Research Article

Classifying EEG for Brain-Computer Interface: Learning Optimal Filters for Dynamical System Features

Le Song[1] and Julien Epps[2]

[1] *School of Information Technologies, The University of Sydney, N.S.W. 2006, Australia*
[2] *National ICT Australia, Locked Bag 9013 Alexandria, N.S.W. 1435, Australia*

Correspondence should be addressed to Le Song, lesong@it.usyd.edu.au

Classification of multichannel EEG recordings during motor imagination has been exploited successfully for brain-computer interfaces (BCI). In this paper, we consider EEG signals as the outputs of a networked dynamical system (the cortex), and exploit synchronization features from the dynamical system for classification. Herein, we also propose a new framework for learning optimal filters automatically from the data, by employing a Fisher ratio criterion. Experimental evaluations comparing the proposed dynamical system features with the CSP and the AR features reveal their competitive performance during classification. Results also show the benefits of employing the spatial and the temporal filters optimized using the proposed learning approach.

1. INTRODUCTION

A brain-computer interface (BCI) is a communication system that relies on the brain rather than the body for control and feedback [1]. Ideally, it should run in a servo mode, allowing the subjects to initiate the communication anytime and anywhere without resorting to external stimuli or triggers. Such an interface not only offers a promising prosthetic device for those severely paralyzed, but also signifies a radically new technology for the general public. Current BCI research is still in its early stage and the emphasis is placed on the design of algorithms to decode a prespecified set of brain states. This involves three main aspects.

Brain states

Only brain states consciously controllable by the subjects are suitable for BCI. Besides, these states should generate distinct, repeatable, and measurable patterns whenever accessed. Among the most commonly used brain states are imaginations of body movements (motor imaginations). Motor imaginations can reliably change the neural activities over sensorimotor cortices. Depending on the part of the body imagined moving, these changes exhibit distinct spatial distributions [2]. Recognition of these patterns can then be translated into control signals, as is the case in this study.

Recording devices

Motor imaginations can be recorded by both electroencephalography (EEG) and magnetoencephalography (MEG). EEG remains the most popular way to record BCI signals and will be the focus of this study. It measures scalp electrical activities diffused from the cortex. Compared to MEG, it is portable and inexpensive. However, EEG can only measure blurred cortical activities due to the diffusion of the skull and the skin. Thus, EEG is normally used for studying cortical patches in the centimeter scale. Furthermore, EEG signals are contaminated by noise from various sources, such as muscle activities and power line interference. Spatial and temporal filters are commonly applied before any further analysis [3, 4].

Decoding algorithms

Prefiltered EEG signals still contain considerable noise, which poses a challenge for its decoding. Statistical machine learning (ML) techniques have been introduced into BCI to combat these variations. Techniques like artificial neural networks, support vector machine (SVM) [5], and Linear Discriminant Analysis [4] have been employed to learn patterns from training EEG signals and then classify new EEG signals. This strategy often results in increased decoding success and

significant shortening of subject training time (from several months down to several days). The most prominent examples include the Berlin BCI [4], the MPI BCI [6], and the Graz BCI [7].

Apart from the classifiers, these ML-based BCIs also differ in the features they extract from EEG signals. The most successfully used features include autoregressive (AR) coefficients [6, 8] and common spatial patterns (CSP) [4, 7]. In this paper, we will employ a novel type of feature based explicitly on the neurophysiology of EEG signals instead. Basically, we consider EEG signals as the outputs of a networked dynamical system. The nodes of this system consist of cortical patches, while the links correspond to neural fibers. A large and complex system like this often generates interesting collective dynamics, such as synchronization in the activities of the nodes, and they result in the change of EEG patterns measured on the scalp. These features from the collective dynamics of the system can be employed for classification [9, 10]. This will be elaborated in Section 2.

To recover the cortical dynamics from the EEG signals, subject-specific spatial and temporal filtering is usually needed [4, 11]. Instead of manually tuning these filters, we propose a common framework in Section 3 to learn them from the data. Our basic idea is to optimize the filters so that the separability of the two classes is improved. Experimental results show that the learned filters not only reduce the classification errors of the dynamical system (DS) features, but also extract physically meaningful information from the EEG signals. Comparisons are also made between the DS features with the learned filters and the CSP and the AR features with manually tuned filters. These comparisons together with further comparisons to other filter learning methods, such as the CSSP [12] and CSSSP [4] methods, demonstrate the competitive performance of our method (Section 4). Finally, the conclusion is given in Section 6.

2. DYNAMICAL SYSTEM FEATURES

The cortex is a highly folded sheet of neurons ($\approx$100 billion neurons) and they self-organize into clusters. These neuronal clusters not only tightly connect with their neighbors, but also communicate with distal clusters through neural fibers. Each cluster is often associated with certain aspect of information processing. The collaboration of these clusters achieves the normal functioning of the brain. In this section, we will first describe a simple mathematical model of the cortex, and then show how it leads to dynamical system features related to motor imaginations.

2.1. Mathematical model of the cortex

Typically, a neuronal cluster will generate electrical oscillations. It has been modeled as an oscillator with phase θ and output s. Its dynamics are governed by a simple phase model [13]:

$$s = f(\theta),$$
$$\dot{\theta} = \omega + g(t), \qquad (1)$$

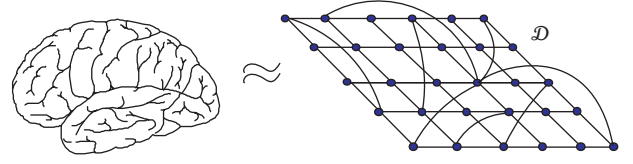FIGURE 1: Networked dynamical system model of the cortex.

where ω is the intrinsic frequency of the oscillation and f is a function 2π-periodic in θ. $g(t)$ is the input to the oscillator. $g(t)$ will accelerate the oscillation if it assumes positive values, and slow it down if negative.

The whole cortex can then be modeled as a networked dynamical system $\mathcal{D}$, as shown in Figure 1. Each node in the system represents a neuronal cluster and each link a neural interaction. The input, $g(t)$, to each neuronal cluster now consists of two parts: influence from other clusters and modulation by subcortical structures [2]. Suppose that the links of the network are represented as an adjacency matrix $\mathbf{G}$ ($\mathbf{G}_{ij} = 1$ if node i and j are connected; $\mathbf{G}_{ij} = 0$ otherwise). Then, the dynamics of a node i take a more specific form:

$$\dot{\theta}_i = \omega_i + \sum_j \epsilon_{ij} \mathbf{G}_{ij}(s_j - s_i) + h_i(t), \qquad (2)$$

where s_i and s_j denote the outputs from node i and j, respectively, $\sum_j \epsilon_{ij} \mathbf{G}_{ij}(s_j - s_i)$ represents the influence from other nodes, and $h_i(t)$ is the subcortical input. Note that there is an added parameter ϵ_{ij} in (2), which controls the strength of the influence from node j to i.

