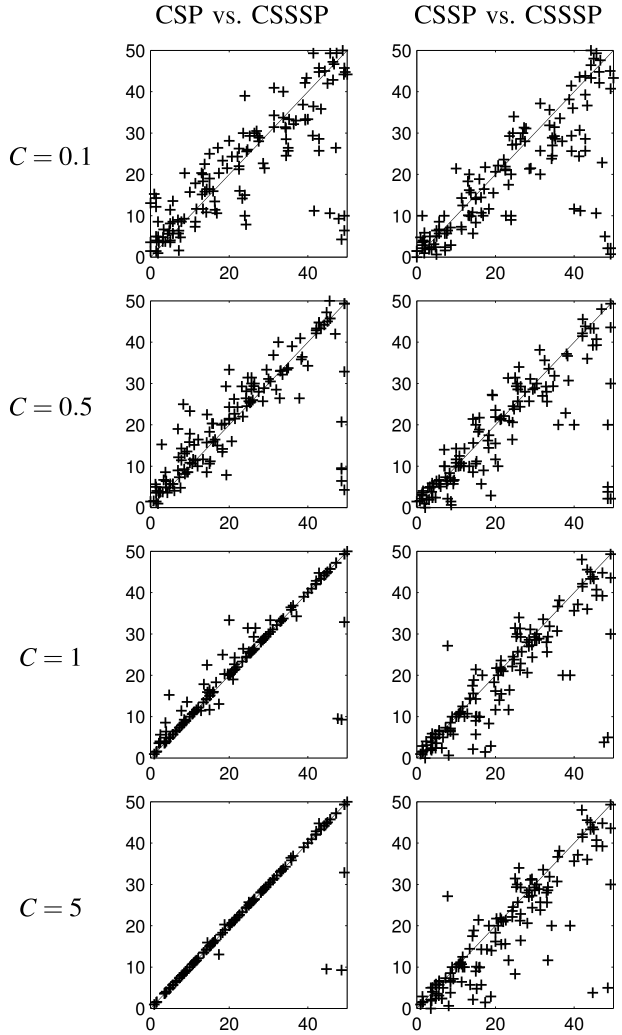

Fig. 7: Each plot shows the validation error of one algorithm against another, in row 1 that is CSP (y-axis) vs. CSSSP (x-axis), in row 2 that is CSSP (y-axis) vs. CSSSP (x-axis). In columns the regularization parameter of CSSSP is varied between 0.1, 0.5, 1 and 5. In each plot a cross above the diagonal marks a dataset where CSSSP outperforms the other algorithm.

frequency peak is very predominant for this subject without having a high discrimination power, a filter is learned which drastically decreases the amplitude in this band, whereas full power at 12 Hz is retained.

Fig. 6 shows the ERD curve for the same subject on one projected CSP channel if one uses data filtered to 7–30 Hz or data additionally temporally filtered with the CSSSP approach. r^2-values (see [30]) are a measure of discriminability between data samples of different classes. In this figure, they reveal the superiority of the suitably filtered data against the normal one.

Applied to all datasets and all pairwise class combinations of the datasets we get the results shown in Fig. 7. First of all, it is obvious that a small choice of the regularization constant C is problematic, since the algorithm then tends to overfit. For high values of C CSSSP tends towards the CSP performance since using frequency filters is penalized too strongly. In between there is a range where CSSSP is significantly better than CSP. Furthermore there are some datasets where the gain when using CSSSP is very significant.

