Spatio-Spectral Filters for Improving the Classification of Single Trial EEG

Steven Lemm, Benjamin Blankertz, Gabriel Curio and Klaus-Robert Müller

***Abstract*— Data recorded in EEG based Brain-Computer Interface experiments is generally very noisy, non-stationary and contaminated with artifacts, that can deteriorate discrimination/classification methods. In this work we extend the Common Spatial Pattern (CSP) algorithm with the aim to alleviate these adverse effects. In particular we suggest an extension of CSP to the state space, which utilizes the method of time delay embedding. As we will show, this allows for individually tuned frequency filters at each electrode position and thus yields an improved and more robust machine learning procedure. The advantages of the proposed method over the original CSP method are verified in terms of an improved information transfer rate (bits per trial) on a set of EEG-recordings from experiments of imagined limb movements.**

***Index Terms*— feature extraction, CSP, classification, BCI**

I. Introduction

THe development of a Brain-Computer Interface (BCI) aims to provide a communication channel from a human to a computer that directly translates brain activity into sequences of control commands. Such a device may give disabled people direct control over a neuro-prosthesis or over computer applications as tools for communicating solely by their intentions that are reflected in their brain signals (e.g. [1]–[9]). We record brain activity by means of multi-electrode electroencephalogram (EEG) which is non-invasive as opposed to invasive work by e.g. [10]–[13]. An ideal BCI should only need short adaptation and preparation times and should yield high information transfer rates. In practice the user is behaving according to a well-defined paradigm (such as movement imagination) that allows for an effective discrimination between different brain states (see e.g. [3], [14]). From the signal processing perspective this requires the definition of appropriate features that can be effectively translated into a control signal, either by simple threshold criteria (cf. [3]), or by means of machine learning, where given some training examples for each brain state, the task is to infer a possibly nonlinear discriminating function to distinguish between different states (e.g. [3], [6]–[9], [15], [16]). Non-invasive data acquisition, makes automated feature extraction challenging, since the signals of interest are 'hidden' in a highly 'noisy' environment as EEG signals consist of a superposition of a large number of simultaneously active brain sources that are typically distorted by artifacts and even subject to non-stationarity. However, outliers and artifacts can strongly distort the classifier performance [15], yielding bad generalization, i.e. the performance on previously unseen data, can become arbitrarily poor. So it is important to strive for robust machine learning and signal processing methods that are as invariant as possible against such distortions (e.g. [16], [17]). This paper contributes an extension to CSP based BCI approaches [18], [19] in order to improve both: the accuracy and the generalization ability of a brain state classifier.

The paper is organized as follow: Section II introduces the underlying neurophysiological principles and elaborates on the mathematical background of CSP. Based on the latter the subsequent section introduces the Common Spatio-Spectral Pattern algorithm, as an extension of CSP to the state space and provides insights into the implications of the extended model. The performance of the two methods is then compared on a broad set of experiments in section IV and a concluding discussion follows.

This work was supported in part by the DFG SFB 618/B4, the Bundesministerium für Forschung (BMBF) Grant FKZ 01IBB02A,B, and the PASCAL Network of Excellence (EU #506778).

S. Lemm is with the Department of Intelligent Data Analysis, FIRST Fraunhofer Institute, 12489 Berlin, Germany and also with the Neurophysics Group, Department of Neurology, Campus Benjamin Franklin, Charité, University Medicine 12200 Berlin, Germany (e-mail: steven.lemm@first.fhg.de)

B. Blankertz is with the Department of Intelligent Data Analysis, FIRST Fraunhofer Institute, 12489 Berlin, Germany

G. Curio is with the Neurophysics Group, Department of Neurology, Campus Benjamin Franklin, Charité, University Medicine Berlin, 12200 Germany

K.-R. Müller is with the Department of Intelligent Data Analysis, FIRST Fraunhofer Institute, 12489 Berlin, Germany and also with the Computer Science Department, University of Potsdam, 14482 Potsdam, Germany

II. Neurophysiological and Mathematical Background

A. Neurophysiology

According to the concept known as homunculus, for each part of the human body there exists a respective region in the motor and somatosensory area of the neocortex. The 'mapping' from the body to the respective brain areas preserves topography, i.e., neighboring parts of the body are represented in neighboring parts of the cortex. While the region of the feet is at the center of the vertex, the left hand is represented lateralized on the right hemisphere and the right hand on the left hemisphere. One possible feature of brain activity, that can be exploited for brain-computer interfaces relies on the neurophysiological observation, that large populations of neurons in the respective cortex are firing in rhythmical synchrony when a subject is not engaged with one of his limbs (movements, tactile senses, or just mental introspection). These are so-called idle rhythms that are attenuated when engagement with the respective limb takes place and that can be measured at the scalp in the EEG as a brain rhythm around 10 Hz (μ-) or 20 Hz (β-rhythm). As the attenuation effect is due to loss of synchrony in the neural populations, it is termed event-related desynchronization (ERD), see [20]. In opposite, the dual effect of an enhanced rhythmic activity is called event-related synchronization (ERS). Such modulations of the μ- and β-rhythm have been reported for different physiological manipulations, e.g., by motor activity, both actual and imagined [21]–[23], as well as by somatosensory stimulation [24]. In order to setup a BCI we will utilize these physiological phenomena, in particular caused by imaginary movements or sensations at different limbs. The discrimination between different limbs, e.g. left hand vs. right hand vs. foot will exploit the dissimilarities in the spatio-spectral topography of the attenuation of the μ and/or β rhythm.