2.2. Desynchronization of neuronal clusters

Two properties of the network of oscillators in (2) are of particular interest to BCI [13].

(i) Without the input $h(t)$, all nodes will settle down into an oscillation of the same frequency ω_0, if the network is connected and the influence ϵ is sufficiently strong (mutual synchronization).

(ii) If the input $h_i(t)$ to node i is sufficiently strong and oscillates at a frequency ω_0, node i will then be forced to oscillate in the same frequency ω_0 (forced synchronization).

These two properties explain well the spatial distribution of the EEG signals during motor imaginations [2].

(i) If no imagination is carried out, the neuronal clusters in the idle sensorimotor cortex tend to synchronize with each other and oscillate in the frequency range of 8–26 Hz (EEG α and β rhythm). The spatial summation of this unison is a strong α (and/or β) rhythm in EEG signals.

(ii) If the subject is actively engaged in motor imaginations, the associated neuronal clusters will be strongly modulated by the subcortical structures. The dynamics of these clusters will then stray away from their former synchronous state. This results in a decrease of α (and/or β) power in EEG signals.

This phenomenon is called event-related desynchronization (ERD) in the neuroscience literature. Depending on the part of the body imagined moving, neuronal clusters at different locations will be active. These clusters desynchronize with other clusters, and the spatial distribution of the desynchronization will be different as the imagination contents change. ERD suggests that the strength of the synchronization between neuronal clusters can be used as features for classification [9, 10].

2.3. Features for motor imaginations

An EEG electrode measures mostly the activities of the neuronal cluster directly underneath it (we will qualify this in Section 3). Suppose that the pairwise synchronization of the measured neuronal clusters can be computed from EEG signals and organized into a matrix $\mathbf{S}$ ($\mathbf{S}$ is symmetric with entry $\mathbf{S}_{ij}$ for clusters i and j). Each entry in $\mathbf{S}$ is a dynamical system feature and the similarity between two EEG signals can then be quantified in terms of these features as follows:

$$k(\mathbf{S}, \widetilde{\mathbf{S}}) = \mathrm{Tr}\left((\mathbf{S} \circ \mathbf{A})^{\mathsf{T}}(\widetilde{\mathbf{S}} \circ \mathbf{A})\right), \qquad (3)$$

where $\mathbf{A}$ is a weighting matrix, $\mathrm{Tr}(\cdot)$ computes the trace of a matrix, and $\circ$ represents element-wise matrix product. Essentially, this measure transforms EEG trials into synchronization features and then computes their similarity based on these features. Since we will use a SVM classifier for our later experiments, $k(\cdot, \cdot)$ can be interpreted as a kernel between EEG trials.

Preliminary analysis of our motor imagination data set (this data set is further explained in Section 4) indicates that the synchronization in our data appears to be either inphase ($\theta_i - \theta_j = 0$) or antiphase ($\theta_i - \theta_j = \pi$). These two types of synchronization can be well detected simply using the covariance. Therefore, classifying EEG signals using the DS features consists of three steps.

(i) Filter EEG signals. This is the step where filter learning techniques are applied. For our method, filters are learned for individual channels. Hence, EEG signals from different channels are filtered differently.

(ii) Compute the entries of $\mathbf{S}$ and apply $\mathbf{A}$. In this paper, $\mathbf{S}$ is simply the sample covariance matrix, and this is computed for each trial separately. Each entry in $\mathbf{S}$ is a DS feature, and the matrix $\mathbf{A}$ is mainly used for selecting the features. For instance, by setting 20 entries of $\mathbf{A}$ (in (3)) to 1 and all others to 0, then only 20 features are used for later classification.

(iii) Compute the kernel $k(\cdot, \cdot)$ (in (3)) for pairs of trials, form the kernel matrix $\mathbf{K}$, and pass it to SVM for classification. The entry in $\mathbf{K}$ corresponding to trial $\mathbf{S}$ and $\widetilde{\mathbf{S}}$ is simply $k(\mathbf{S}, \widetilde{\mathbf{S}})$ (as in (3)).

3. LEARNING OPTIMAL FILTERS

Filtering EEG signals is important for later classifications. Due to the diffusion of the skull and skin, an EEG electrode actually measures a mixture of signals from several neuronal clusters. Spatial filters, such as a Laplacian filter, are usually applied to concentrate the signals to a single neuronal cluster. Furthermore, EEG signals are contaminated by various noises, such as electrical signals from muscle movements. Our interest lies in oscillation in the frequency range of 8–26 Hz (α and β rhythm). Bandpass filtering is usually needed to suppress other signals.

As previous BCI researchers have experienced [4], the optimal filters for each subject are very different, and it is quite inconvenient to manually choose these filters. Attempts have been made to learn these filters from the training EEG data. Pioneering works have been reported in [4, 12], where FIR (temporal) filters are learned for the CSP features to improve the separability of the two classes. Our work is inspired by their ideas, but our approach is different in two aspects. First, our approach is directed to the dynamical system features. Second, we have proposed a common framework for the learning of both the spatial and the temporal filters. In the following sections, the common framework is first described before it is specialized into the spatial and the temporal filter learning.

3.1. New framework

Our filter learning framework involves three steps: (i) quantify the quality of a feature using the Fisher ratio; (ii) express the Fisher ratio using the filter parameters; (iii) and then maximize the Fisher ratio with respect to the filter parameters. Given the data and the filter parameter $\mathbf{a}$, our framework can be formulated mathematically as follows

$$\max_{\mathbf{a}} Q(\mathbf{a}) = \frac{(\mu_+(\mathbf{a}) - \mu_-(\mathbf{a}))^2}{\sigma_+^2(\mathbf{a}) + \sigma_-^2(\mathbf{a})}, \qquad (4)$$

where Q is the fisher ratio, μ the mean value of a feature, and σ its standard deviation (the subscripts $+$ and $-$ restrict computation to positive and negative classes, resp.). Higher values of Q usually indicate better separation of the two classes. This learning framework can be applied to various problems. However, only local optimum can be guaranteed for the solution, since Q is in general not convex in terms of $\mathbf{a}$. This is also the case in learning the filters for the DS features. To find an optimal solution efficiently, we will employ the subspace optimization technique.