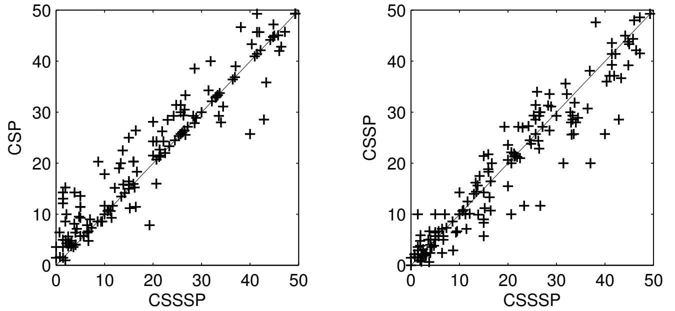

Fig. 8: The figures compare the test error for all datasets of CSSSP on the x-axis vs. CSP (left figure) and CSSP (right figure) on the y-axis. Note that all model parameters (τ for CSSP and C for CSSSP) are chosen on the training set using a cross validation procedure. For crosses above the diagonal CSSSP outperforms the other algorithms.

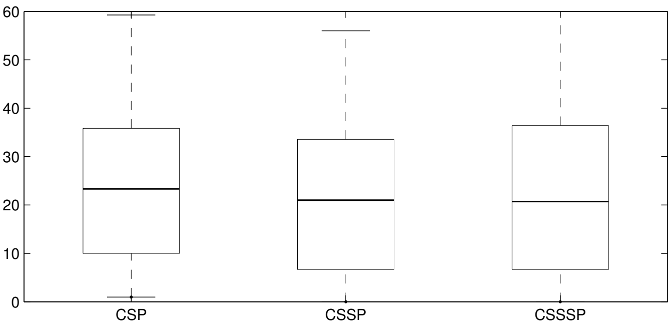

Fig. 9: On the left the boxplots for the results of Fig. 8 for all three algorithms are shown. Here the median-value, the minimum and maximum values and the 25% and 75%-percentiles are marked.

Compared to CSSP the situation is similar, namely CSSSP outperforms the CSSP in many cases and on average, but there are also a few cases, where CSSP is better.

An open issue is the choice of the parameter C. If we choose it constant at 1 for all datasets then Fig. 7 shows that CSSSP will typically outperform CSP (see [1]). Compared to CSSP both cases appear, namely that CSSP is better than CSSSP and vice versa.

A more refined way is to choose C individually for each dataset. One way to accomplish this choice is to perform cross-validations for a set of possible values of C and to select the C with minimum cross-validation error. This was done again for all datasets. The results are shown in Fig. 8. One can observe that there are many datasets where CSSSP outperforms CSP and CSSP, but there are also a few where one of the other algorithms is better. In these cases CSSSP overfits due to the wrong choice of the parameter C. However, according to a Wilcoxon Rank test CSSSP significantly exceeds the other algorithms ($p < 0.005$). In Fig. 9 these results are also shown as box plots with median, minimum and maximum value and 25 %- and 75 %-percentile. Again the superiority of CSSSP against CSSP and CSP is clearly observable: The median classification error rate for CSSSP is 20.7%, for CSSP 21.0% and for CSP 23.3%, i.e., the median classification error rate for CSSSP is 11% lower than for CSP and competative to CSSP.

Note that one could in principle also use the r^2-values shown in Fig. 2 to determine a filter. Based on some heuristic on these values one could choose a global frequency band and apply the corresponding bandpass filter to the data before calculating CSP. However, CSSSP outperforms this algorithm by 9% in median classification error rate, too.

VIII. Concluding discussion

In past BCI research the CSP algorithm has proven to be very sucessful in determining spatial filters which extract discriminative brain rhythms. However the performance can suffer when non-discriminative brain rhythms with an overlapping frequency range interfere. The presented CSSSP algorithm successfully overcomes this problem by optimizing simultaneously the spatial and spectral filters. The trade-off between flexibility of the estimated frequency filter and the danger of overfitting needs to be controlled and is accounted for by CSSSP using a regularizing sparsity constraint. The successfulness of the proposed algorithm when comparing it to the original CSP and to the CSSP algorithm was demonstrated on a corpus of 60 EEG data sets recorded from 22 different subjects. Apart from the excellent classification performance seen when applying CSSSP, an advantage is that an interpretable spatial and temporal filter is learned from data (see Fig. 5). It allows to clearly reveal discriminating parts in the spectrum and thus to contribute to a better understanding of the mechanisms a subject uses for, say, an imagination task. When developing a new paradigm, CSSSP can thus be useful to optimize paradigm design and subject instructions. Finally, we would like to remark that CSSSP is – although very well suited to single trial EEG analysis – a universal signal processing algorithm not limited to the analysis of brain signals: It is a general purpose method that can be readily applied whenever it is necessary to construct spatial and temporal filters for multivariate time-series analysis.