The strength of the sensorimotor idle rhythms as measured by scalp EEG is known to vary strongly between subjects. This introduces a high inter-subject variability on the accuracy with which an ERD-based BCI system works. That is reflected in a high varying classification accuracy for different subjects. Hence [19], [25] suggested to combine the oscillation feature of ERD with another feature reflecting imagined or intended movements, the movement related potentials (MRP), denoting a negative DC shift of the EEG signals in the respective cortical regions. In this paper we are focusing on

the improvement of the classification based only on the oscillatory feature (ERD/S), nevertheless the suggested algorithm can be straight forwardly integrated into the combination framework.

B. Common Spatial Pattern

The common spatial pattern (CSP) algorithm [26] is highly successful in calculating spatial filters for detecting ERD/ERS effects [27] and for ERD-based BCIs, see [18] and has been extended to multi-class problems in [19]. Given two distributions in a high-dimensional space, the CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and that at the same time minimize variance for the other class. After having bandpass filtered the EEG signals in the frequency domain of interest, high or low signal variance reflect a strong respective a weak (attenuated) rhythmic activity. Let us take the example of discriminating left hand vs. right hand imaginary movement. According to Section II-A, if the EEG is first preprocessed in order to focus on the μ and β band, i.e. bandpass filtered in the frequency range 7–30 Hz, then a signal projected by a spatial filter focusing on the left hand area is characterized by a strong motor rhythm during the imagination of right hand movements (left hand is in idle state), and by an attenuated motor rhythm if movement of the left hand is imagined. This can be seen as a simplified exemplary solution of the optimization criterion of the CSP algorithm: maximizing variance for the class of right hand trials and at the same time minimizing variance for left hand trials. Furthermore the CSP algorithms calculates the dual filter that will focus on the area of the right hand (and it will even calculate several filters for both optimizations by considering orthogonal subspaces).

To be more precise, let $X^k = \left(X^k_{c,t}\right),\ c = 1,\ldots,C,\ t = t_0,\ldots,T$ denote the (potentially bandpass filtered) EEG recording of the k-th trial, where C is the number of electrodes. Correspondingly $Y^k \in \{1;2\}$ represents the class-label of the k-th trial. Using this notation then the two class-covariance matrices are given as,

$$\Sigma_1 = \langle X^k {X^k}^\top \rangle_{\{k:Y^k=1\}} \quad \text{and} \quad \Sigma_2 = \langle X^k {X^k}^\top \rangle_{\{k:Y^k=2\}} \tag{1}$$

$$W\Sigma_1 W^\top = D \qquad \text{and} \qquad W\Sigma_2 W^\top = I - D. \tag{2}$$

This can be accomplished in the following way: First *whiten* the matrix $\Sigma_1 + \Sigma_2$, i.e., determine a matrix P such that

$$P(\Sigma_1 + \Sigma_2)P^\top = I. \tag{3}$$

This decomposition can always be found due to positive definiteness of $\Sigma_1 + \Sigma_2$. Second define $S_1 = P\Sigma_1 P^\top$ and $S_2 = P\Sigma_2 P^\top$ respectively and calculate an orthogonal matrix R and a diagonal matrix D by spectral theory such that

$${S_1}^\top = RDR^\top. \tag{4}$$

From $S_1 + S_2 = I$ it follows that ${S_2}^\top = R(I-D)R^\top$. Note that the projection given by the p-th row of matrix R has a relative variance of d_p (p-th element of D) for trials of class 1 and relative variance $1 - d_p$ for trials of class 2. If d_p is close to 1 the filter given by the p-th row of R maximizes variance for class 1, and since $1 - d_p$ is close to 0, minimizes variance for class 2. The final decomposition, that satisfies Eq.(2) can be obtained from,

$$W := R^\top P. \tag{5}$$

Using this decomposition matrix W the EEG recordings X^k are projected onto

$$Z^k = WX^k. \tag{6}$$

The interpretation of W is two-fold, the rows of W are the stationary spatial filters, whereas the columns of W^{-1} can be seen as the common spatial patterns or the time-invariant EEG source distribution vectors.

III. Common Spatio-Spectral Pattern (CSSP)

A. Feature Extraction

In this section we will extend the CSP algorithm to the state space. Therefore we first introduce an extension to the state space for Eq.(6), subsequently we discuss its consequences for the optimization problem and give a mathematical interpretation.