The filter learning is performed on each pair of EEG electrodes separately. For a pair, two filters are learned, one for each electrode. Suppose that the parameters of the two filters are $\mathbf{a}$ and $\mathbf{b}$, respectively. It turns out that for both the spatial and the temporal filtering, Q assumes a form biquadratic in $\mathbf{a}$ and $\mathbf{b}$. For instance, if $\mathbf{b}$ is fixed, Q becomes the quotient between $\mathbf{a}^{\mathsf{T}}\mathbf{V}(\mathbf{b})\mathbf{a}$ and $\mathbf{a}^{\mathsf{T}}\mathbf{W}(\mathbf{b})\mathbf{a}$, where $\mathbf{V}(\mathbf{b})$ and $\mathbf{W}(\mathbf{b})$ are matrices quadratic in $\mathbf{b}$. The optimal $\mathbf{a}$ can then be obtained by solving the following constrained optimization problem:

$$\max_{\mathbf{a}} \mathbf{a}^{\mathsf{T}}\mathbf{V}(\mathbf{b})\mathbf{a}, \quad \text{s.t. } \mathbf{a}^{\mathsf{T}}\mathbf{W}(\mathbf{b})\mathbf{a} + \gamma\mathbf{b}^{\mathsf{T}}\mathbf{b}\mathbf{a}^{\mathsf{T}}\mathbf{a} = c. \qquad (5)$$

Note that the additional term $\gamma\mathbf{b}^{\mathsf{T}}\mathbf{b}\mathbf{a}^{\mathsf{T}}\mathbf{a}$ does not originate from Q. It is a regularized product of the norms of $\mathbf{a}$ and $\mathbf{b}$, and the strength of this regularization is controlled by γ.

```
Input: random initial values of a and b
Output: optimized value of a and b
(1) repeat
(2)     ã ← a
(3)     b̃ ← b
(4)     compute V(a), W(a), Ṽ(a), and W̃(a)
(5)     b ← eigenvector of Ṽ(a)b = λW̃(a)b
(6)     compute V(b), W(b), Ṽ(b), and W̃(b)
(7)     a ← eigenvector of Ṽ(b)a = λW̃(b)a
(8) until ‖a − ã‖² < ε and ‖b − b̃‖² < ε
```

ALGORITHM 1: Learning optimal filter $\mathbf{a}$ and $\mathbf{b}$.

Using the Lagrange multiplier method (let the multiplier be λ), the optimal $\mathbf{a}$ can be derived from the following generalized eigenvector problem:

$$\widetilde{\mathbf{V}}(\mathbf{b})\mathbf{a} = \lambda \widetilde{\mathbf{W}}(\mathbf{b})\mathbf{a}, \qquad (6)$$

where

$$\begin{aligned}\widetilde{\mathbf{V}}(\mathbf{b}) &= \mathbf{V}(\mathbf{b}) + \mathbf{V}(\mathbf{b})^\mathsf{T}, \\ \widetilde{\mathbf{W}}(\mathbf{b}) &= \mathbf{W}(\mathbf{b}) + \mathbf{W}(\mathbf{b})^\mathsf{T} + 2\gamma\mathbf{b}^\mathsf{T}\mathbf{b}\mathbf{I}.\end{aligned} \qquad (7)$$

The optimal $\mathbf{a}$ is then the generalized eigenvector corresponding to the largest eigenvalue. Similarly, $\mathbf{b}$ can be optimized by fixing $\mathbf{a}$. Local maxima can then be found by optimizing $\mathbf{a}$ and $\mathbf{b}$ alternately (Algorithm 1). In our experiments, the solution usually changes very little after two iterations, and henceforth only two iterations are used. To specialize this algorithm into the learning of the spatial and the temporal filters, we only need to derive the exact forms of $\mathbf{V}(\mathbf{a})$, $\mathbf{W}(\mathbf{a})$, $\mathbf{V}(\mathbf{b})$, and $\mathbf{W}(\mathbf{b})$ for these two cases, respectively.

3.2. Learning spatial filters

Studies show that the spherical spline Laplacian filter is useful for the study of cortical dynamics [14]. This method models the shape of the head as a unit sphere and uses orthogonal bases on the sphere to spatially interpolate EEG signals [3]. The filtering is then achieved by computing the analytical Laplacian of the interpolation function. This filters only high-passes EEG signals, and is unable to emphasize interesting signals in the middle frequency range [11] (also see Figure 2). This section will start with a reformulation of the spherical spline Laplacian filter, which leads to a class of spatial filters. The exact forms of $\mathbf{V}(\mathbf{a})$, $\mathbf{W}(\mathbf{a})$, $\mathbf{V}(\mathbf{b})$, and $\mathbf{W}(\mathbf{b})$ are then derived.

For square integrable functions on a sphere, the Legendre polynomials $p_n(\cdot)$ evaluated at $\cos\theta$ constitute a set of orthogonal bases. The parameter n is the degree of the polynomial and it controls the spatial frequency of a basis. A p_n with larger n will generally represent higher spatial frequency. θ is the latitudinal (zonal) angle. In this study, a maximum of $n = 20$ is used for the interpolation of EEG signals (due to the low spatial variation of EEG signals).

Suppose that a position on the unit sphere is $\mathbf{e}$, and the position of the ith EEG electrode is $\mathbf{e}_i$. Let $\cos(\mathbf{e}, \mathbf{e}_i)$ denote the cosine of the angle between $\mathbf{e}$ and $\mathbf{e}_i$, we can construct a matrix $\mathbf{P}(\mathbf{e})$ with entries:

$$(\mathbf{P}(\mathbf{e}))_{in} = \frac{1}{4\pi}\frac{2n+1}{(n(n+1))^4}p_n(\cos(\mathbf{e}, \mathbf{e}_i)), \qquad (8)$$

where i ranges through the index of the electrodes, and $n = 1 \cdots 20$. Then, EEG signals at position $\mathbf{e}$ can be interpolated as follows:

$$\mathbf{u}(\mathbf{e}) = \mathbf{c}_0 + \mathbf{C}^\mathsf{T}\mathbf{P}(\mathbf{e})\mathbf{1}, \qquad (9)$$

where $\mathbf{u}(\mathbf{e})$ is a vector with each dimension corresponding to a time point. $\mathbf{1}$ is a vector of all ones. $\mathbf{c}_0$ (a vector of the same size as $\mathbf{u}(\mathbf{e})$) and $\mathbf{C}^\mathsf{T}$ (a matrix with 20 columns and the same number of rows as $\mathbf{u}(\mathbf{e})$) are the interpolation coefficients estimated from actual EEG signals. The solution of these coefficients can be found using two constraints [3]: (i) the interpolated function has to pass the actual EEG measurements; (ii) $\mathbf{C}^\mathsf{T} \cdot \mathbf{1} = \mathbf{0}$. Our formulation in (9) is equivalent to (1) in Perrin's original formulation [3]. The difference is that (9) describes the interpolation for each time point of a time series rather than that of a single time point. Our reformulation simply stacks separate interpolation for each time point into a matrix notation. This provides us with insight to how spatial filtering is performed.