Acknowledgments

We thank S. Lemm for helpful discussions.

References

[1] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Optimizing spatio-temporal filters for improving brain-computer interfacing," in *Advances in Neural Inf. Proc. Systems (NIPS 05)*, vol. 18, 2006, accepted.

[2] 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.

[3] E. A. Curran and M. J. Stokes, "Learning to control brain activity: A review of the production and control of EEG components for driving brain-computer interface (BCI) systems," *Brain Cogn.*, vol. 51, pp. 326–336, 2003.

[4] A. Kübler, B. Kotchoubey, J. Kaiser, J. Wolpaw, and N. Birbaumer, "Brain-computer communication: Unlocking the locked in," *Psychol. Bull.*, vol. 127, no. 3, pp. 358–375, 2001.

[5] 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.

[6] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, R. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, "Current trends in Graz brain-computer interface (BCI)," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 216–219, June 2000.

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

[8] L. Trejo, K. Wheeler, C. Jorgensen, R. Rosipal, S. Clanton, B. Matthews, A. Hibbs, R. Matthews, and M. Krupka, "Multimodal neuroelectric interface development," *IEEE Trans. Neural Sys. Rehab. Eng.*, no. 11, pp. 199–204, Jun 2003.

[9] 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.

[10] W. D. Penny, S. J. Roberts, E. A. Curran, and M. J. Stokes, "EEG-based communication: A pattern recognition approach," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 214–215, June 2000.

[11] J. D. R. Millán and J. Mouriño, "Asynchronous bci and local neural classifiers: An overview of the adaptive brain interface project," *IEEE Trans. Neural Sys. Rehab. Eng.*, vol. 11, no. 2, pp. 159–161, 2003.

[12] J. D. R. Millan, "Brain-computer interfaces," in *Handbook of Brain Theory and Neural Networks*, 2nd ed. MIT Press, 2002.

[13] G. E. Birch and S. G. Mason, "Brain-computer interface research at the Neil Squire Foundation," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 193–195, June 2000.

[14] 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.

[15] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improved classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, 2005.

[16] H. Jasper and W. Penfield, "Electrocorticograms in man: Effects of voluntary movement upon the electrical activity of the precentral gyrus," *Arch. Psychiat. Nervenkr.*, vol. 183, pp. 163–174, 1949.

[17] H. Jasper and H. Andrews, "Normal differentiation of occipital and precentral regions in man," *Arch. Neurol. Psychiat. (Chicago)*, vol. 39, pp. 96–115, 1938.

[18] H. Berger, "Über das Elektroenkephalogramm des Menschen," *Arch. Psychiat. Nervenkr.*, vol. 99, no. 6, pp. 555–574, 1933.

[19] F. H. da Silva, T. H. van Lierop, C. F. Schrijer, and W. S. van Leeuwen, "Organization of thalamic and cortical alpha rhythm: Spectra and coherences," *Electroencephalogr. Clin. Neurophysiol.*, vol. 35, pp. 627–640, 1973.

[20] G. Pfurtscheller and F. H. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov 1999.

[21] 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.

[22] ——, "Increase information transfer rates in BCI by CSP extension to multi-class," in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16, pp. 733–740.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. San Diego: Academic Press, 1990.

[24] Z. J. Koles and A. C. K. Soong, "EEG source localization: implementing the spatio-temporal decomposition approach," *Electroencephalogr. Clin. Neurophysiol.*, vol. 107, pp. 343–352, 1998.

[25] 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, June 2004.

[26] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[27] 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.