The concept of deterministic low-dimensional chaos has proven to be fruitful in the understanding of many complex phenomena despite the fact that very few natural systems have actually been found to be low-dimensional deterministic in the sense of theory. Also a number of attempts have been made to analyze various aspects of EEG time series in the context of nonlinear deterministic dynamical systems.

Determinism in a strict mathematical sense means that there exist an autonomous dynamical system, defined typically by a first order differential equation $\dot{\mathbf{y}} = f(\mathbf{y})$ in a state space $\Gamma \subset R^D$, which is assumed to be observed through a single measurable quantity $s = h(\mathbf{y})$. The system thus possesses D natural variables, but the measurement is usually a nonlinear projection onto a scalar value. In order to recover the deterministic properties of such a system, we have to reconstruct an equivalent of the state space Γ. Therefore the time delay embedding method is one way to do so. From the sequence of scalar observations $s_1, s_2, \ldots, s_N$ overlapping vectors $\mathbf{s}_n = (s_n, s_{n-\tau}, \ldots, s_{n-(m-1)\tau})$ are formed, with τ as the delay time. Then according to Takens Theorem [28] under certain conditions, i.e. for mathematically perfect, noise free observations s_n and m sufficient large, there exist a one-to-one relation between $\mathbf{s}_n$ and the unobserved vectors $\mathbf{y}_n$. Thus the attractor of any non-linear dynamic can be reconstructed in the state space using an appropriate delay coordinate function.

Since it is not our aim to reconstruct the entire dynamics of the EEG-signal, but rather to extract *robust* (invariant) features, we extend Eq. (6) just by one delayed coordinate, i.e.,

$$Z^k = W^{(0)}X^k + W^{(\tau)}\delta^\tau X^k. \tag{7}$$

Where, for notational convenience, δ^τ denotes the delay operator, i.e,

$$\delta^\tau (X_{\cdot,t}) = X_{\cdot,t-\tau}. \tag{8}$$

Once again, the optimization criterion is to find projections $W^{(0)}$ and $W^{(\tau)}$ such that signal variance of different Z_p discriminates two given classes best, i.e. maximizing the variance for one class while minimizing it for the opposite class. In order to use the identical mathematical concepts, introduced in section II-B, we append the delayed vectors $\delta^\tau X^k$ as additional channels to X^k, i.e.

$$\hat{X}^k = \begin{pmatrix} X^k \\ \delta^\tau X^k \end{pmatrix}. \tag{9}$$

Then the optimization criterion can be formulated equivalent to Eq. (2) using the class covariance matrices $\hat{\Sigma}_l, l \in \{1,2\}$ obtained from $\hat{X}^k$. Following the steps of Eq. (3)–(5), this yields a solution to this modified optimization problem, especially a decomposition matrix $\hat{W}$, whose columns divide in two submatrices: $\hat{W}^{(0)}$ that applies to X^k and $\hat{W}^{(\tau)}$ that applies to the delayed channels $\delta^\tau X^k$, i.e., $\hat{W}\hat{X}^k = \left(\hat{W}^{(0)}\,\hat{W}^{(\tau)}\right)\hat{X}^k$.

Based on this, we will now further explore the implications of this decomposition. Especially we will derive an interpretation into a spatial and a spectral filter. Therefore let w denote the p-th row

of the decomposition matrix $\hat{W}$, then the projected signal $\hat{Z}_p^k = w\hat{X}^k$ can be expressed as

$$\begin{aligned}\hat{Z}_p^k &= w^{(0)}X^k + w^{(\tau)}\delta^\tau X^k \\ &= \sum_{c=1}^{C} w_c^{(0)}X_{c,\cdot}^k + w_c^{(\tau)}\delta^\tau X_{c,\cdot}^k \\ &= \sum_{c=1}^{C}\gamma_c\left(\frac{w_c^{(0)}}{\gamma_c}X_{c,\cdot}^k + \frac{w_c^{(\tau)}}{\gamma_c}\delta^\tau X_{c,\cdot}^k\right), \end{aligned} \tag{10}$$

where $(\gamma_c)_{c=1,\ldots,C}$ is a pure spatial filter and $(\frac{w_c^{(0)}}{\gamma_c}, \overbrace{0,\ldots,0}^{\tau-1}, \frac{w_c^{(\tau)}}{\gamma_c})$ defines a Finite Impulse Response (FIR) filter for each electrode c. This decomposition into a spatial and a FIR filter is not unique, but there exists a very intuitive partitioning, that we will use throughout this paper, i.e.

$$\gamma_c := \frac{\sqrt{{w_c^{(0)}}^2 + {w_c^{(\tau)}}^2}}{\widetilde{\text{sign}}\left(w_c^{(0)}\right)}, \tag{11}$$

where

$$\widetilde{\text{sign}}(w) = \begin{cases} -1, & w < 0 \\ +1, & w \geq 0 \end{cases}. \tag{12}$$

The use of the signed norm γ of the coefficients vector as spatial filter allows us to examine the origin of the projected source signals since each column of the inverse of the entire spatial projection matrix $\Gamma = (\gamma)_{p,c}$ corresponds to the coupling strength of one source with the electrodes. Note that γ_c maps the non-zero coefficients of corresponding FIR filter on to one half of the two dimensional unit-sphere. Consequently we can easily parameterize the FIR filters by the angle $\phi_c^{(\tau)}$, defined as

$$\phi_c^{(\tau)} := \text{atan}\left(\frac{w_c^{(0)}}{w_c^{(\tau)}}\right) \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \tag{13}$$

Fig. 1 and 2 illustrate the effect of these FIR filters by means of the resulting filter responses curves for various values of τ and at different angles $\phi^{(\tau)}$. Note that at each electrode the FIR filter is additional to the global bandpass filter, that focuses on the frequency band of interest.