Spatial filtering of EEG signals can then be achieved by simply removing the DC component $\mathbf{c}_0$ and reweighting other frequency components (the bases). Suppose that the filter (weighting) is $\mathbf{a}$. Thus, spatial filtering can be computed as follows:

$$\widetilde{\mathbf{u}}(\mathbf{e}_i) = \mathbf{C}^\mathsf{T}\mathbf{P}(\mathbf{e}_i)(\mathbf{1} \circ \mathbf{a}) = \mathbf{C}^\mathsf{T}\mathbf{P}(\mathbf{e}_i)\mathbf{a}. \qquad (10)$$

The spherical spline Laplacian filter can be obtained by simply setting entries of $\mathbf{a}$ to $-n(n+1)$ (equivalent to [3, equation (5)]). With formula (10), other types of filtering can also be implemented by varying $\mathbf{a}$. For example, a bell-shaped bandpass filter can be obtained by setting the entries of $\mathbf{a}$ to $\exp(-\kappa n(n+1))n(n+1)$ (κ is a parameter controlling the width and the peak). These two filters are illustrated in Figure 2. Note that the weights in the figures are normalized into the range between 0 and 1.

Suppose that filter $\mathbf{a}$ and $\mathbf{b}$ are applied to electrode $\mathbf{e}_i$ and $\mathbf{e}_j$, respectively, the covariance between the two filtered EEG signals can then be computed as

$$\mathsf{cov}^{ij} = \frac{1}{l}\widetilde{\mathbf{u}}(\mathbf{e}_i)^\mathsf{T}\widetilde{\mathbf{u}}(\mathbf{e}_j) = \frac{1}{l}\mathbf{a}^\mathsf{T}\mathbf{P}^\mathsf{T}(\mathbf{e}_i)\mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{P}(\mathbf{e}_j)\mathbf{b}, \qquad (11)$$

where l is the number of time points. Further, denote $\widetilde{\mathbf{C}}^{ij} = \mathbf{P}^\mathsf{T}(\mathbf{e}_i)\mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{P}(\mathbf{e}_j)$. (Since the following derivations are the same for each pair of electrodes, the superscripts ij are dropped henceforth for convenience.) Then, μ_+ in (4) can be computed as follows:

$$\mu_+ = \frac{1}{m}\sum_{k \in +}\mathsf{cov}_k = \mathbf{a}^\mathsf{T}\left(\frac{1}{ml}\sum_{k \in +}\widetilde{\mathbf{C}}_k\right)\mathbf{b} = \mathbf{a}^\mathsf{T}\mathbf{D}_+\mathbf{b}, \qquad (12)$$
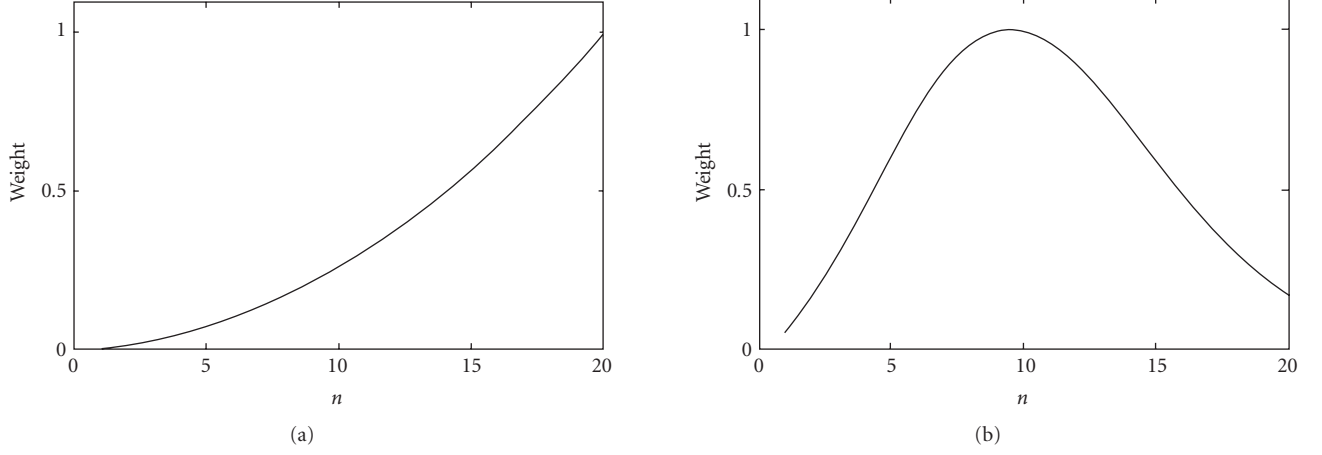
FIGURE 2: (a) Spherical spline Laplacian filter and (b) a bell-shaped filter.

where $k \in +$ means that the index ranges through all m trials in the positive class. (Suppose that the negative class also has m trials). The variance σ_+ can be computed as follows:

$$(\sigma_+)^2 = \frac{1}{m} \sum_{k \in +} (\mathrm{cov}_k - \mu_+)^2 = \mathbf{a}^\mathsf{T} \mathbf{E}_+(\mathbf{b})\mathbf{a}, \qquad (13)$$

where

$$\mathbf{E}_+(\mathbf{b}) = \frac{1}{m^2 l^2} \left(\left(\sum_{k \in +} \widetilde{\mathbf{C}}_k \right) \mathbf{b} \right)^2 - \frac{1}{m l^2} \sum_{k \in +} (\widetilde{\mathbf{C}}_k \mathbf{b})^2. \qquad (14)$$

Similarly, $\mu_- = \mathbf{a}^\mathsf{T} \mathbf{D}_- \mathbf{b}$ and $(\sigma_-)^2 = \mathbf{a}^\mathsf{T} \mathbf{E}_-(\mathbf{b})\mathbf{a}$. $\mathbf{V}(\mathbf{b})$ and $\mathbf{W}(\mathbf{b})$ can then be derived as follows:

$$\begin{aligned} (\mu_+ - \mu_-)^2 &= \mathbf{a}^\mathsf{T} (\mathbf{D}_+ \mathbf{b} - \mathbf{D}_- \mathbf{b})^2 \mathbf{a} = \mathbf{a}^\mathsf{T} \mathbf{V}(\mathbf{b})\mathbf{a}, \\ (\sigma_+)^2 + (\sigma_-)^2 &= \mathbf{a}^\mathsf{T} (\mathbf{E}_+(\mathbf{b}) + \mathbf{E}_-(\mathbf{b}))\mathbf{a} = \mathbf{a}^\mathsf{T} \mathbf{W}(\mathbf{b})\mathbf{a}. \end{aligned} \qquad (15)$$

Since $\mathbf{a}$ and $\mathbf{b}$ are symmetric, $\mathbf{V}(\mathbf{a})$ and $\mathbf{W}(\mathbf{a})$ can be derived analogously by exchanging the positions of $\mathbf{a}$ and $\mathbf{b}$ and transposing $\widetilde{\mathbf{C}}_k$ in (12)–(15). Substituting $\mathbf{V}(\mathbf{a})$, $\mathbf{W}(\mathbf{a})$, $\mathbf{V}(\mathbf{b})$, and $\mathbf{W}(\mathbf{b})$ into Algorithm 1 will then produce the optimal filters.