[28] 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, pp. 165–169, 2003.

[29] 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.

[30] B. Winer, Ed., *Statistical Principles in Experimental Design*, 2nd ed. New York: McGraw-Hill, 1962.

[31] 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.

Guido Dornhege was born in Werne, Germany, in 1976. He received the Diploma degree in mathematics 2002 from University of Münster, Germany. He conducted studies of Maass cuspforms. Since 2002 he is member of the intelligent data analysis (IDA) group at Fraunhofer-FIRST in Berlin working in the Berlin Brain-Computer Interface (BBCI) project. He received the PhD degree in computer science 2006 from University of Potsdam, Germany. His scientific interests are in the field of analysis of biomedical data by machine learning techniques.

Benjamin Blankertz received the Diploma degree in mathematics 1994 and the Ph.D. in mathematical logic in 1997, both from University of Münster, Germany. He conducted studies in computational models for perception of music and computer-aided music analysis. Since 2000 he is with the intelligent data analysis (IDA) group at Fraunhofer FIRST in Berlin working in the Berlin Brain-Computer Interface (BBCI) project. His scientific interests are in the fields of machine learning, analysis of biomedical data, and psychoacoustics.

Matthias Krauledat

Florian Losch received the Dr. med. degree with the analysis of signal behaviour, propagation and topographic determination in the somatosensory system measured by EEG and MEG. After working about EEG-EMG coherence in patients with writer’s cramp he joined again the Neurophysics Group at the Department of Neurology of the Campus Benjamin Franklin, Charite - University Medicine Berlin in 2003 to work in the Berlin Brain Computer Interface (BBCI) project. Motor-associated signal behaviour under varying frequencies and speed, role of inhibition and influence of sensory feedback in signal development are the topics of current interest.

Gabriel Curio received the Dr. med. degree with a thesis on attentional influences on smooth pursuit eye movements and holds Board specializations in Neurology and Psychiatry. Since 1991 he is leading the Neurophysics Group at the Department of Neurology of the Campus Benjamin Franklin, Charité – University Medicine Berlin. His main interest is to integrate the neurophysics of noninvasive electromagnetic brain monitoring with both basic and clinical neuroscience concepts. Recent research interests of Dr. Curio include spike-like acitivities in somatosensory evoked brain responses, neuromagnetic detection of injury currents, magnetoneurography, the comparison of cortical processing of phonems versus musical chords, speech-hearing interactions, single-trial EEG/MEG analysis and and brain-computer interfacing. Since 1998 Dr. Curio serves as member of the Technical Commission of the German Society for Clinical Neurophysiology, and since 2003 as Section Editor for the IFCN journal ’Clinical Neurophysiology’.

Klaus-Robert Müller received the Diplom degree in mathematical physics 1989 and the Ph.D. in theoretical computer science in 1992, both from University of Karlsruhe, Germany. From 1992 to 1994 he worked as a Postdoctoral fellow at GMD FIRST, in Berlin where he started to built up the intelligent data analysis (IDA) group. From 1994 to 1995 he was a European Community STP Research Fellow at University of Tokyo in Prof. Amari’s Lab. From 1995 on he is department head of the IDA group at GMD FIRST (since 2001 Fraunhofer FIRST) in Berlin and since 1999 he holds a joint associate Professor position of GMD and University of Potsdam. In 2003 he became full professor at University of Potsdam. He has been lecturing at Humboldt University, Technical University Berlin and University of Potsdam. In 1999 he received the annual national prize for pattern recognition (Olympus Prize) awarded by the German pattern recognition society DAGM. He serves in the editorial board of Computational Statistics, IEEE Transactions on Biomedical Engineering and in program and organization committees of various international conferences. His research areas include statistical physics and statistical learning theory for neural networks, support vector machines and ensemble learning techniques. He contributed to the field of signal processing working on time-series analysis, statistical denoising methods and blind source separation. His present application interests are expanded to the analysis of biomedical data, most recently to brain computer interfacing and genomic data analysis.