This additional property of the decomposition matrix allows for a fine tuning of the overall frequency filters, e.g. an adaptation to the spectral EEG peaks.

B. Classification

Finally, the features used for the classification are obtained by decomposing the EEG according to Eq.(6) respectively (7). Typically one would retain only a small number $2m$ of projections that contain most of the discriminative information between the two classes, i.e. the signal variance. These projections are given by the rows of $\hat{W}$ that correspond to the m largest and m smallest eigenvalues d_{p_i}. Based on the projected single trials $Z_{p_i,t}^k, i = 1,\ldots,2m$, a classifier is estimated on the log-transformed signal variances, i.e.

$$f_i^k = \log\left(\text{var}\left(Z_{p_i,\cdot}^k\right)\right). \tag{14}$$

Specifically we applied a linear discriminant analysis (LDA) as classification model. Since the introduced delay τ appears as an underlying hyper-parameter in the overall optimization scheme, it has to be subject of a model selection procedure in order to find the optimal τ for the specific classification task. Note that using $\tau = 0$ in the set of feasible hyper-parameters, will incorporate the original CSP algorithm into the model selection procedure.

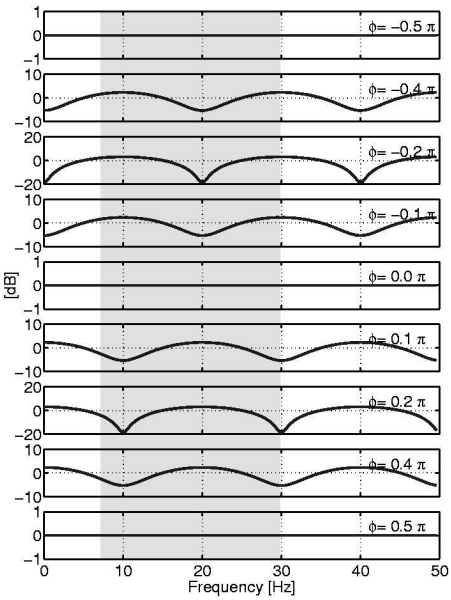

Fig. 1: Magnitude responses of the FIR filters at different values of $\phi^{(\tau)}$ for a single fixed delay τ =50 ms. Increasing $\phi^{(\tau)}$ from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$ keeps the position of the extreme values in the frequency domain fixed, but turns maxima into minima. Since a minimum corresponds to a suppression of the spectral information at this frequency, the FIR filter for $\phi^{(\tau)} = -\frac{2}{5}\pi$ focuses mainly on $\{10, 30, 50\}$ Hz. In opposite the filter given by $\phi^{(\tau)} = \frac{1}{5}\pi$ can be associated with the contrary effect, i.e. it cancels these frequencies. The shaded region denotes the frequency range (7–30 Hz) of interest, that defines the bandpass filter used for preprocessing the data in the experimental session.

C. Online Applicability

A major concern in online applications is to implement an algorithm as efficient as possible. In case of a CSP based classification procedure the involved operations such as the bandpass filtering and the spatial projections define the bottleneck for the processing speed. Especially the filtering of each EEG channel (64-128) is quite time consuming and dramatically effects the overall processing speed. But fortunately the bandpass filtering, if realized by convolution with h, an Infinite Impulse Response (IIR) filter ($h \star X$), and the spatial filtering (WX) are strictly linear, and it follows immediately that these operations can be executed in an arbitrary order, i.e.

$$W(h \star X) = h \star (WX). \tag{15}$$

This implies, that for an online application one does not need to bandpass filter all the channels before projecting onto the few spatial features. Instead we can first apply the spatial projection and filter only the few resulting signals with the desired bandpass, what makes the resulting algorithm applicable online.

In case of CSSP, due to the embedding operation δ^τ this does not hold in general. But we can easily work around this and are allowed to exchange the operations in the following manner:

$$\hat{W}\begin{pmatrix} h\star X \\ \delta^\tau(h\star X)\end{pmatrix} = \begin{pmatrix} I_C \\ I_C\end{pmatrix}^\top \begin{pmatrix} h\star\hat{W}^{(0)}X \\ \delta^\tau(h\star(\hat{W}^{(\tau)}X))\end{pmatrix}. \tag{16}$$

Where I_C denotes the C-dimensional identity-matrix. Again one can arbitrarily exchange the order of first decomposing the EEG using $W^{(0)}$ and $W^{(\tau)}$ and then bandpass filtering the projected channels, as long as the delay operator and the final summation is applied afterwards. From this it can be directly seen, that the computational demands are just doubled compared to the original CSP. Hence the proposed extended model is applicable online as well.