3.3. Learning temporal filters

Unlike [4] that formulated the learning of the temporal filters in the time domain (FIR filter), our formulation works directly in the frequency domain. The basic idea of our approach is to place weighting directly on the complex coefficients of the discrete Fourier transformation (DFT).

Weighting the frequency components of an EEG signal $\mathbf{u}(\mathbf{e}_i)$ will transform it to

$$\widetilde{\mathbf{u}}(\mathbf{e}_i) = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{u}(\mathbf{e}_i)) \circ \mathbf{a}), \qquad (16)$$

where $\mathbf{a}$ is the filter (weighting), and $\mathcal{F}$ represents the forward DFT ($\mathcal{F}^{-1}$, the inverse DFT). Suppose that filters $\mathbf{a}$ and $\mathbf{b}$ are applied to EEG electrodes $\mathbf{e}_i$ and $\mathbf{e}_j$, respectively. The covariance of the filtered signals can then be computed as follows:

$$\begin{aligned} \mathrm{cov} &= \frac{1}{l} \widetilde{\mathbf{u}}(\mathbf{e}_i)^\mathsf{T} \widetilde{\mathbf{u}}(\mathbf{e}_j) \\ &= \frac{1}{l} (\mathcal{F}^{-1}(\mathcal{F}(\mathbf{u}(\mathbf{e}_i)) \circ \mathbf{a}))^\mathsf{T} (\mathcal{F}^{-1}(\mathcal{F}(\mathbf{u}(\mathbf{e}_i)) \circ \mathbf{b})). \end{aligned} \qquad (17)$$

(Note that the superscripts are dropped for convenience.) Computation (17) is inefficient, since two forward and inverse DFTs are needed. The computation, however, can be reduced using the correlation theorem. This theorem states that the covariance between two signals $\mathbf{u}(\mathbf{e}_i)$ and $\mathbf{u}(\mathbf{e}_j)$ is equal to $(\mathcal{F}(\mathbf{u}(\mathbf{e}_i)))^* \mathcal{F}(\mathbf{u}(\mathbf{e}_j))$ ($*$ denotes conjugate transpose). Thus, (17) can be simplified to:

$$\mathrm{cov} = \frac{1}{l} \mathbf{a}^\mathsf{T} \mathrm{Diag}((\mathcal{F}(\mathbf{u}(\mathbf{e}_i)))^* \circ \mathcal{F}(\mathbf{u}(\mathbf{e}_j)))\mathbf{b}, \qquad (18)$$

where $\mathrm{Diag}(\cdot)$ transforms its vector argument into a diagonal matrix. Formula (18) requires only two DFT computations, and hence it is more efficient.

The derivations for $\mathbf{V}(\mathbf{a})$, $\mathbf{W}(\mathbf{a})$, $\mathbf{V}(\mathbf{b})$, and $\mathbf{W}(\mathbf{b})$ become straightforward if we compare (18) with (11). By setting $\widetilde{\mathbf{C}}^{ij} = \mathrm{Diag}((\mathcal{F}(\mathbf{u}(\mathbf{e}_i)))^* \circ \mathcal{F}(\mathbf{u}(\mathbf{e}_j)))$, they can be obtained from (12)–(15). Substituting these matrices into Algorithm 1 produces the optimal filters.

4. RESULTS AND COMPARISON

The dynamical system (DS) features and the filter learning approach are evaluated using data set IVa from the Berlin BCI group [8]. This data set contains EEG signals (118 channels, sampled at 100 Hz) for five healthy subjects (labeled "aa," "al," "av," "aw" and "ay," resp.). During the recordings, they were prompted by visual cues to imagine for 3.5 seconds

TABLE 1: Averaged dynamical system features (DS) and common spatial patterns (CSP) for five subjects ("aa," "al," "av," "aw," and "ay") over three-time windows (0–0.5 second, 0.5–2.5 seconds, and 3.5–4.0 seconds) during the motor imagination. The top 20 most discriminative DS features are shown as edges connecting the corresponding electrodes (dots). The most discriminative CSPs for right-hand imagination are plotted as color maps.
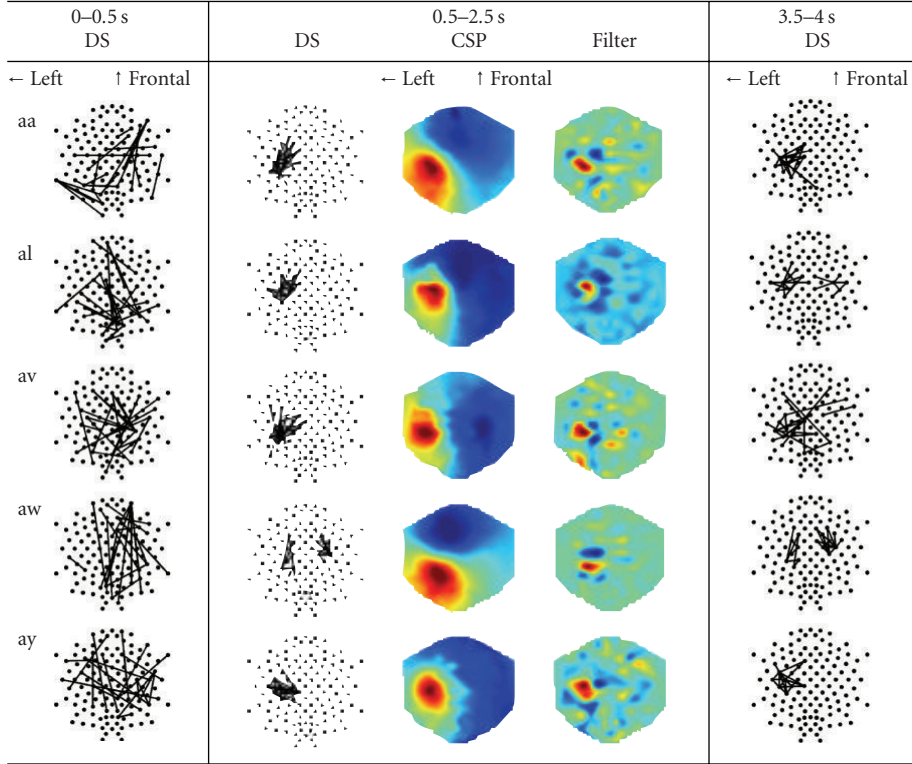
either right-hand (the positive class) or right-foot movements (the negative class). Our classification analysis will focus on the data between 0.5 seconds and 2.5 seconds (i.e., 200 time points for each channel), since in an online BCI setting a sliding window seldom exceeds 2 seconds [4]. For convenience, the period between 0 s and 1 s of a trial is called imagination preparation stage, and the period between 3.5 seconds and 4.0 seconds is called postimagination stage. Each type of imagination was carried out 140 times. Thus, there are 280 labeled trials in total for each subject (note: each trial is a multivariate time series of 118 dimensions). The task is to classify the type of the imagination for each trial in an offline fashion.