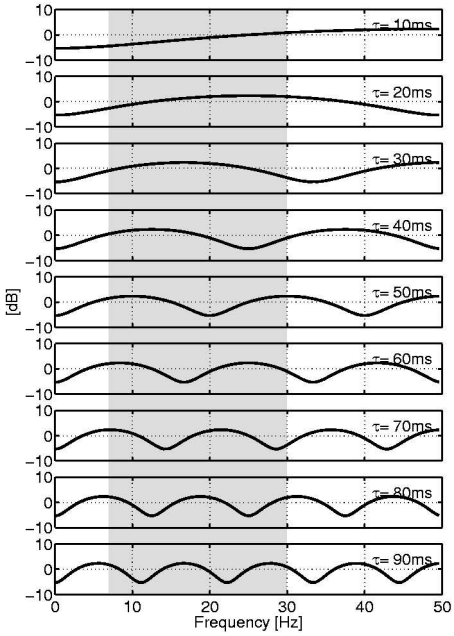

Fig. 2: Magnitude responses of the FIR filters at different values of τ for a single fixed angle $\phi = -\frac{2}{5}\pi$. Increasing τ changes the position and increases the number of the extreme values in the frequency domain. Since a minimum corresponds to a suppression of the spectral information at this frequency, the FIR filters for different τ focuses on different sub-bands of the frequency spectrum. The shaded region denotes the frequency range (7–30 Hz) of the additional bandpass filter used for preprocessing the data in the experimental session.

IV. Application

A. Experimental Design

In this section we apply the CSP and the proposed CSSP algorithm to data from 22 EEG experiments of imaginary limb movements performed by 8 different subjects and compare the resulting classification performances. The investigated mental tasks were imagined movements of the left hand (*l*), the right hand (*r*), and the right foot (*f*). Two experiments were carried out with only 2 classes *l* and *r*. In this study we investigate all resulting two-class classification problems, i.e. all possible combination of two classes (*l*-*r*,*l*-*f* and *r*-*f*), yielding 62 different classification tasks.

All experiments start with training sessions in which the subjects performed mental motor imagery tasks in response to a visual cue. In such a way examples of brain activity can be obtained during the different mental tasks. In the original experiment, these recorded single trials were then used to train a classifier which was in a second sessions applied online to produce a feedback signal for (unlabeled) continuous brain activity (results will be reported elsewhere).

In this off-line study we will only take data from the first (training session) into account to evaluate the performance of the algorithms under study. This reflects, that if any feedback is provided to the subject, he/she will adapt to the feedback (output of the classifier) itself. Hence the data obtained from a feedback session are biased towards the specific classifier used to produce the particular feedback and for that reason we decided to exclude the data of the feedback session for the evaluation process.

During the experiment the subjects were sitting in a comfortable chair with arms lying relaxed on the armrests. In the training period every 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. The brain activity was recorded from the scalp at a sampling rate of 100 Hz with multi-channel EEG amplifiers using 32, 64 resp. 128 channels. Additional to the EEG channels, we recorded the electromyogram (EMG) from both forearms and the right leg as well as horizontal and vertical electrooculogram (EOG) from the eyes. The EMG and EOG channels were exclusively used to ensure 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 should be constrained to operate on the CNS (central nervous system) activity only. For each involved mental task we obtained between 120 and 200 single trials of recorded brain activity.

B. Training

After choosing all channels except the EOG and EMG and a few outermost channels of the cap that are known to have non-stationary signal quality due to changing conductive properties, we applied a causal band-pass filter from 7–30 Hz to the data, which encompasses both the μ- and the β-rhythm. The single trials were extracted from the temporal frame 750–3500 ms after the presentation of the visual stimulus, since during this period discriminative brain patterns are present in most subjects. On these preprocessed single trials we perform the feature extraction by the CSP and the proposed CSSP method separately. For each method we project the data to the three most informative directions of each class, yielding a 6-dimensional subspace. For these six dimensions we calculate the logarithms of the variances as feature vectors. Finally we apply a linear discriminant analysis (LDA) to the feature vectors of all trials, to find a separating hyperplane. Note that in contrast to [6], [14] we omitted the regularization (cf. [29]), since the dimensionality of the features is rather small compared the number of training examples.

In order to compare the results of the two methods (CSP vs. CSSP with selected τ) we split the data set in two. On the (chronological) first half we perform the training of the classifier, i.e. the feature extraction, model selection and the LDA. The performance of the estimated models were then evaluated on the second half of the data, to which both algorithms had have no access before. In the following we will refer to these halves of the data set as “training data” and “test data”.