Two sets of experiments were conducted in our evaluations. They are designed to reveal two major aspects: (i) using DS features for classification, and (ii) learning spatial and temporal filters. We will first describe procedures common to these experiments. All our classifications are carried out using SVM and the errors are obtained from 10×10 fold cross-validation. An identical temporal prefiltering (bandpass between 8–40 Hz) is applied to all subjects. In the case of the DS features, an identical spatial prefiltering (a bell-shaped bandpass filter $\exp(-\kappa n(n + 1))n(n + 1)$ with $\kappa = 0.01$) is also applied for all subjects. Furthermore, only the top 20 DS features (in terms of their Fisher ratios) are used for classification.

4.1. Dynamical system features

4.1.1. Where are the discriminative dynamical system features?

The dynamical system (DS) features during motor imagination (0.5–2.5 s) are scored by Fisher ratio for each fold of the cross-validation, and these scores are further averaged over the folds. The top 20 most discriminative DS features are plotted in the second column of Table 1. For comparison, typical common spatial patterns (CSPs) for the right-hand imagination (corresponding to the smallest generalized eigenvalues) are also shown beside the DS features.

For four of the five subjects ("aa," "al," "av," and "ay"), the DS features share clear pattern across the subjects—they tightly concentrate on the area in charge of right-hand imagination (left motor cortex, hand region in the Homunculus). This phenomenon can be well explained by the theory of event-related desynchronization (ERD): as the hand region in the left motor cortex is actively engaged in imagination, its neuronal activities deviate from those of the neighboring cortices; and such localized spatial discordance results in the tight cluster of the DS features.

Furthermore, the typical common spatial patterns (CSP) also show nice agreement with the DS features. The areas of the ERD correspond to the peaks in the CSPs.

Table 2: Classification errors (%) of the CSP, the AR, and the DS features with optimized filters.

Sb	CSP	AR	DS + S + T
aa	**8.5 ± 5.4**	10.5 ± 6.0	*9.5 ± 5.7*
al	**0.8 ± 1.8**	*1.6 ± 2.5*	2.7 ± 3.1
av	29.1 ± 8.2	*23.3 ± 7.6*	**21.5 ± 7.6**
aw	**3.1 ± 2.8**	7.7 ± 3.8	6.5 ± 4.5
ay	**5.3 ± 3.8**	9.5 ± 4.4	*8.5 ± 5.0*

Table 3: Classification errors (%) of the combinations of the CSP, the AR, and the DS features.

Sb	CSP + AR	CSP + DS	AR + DS	ALL
aa	*7.6 ± 5.0*	7.3 ± 5.1	7.7 ± 4.7	7.3 ± 4.9
al	1.6 ± 2.3	**0.9 ± 1.9**	1.6 ± 2.5	*1.5 ± 2.2*
av	22.3 ± 7.4	22.5 ± 7.8	21.4 ± 7.4	*21.6 ± 7.1*
aw	3.5 ± 3.2	2.8 ± 3.1	5.2 ± 3.8	*3.4 ± 3.2*
ay	8.9 ± 4.6	**5.5 ± 4.3**	9.1 ± 4.6	*8.7 ± 4.5*

Beside the similarity also revealed in the figures is the difference of the DS features across subjects. Especially for subject "aw," half of the DS features locate in the contralateral hand region. A plausible explanation is that the subject may have imagined movements of both hands.

4.1.2. How do dynamical system features evolve over time?

The top 20 DS features in the imagination preparation stage (0–0.5 s) and the postimagination stage (3.5–4.0 s) are scored similarly and plotted, respectively, in the first and the third column of Table 1. These figures provide us an idea of the evolution of the DS features over time.

During the preparation stage, the DS features scatter around the whole scalp. They mostly connect distal regions of the brain; other than that, no clear pattern is shared across subjects. In fact, these DS features provide classifications only slightly better than random (the errors are not reported). This implies that the DS features within this period do not contain useful information for classification.

During the imagination, tight clusters of DS features are formed and they lead to good classification. Then, as the subjects are signaled to stop their imaginations (3.5–4.0 s), the clusters start to diffuse into wider areas of the brain. Such trend is most clearly revealed in subject "av," where parts of the DS features are replaced by long range connections across hemispheres of the brain.

The formation and the dissolution of clusters over the course of an imagination present a unique characteristic for the DS features. Potentially, such pattern can be exploited for online detection of motor imagination.

4.1.3. Dynamical system features are competitive

The DS features obtained with learned filters were compared to the CSP and the AR features obtained with manually cho-sen parameters. The parameters for the CSP features (filtering frequency, selected channels, and the number of projection subspaces) and the AR features (filtering frequency, selected channels, and the order of the AR model) were tuned according to the winning entry of BCI competition III [15]. The results are shown in Table 2.

Overall, the CSP features perform the best, the DS features follow, and the AR features produce lower accuracy. Furthermore, the DS features often obtain the best (highlighted in bold) or the second best place (highlighted in italic). Especially for subject "av," the DS features outperform the CSP features by 6%. It is important to note that the parameters for the CSP and AR features have been tuned manually and intensively, while the results for the DS features are obtained with exactly the same starting parameters. This shows the usefulness of the DS features and our filter learning approach.

4.1.4. Dynamical system features extract complementary information

The CSP, AR, and DS features are computed differently from the EEG signals. An interesting question is whether they complement each other during classification. To investigate this, we combine more than two types of features (CSP + AR, CSP + DS, AR + DS, and ALL three) using the META scheme described by [8]. The classifications of the combined features are presented in Table 3. The combination with the smallest error for each subject is highlighted in bold and the second place in italic. Furthermore, we surround an error with a box, if it is the smallest ever (in Tables 2 and 3) for a subject.

The DS features indeed complement the CSP and the AR features, as is evidenced by the further reduction of errors in subject, "aa," "av," and "aw." The reduction, however, is not large (the largest being around 1% for subject "aa"). Furthermore, the combination of all three types of features does not necessarily further reduce the errors. This happens when the best features have already extracted almost all information about the separability of the two classes. Additional features may only provide redundant or even conflicting information for the classification. This is very likely in our case since we have optimized each type of features intensively. Finally, our results suggest that the combination of the CSP and the DS features performs the best, and the DS features complement the CSP features better than the AR features.