To select the best CSSP model for each binary classification problem, i.e. find the optimal τ, we estimate the performance of the algorithms by means of a leave one (trial) out (LOO) cross-validation

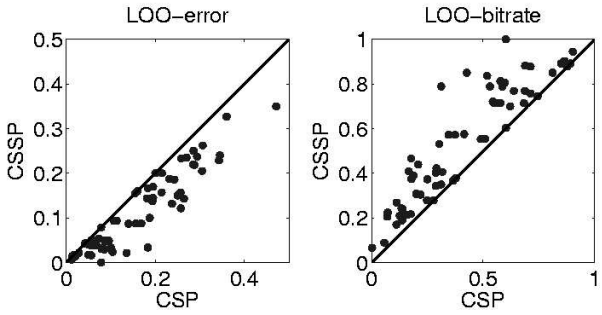

Fig. 3: The left panel compares the LOO classification error on the training data of the CSP based model and the selected CSSP based model for 62 binary classification problems. In either case the error decreases. For 6 out of 62 datasets the model selection suggest to keep the CSP-based model, indicated by dots on the diagonal line, in any other case the CSSP-based classification improves the LOO-training error. The right panel shows the identical information but instead the error the corresponding information transfer rate measured in bits per trial is given.

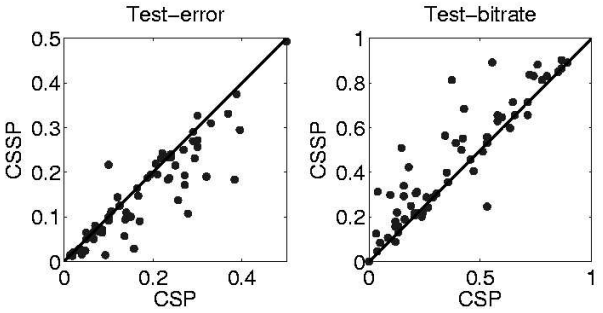

Fig. 4: The left panel compares the classification error of the CSP based model and the selected CSSP based model on the test data taken from the second half of the experiment. Except for a very few cases the CSSP based classification model improves the test error. Again the right panel provides the same information in terms of the corresponding information transfer rate per trial.

procedure on the corresponding training data. Especially we run a model selection procedure over $\tau = 0,..15$. Since the CSP resp. the CSSP algorithm make explicit use of the label information, these techniques have to be repeatedly applied to each LOO training set within the cross-validation procedure. Otherwise the cross-validation error could underestimate the generalization error. Fig. 3 compares the LOO error respective the information transfer rate of the CSP model and the best CSSP on the training data for each experiment. Note that in this particular case CSSP always performs the same or better than the CSP on the training data, since the CSP model is part of the model selection procedure ($\tau = 0$).

C. Results

After the model selection the best CSSP and the CSP based models for each classification problem were then finally trained on the entire training data and afterwards applied to the corresponding test data, obtained from the second unseen half of the training session of each experiment. A comparison of the resulting test errors/bitrate for all datasets is summarized in Fig. 4. The results on that database

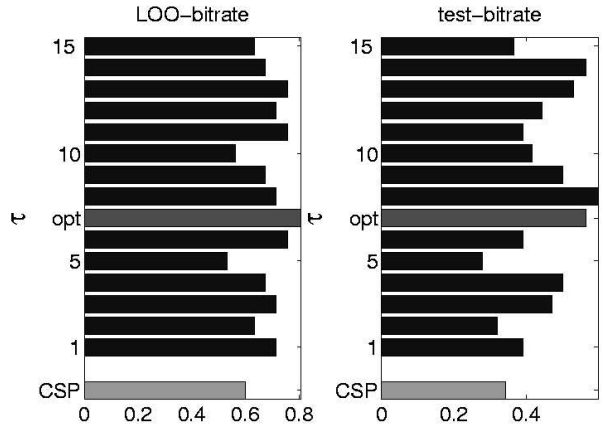

Fig. 5: The left panel gives the resulting LOO information transfer rate in bits per trial on the training set for one exemplarily chosen dataset at all hyper-parameter $\tau = 0,\ldots,15$ (at a sampling rate of 100 Hz this corresponds to $0,\ldots 150$ ms). The right panel shows the corresponding information transfer rate on the test data. For that particular classification problem almost any delay parameter would improve the standard CSP-based classification, shown in the lowest row of each panel. The best model that has been selected by the model selection procedure based on the informations provided in the left panel corresponds to $\tau = 7$.

strongly suggest that the proposed algorithm outperforms the CSP-based approach, in terms of an improved classification accuracy and an increased generalization ability.

For further illustrations of the properties of the proposed method, we will now pick one specific dataset, especially we will focus on a classification task of imaginary foot and right hand movement that serve this purpose best. For this selected dataset we will relate the spatial filters found by the CSSP method to those of the CSP-method and will discuss the impact of the additional spectral filters (cf Eq. (13)).