4.2. Learned filters

4.2.1. Learned filters improve classification

For each pair of EEG electrodes (equivalent to a DS feature), the optimal spatial and temporal filters were learned sequentially. In Table 4, we present the classification errors using the following: (i) the DS features without the spatial and the temporal filter optimization (DS column); (ii) the DS features only with the spatial filter optimization (DS + S column); (iii) the DS features only with the temporal filter optimization (DS + T column); (iv) the DS features with both the

TABLE 4: Classification errors (%) of the DS features before and after applying the learned filters.

Sb	DS	DS + S	DS + T	DS + S + T
aa	16.7 ± 7.2	14.6 ± 7.0	**9.7 ± 5.7**	9.5 ± 5.7
al	3.7 ± 3.3	**3.2 ± 3.2**	3.6 ± 3.4	2.7 ± 3.1
av	27.3 ± 7.9	25.1 ± 8.0	**21.4 ± 7.9**	21.5 ± 7.6
aw	13.1 ± 6.0	12.1 ± 5.7	**7.5 ± 4.4**	6.2 ± 4.5
ay	11.0 ± 5.3	**9.6 ± 5.0**	9.7 ± 5.1	8.5 ± 5.0

spatial and the temporal filter optimization (DS + S + T column). Note that for all four comparisons prefilters have already been applied in both temporal and spatial domains.

The results demonstrate that both the learned spatial and temporal filters improve the classification (DS + S and DS + T columns). Although there is no absolute winner in the two types of filters, when applied separately, the temporal filters outperform the spatial filters in general (the winning filter for each subject is highlighted in bold). Especially for subjects "aa" and "aw," the temporal filters reduce about 5% more errors than the spatial filters.

The combined application of the learned filters almost always further reduces the errors (only subject "av" slightly violates this rule). The maximum reduction is around 7% (for subject, "aa" and "aw"). The errors obtained (DS + S + T column) are now lower than 10% for 4 of the 5 subjects (except "av"). It seems that the learned filters help less for some subjects ("al" and "ay"). The reason can be that the prefiltering is already near the optimal solution.

The classification for subject "av" has the largest error. Our preliminary studies indicate that the most responsive frequency range of this subject shifts above 26 Hz (contrary to the usual 8–26 Hz). While most energy in the EEG signals concentrates below 26 Hz, this makes it difficult to extract good features for the subject.

4.2.2. Learned filters extract meaningful information

Several details related to Section 4.2.1 are clarified here. The spatial and the temporal filters can be interpreted as weighting in the corresponding frequency domain. We have further restricted them to be polynomial models in our experiments. The results in Table 4 are obtained with polynomial functions of degree 6 (for both the spatial and the temporal filter learnings). The regularization parameters γ for the spatial and the temporal filters are 10^{-7} and 10^{-13}, respectively. For the case of the temporal filter, a bell-shaped prefilter is also applied $(-\exp(-\kappa n(n+1))n(n+1)$ with $\kappa = 0.001$ for all subjects). Note that the filters are always learned in pairs, that is, one for each channel in a pair. We will illustrate the learned filters in two ways.

The first way is the joint effect of the bell-shaped prefilter and a learned filter from a single channel. Since the learned filter is always applied after the prefiltering, we will show the shape of the prefilter, the learned filter, and their multiplication in one picture (Figures 3(a) and 4(a)).

The second way is the joint effect of the overall filtering from two channels. Since a DS feature is bilinear in the filters applied to the two channels, our optimization in Algorithm 1 only has exact control over their multiplicative effect. Therefore, we will illustrate the filtering effects for two channels and their multiplication in one picture (Figures 3(b) and 4(b)).

Figure 3(a) shows a learned spatial filter (thin line, bow-shaped) and the prefilter (thin line, bell-shaped) for one channel. Although both filters are simple, their multiplicative effect creates a double-peak characteristics (dotted line). This is equivalent to emphasizing the frequency contributions under these two peaks. The overall effect of the learned filters from two channels (dotted lines in Figure 3(b)) is also double peaked (thick line in Figure 3(b)). We believe that these peaks are somehow related to the electrode spacing on the scalp. It is likely that the learned filters weight the information from the actual electrodes more heavily than that from the interpolated positions.

For the temporal filters, we will interpret the learned filters in terms of their effects on the power spectrum. Hence, only the absolute values of the weighting are displayed. The final filter for an example channel (dotted line in Figure 4(a); it is the multiplication of a prefilter and a learned filter, both in thin lines) does not appear to emphasize the motor imagination signals (i.e., ERD in 8–26 Hz). The meaning, however, becomes clearer when we examine the filters from two channels together. In Figure 4(b), the filters from two channels are shown in dotted lines and their multiplication in thick line. The multiplication creates the strongest peak within 10–18 Hz, and a second strongest peak within 18–28 Hz. This corresponds well to the most responsive frequency range of the motor imaginations.

Note that one can *not* simply replace individual filters, **a** and **b**, from a pair of electrodes by the square root of their multiplication. This is because the two filters **a** and **b** always appear in the form of $\mathbf{ba}^{\mathsf{T}}$ in the objective and the constraint of (5). For instance, one can show that according to (15)

$$\mathbf{a}\mathbf{V}(\mathbf{b})\mathbf{a} = \mathrm{Tr}\left(\left((\mathbf{D}_+ - \mathbf{D}_-)\mathbf{ba}^{\mathsf{T}}\right)^2\right). \tag{19}$$

Therefore, only when two pairs of filters, **a** and **b** versus **a**′ and **b**′, produce the same outer product (i.e., $\mathbf{ba}^{\mathsf{T}} = \mathbf{b}'\mathbf{a}'^{\mathsf{T}}$), they can be equated with each other. In Figures 3(b) and 4(b), we only showed the diagonal of $\mathbf{ba}^{\mathsf{T}}$ to produce a concise summary of their joint filtering effect. One should keep in mind that the learned filters have further effect beyond what is visualized here.