To show the general performance of the CSSP-based algorithm, we trained CSSP models for all values of $\tau \in \{0,\ldots,15\}$ on the training data and applied them to the test data. Fig.5 shows the information transfer rate on the training and on the test data for all models ($\tau = 0,\ldots,15$) for the selected dataset, where $\tau = 0$ corresponds to the original CSP algorithm. In that particular case almost any additional delay improves the classification both on the training data as well as on the unseen test data. According to the model selection procedure, described in section IV-B, the model with the highest LOO bitrate (lowest LOO error) on the training data ($\tau = 7$) would be chosen for the final application. The spatial and spectral filters of the selected

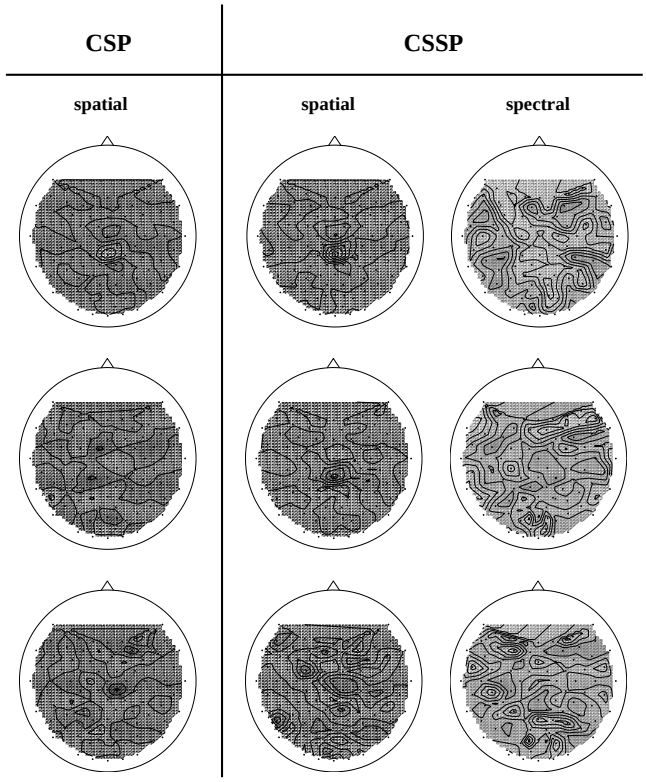

Fig. 6: The scalp maps show the three spatial and spectral filters for class *feet* in descending order of the eigenvalues for both the CSP and the CSSP method. The filters were calculated for a classification of *foot* and *right hand* movements. The spectral filters are gray-scale-coded according Eq.(13) in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Note that the first spatial filters are almost identical for CSP and CSSP, but already those for the second largest eigenvalue diverges. While the spatial filters found by CSP exhibit no clear structure, the second spatial CSSP filter resembles the first one. The main difference in the projection occurs only in the spectral filter, where these filters have opposite sign in the central region, indicating that at the same location different spectral information is exploited for classification.

model for each class are visualized in Fig. 6 and 7. These figures also provide the filters found by CSP method. The first two spectral

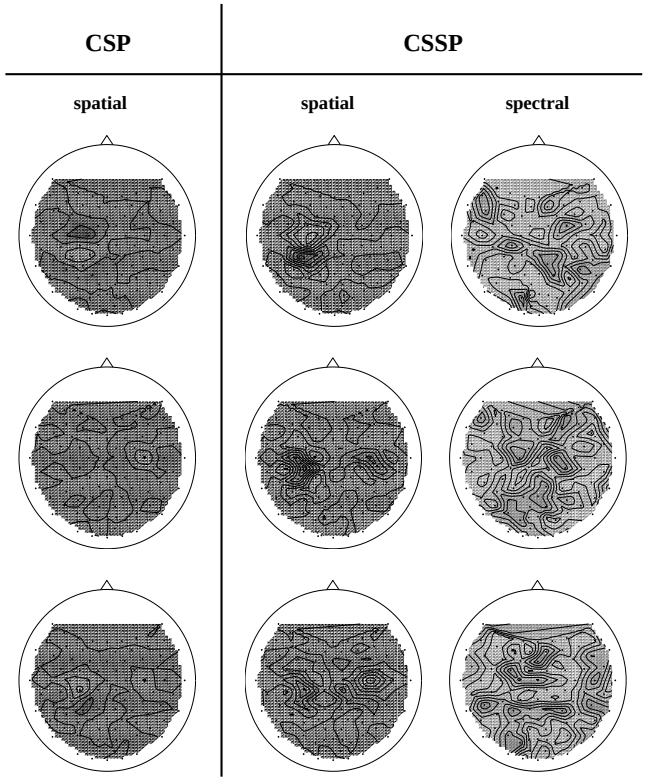

Fig. 7: The scalp maps correspond to the spatial and spectral filters of the three leading eigenvalues for class *right hand* of the CSP and the CSSP method in descending order. The filters were calculated for a classification of *foot* and *right hand* movements. Again the first spatial filters are almost identical for CSP and CSSP (except for the sign). For the second largest eigenvalue the CSSP filters work still at the same location, whereas the CSP filter exhibit no clear focal point at the area of the left motor cortex.

filters for class *foot* supply insight how additional spectral information is exploited. The corresponding spatial filters found by the CSSP method are almost identical and focusing on the central region, while the spectral filters have opposite signs in this area. This indicates, that spectral information of disjoint frequency bands is obtained from the same location. If we look at the FIR filter that corresponds to $\tau = 7$ in Fig. 2, it turns out that its maxima and minima respectively are roughly at 14,21,28 Hz in the bandpass selected frequency range. Remember that for different signs of $\Phi^{(\tau)}$ the maxima and the minima are exchanged. Aggregating these facts, the first spectral filter focuses on β band, whereas the second spectral filter has its focus close to the upper α band (11–13 Hz). So instead of having a spatial projection onto a broad band (7–30 Hz) signal as a solution given by the CSP, CSSP can split the information furthermore by projecting onto two signals of the same local origin, but stemming from different subbands, such that each projection fulfills the optimization criterion of maximizing the variance for one class, while having minimal variance for the other class.

In such a way the CSSP algorithm is not only able to automatically adapt to the spectral EEG characteristics of a subject, but also to treat different spectral information, originating from closely adjoint (or identical) focal areas independently. Summarizing, this yields an improved spatio-spectral resolution of the discriminative signals and hence can improve the robustness and accuracy of the final classification.

V. Conclusion

The paper utilized the method of delay embedding in order to extend the CSP-algorithm to the state space. The advantages of the proposed method were proved by its application to the classification of imagined limb movements on a broad set of experiments. We found that the CSSP algorithm introduced here outperforms the current state-of-the-art CSP algorithm in terms of classification accuracy as well as generalization ability.

It is worth to mention, that in principle it is possible to further extend the suggested model by incorporating more than one temporal delay. But this will come at the expense of a quadratically increasing number of parameters for the estimation of the covariance matrices, while the number of single trials for training remain the same. Hence and in consistence with our observations, this approach will tend to over fit the training data, i.e. the simultaneous diagonalization of the class covariance matrices finds directions that explain the training data best, but might have poor generalization ability. But this directly raises an interesting question for further studies, i.e. how to appropriately regularize the existing diagonalization methods to achieve better generalization.

VI. Acknowledgements

The authors would like to thank Guido Dornhege and Matthias Krauledat for valuable discussions.

References

[1] J. del R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Noninvasive brain-actuated control of a mobile robot by human EEG," *IEEE Trans Biomed Eng*, vol. 51, no. 6, pp. 1026–33, 2004.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] S. P. Levine, J. E. Huggins, S. L. BeMent, R. K. Kushwaha, L. A. Schuh, M. M. Rohde, E. A. Passaro, D. A. Ross, K. V. Elsievich, and B. J. Smith, "A direct brain interface based on event-related potentials," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 2, pp. 180–185, 2000.

[12] D. Taylor, S. Tillery, and A. Schwartz, "Direct cortical control of 3d neuroprosthetic devices," *Science*, vol. 5574, no. 296, pp. 1829–32, Jun 2002.

[13] E. Leuthardt, G. Schalk, J. Wolpaw, J. Ojemann, and D. Moran, "A brain-computer interface using electrocorticographic signals in human," *Journal of Nerual Engineering*, vol. 1, no. 2, pp. 63–71, Jun 2004.

[14] 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.

[15] 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.

[16] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[17] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller, "Invariant feature extraction and classification in kernel spaces," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K.-R. Müller, Eds., vol. 12. MIT Press, 2000, pp. 526–532.

[18] 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.

[19] 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, 2004.

[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] H. Jasper and W. Penfield, "Electrocorticograms in man: Effect of voluntary movement upon the electrical activity of the precentral gyrus," *Arch. Psychiatrie Zeitschrift Neurol.*, vol. 183, pp. 163–74, 1949.

[22] G. Pfurtscheller and A. Arabibar, "Evaluation of event-related desynchronization preceding and following voluntary self-paced movement," *Electroenceph. clin. Neurophysiol.*, vol. 46, pp. 138–46, 1979.

[23] A. Schnitzler, S. Salenius, R. Salmelin, V. Jousmäki, and R. Hari, "Involvement of primary motor cortex in motor imagery: a neuromagnetic study," *Neuroimage*, vol. 6, pp. 201–8, 1997.

[24] V. Nikouline, K. Linkenkaer-Hansen, H. Wikström, M. Kesäniemi, E. Antonova, R. Ilmoniemi, and J. Huttunen, "Dynamics of mu-rhythm suppression caused by median nerve stimulation: a magnetoencephalographic study in human subjects," *Neuroscience Letters*, vol. 294, 2000.

[25] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Increase information transfer rates in BCI by CSP extension to multi-class," in *Advances in Neural Inf. Proc. Systems (NIPS 03)*, vol. 16, 2004, in press.

[26] K. Fukanaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[27] 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.

[28] F. Takens, "Detecting strange attractors in fluid turbulence," in *Dynamical Systems and Turbulence*, D. Rand and L. Young, Eds. Berlin: Springer-Verlag, 1981, pp. 366–381.

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