4.2.3. Learned filters are competitive

The DS features obtained with the learned filters were compared to the CSP features produced by the CSSP [12] and the CSSSP [4] methods. These two methods are also designed to remove the manual filter tuning, and they have incorporated the filter learning into the original CSP method. The comparisons are presented in Table 5.
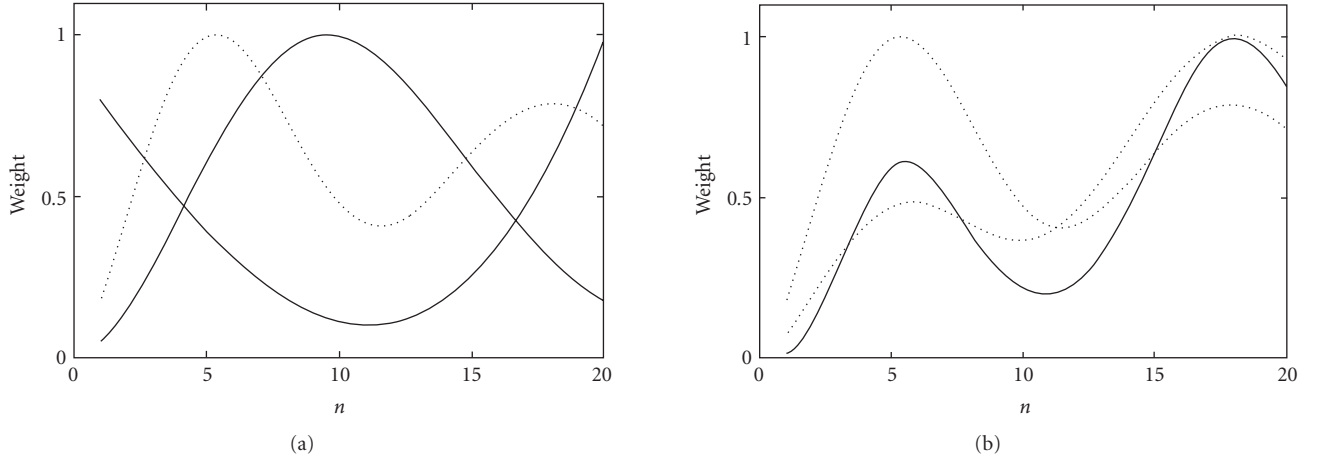
(a)

(b)

FIGURE 3: Illustration of spatial filters: (a) the prefilter (thin line, bell-shaped), a learned filter (thin line, bow-shaped), and their multiplication (dotted line); (b) learned filters from a pair of channels (dotted lines) and their multiplication (thick line).
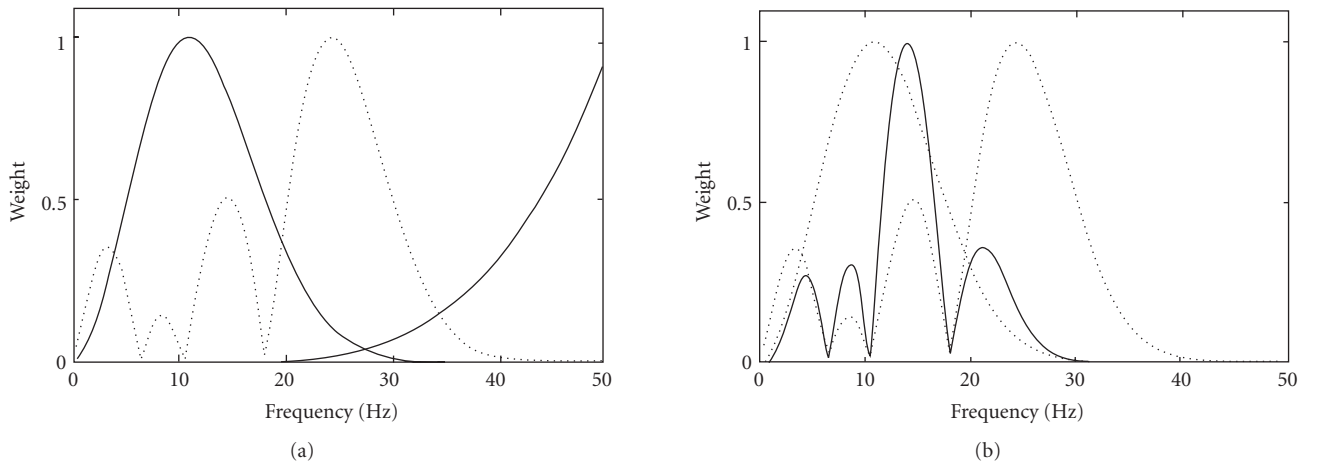

(a)

(b)

FIGURE 4: Illustration of temporal filters: (a) the pre-filter (thin line, bell-shaped), a learned filter (thin line, wedge-shaped), and their multiplication (dotted line); (b) learned filters from a pair of channels (dotted lines) and their multiplication (thick line).

TABLE 5: Classification errors (%) of the CSSP, the CSSSP, and the DS + S + T methods.

Sb	CSSP	CSSSP	DS + S + T
aa	14.6 ± 6.2	*11.6 ± 6.3*	**9.5 ± 2.1**
al	*2.3 ± 3.0*	**2.1 ± 2.7**	2.7 ± 3.1
av	32.6 ± 7.6	*31.8 ± 7.7*	**21.5 ± 7.6**
aw	**3.5 ± 3.3**	6.5 ± 4.3	6.5 ± 4.5
ay	**6.0 ± 3.9**	10.5 ± 5.7	*8.5 ± 5.0*

It can be seen that the three methods are quite competitive. Each method has its best performance in certain subjects. Notably, our method does the best in subject "av," outperforming the other two methods by about 10%. As mentioned earlier, the most responsive frequency range of "av" shifts above the normal α and β bands (8–26 Hz). This seems to suggest that for such BCI "abnormal," the DS features may be a better choice for the classification task.

5. DISCUSSION

5.1. *Relation to other filter learning methods*

In Section 4, a bell-shaped spatial filter is applied as a preprocessing for the DS features. Equivalently, this prefilter can be viewed as a weighting on the electrodes. Spatially, it resembles a Mexico hat, a positive peak surrounded by a ring of negative peaks (as illustrated in Figure 5(a)).

Our filter learning method further optimizes this prefilter by modifying its shape in the frequency domain (e.g., Figure 3(a)). After the optimization, the spatial influence of the resulting filter remains similar to the prefilter (Figure 5(b)). However, the separation between the positive and the negative peaks of the learned filter increases. This allows signals of lower spatial frequency to pass. Such adaptation helps the filter to extract more discriminative signals from the EEG signals.

An interesting observation is that the spatial filters obtained from the CSP method locally resemble the prefilter we
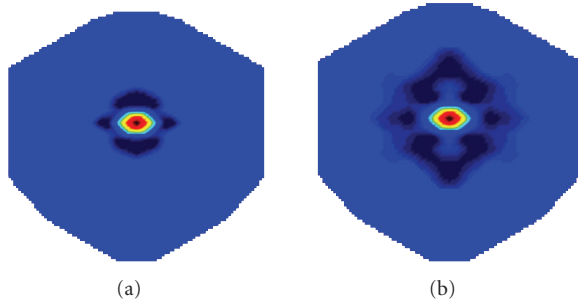
(a) (b)

FIGURE 5: (a) A bell-shaped spatial filter; (b) a learned spatial filter.

applied for the DS features. As shown in the middle column of Table 1, the filters learned by the CSP method emphasize the electrode directly above the neuronal cluster in charge of the imagination; at the same time, they suppress the contribution from adjacent electrodes.

While our filter learning method employs the prefilter as a prior knowledge and successively refines this knowledge locally, the CSP method arrives at similar results by computing a global filter instead. In the cases where this prior knowledge is accurate, we expect that better filters can be obtained by our method, which eventually leads to lower classification error (e.g., the classification error for subject "av" in Table 2).

5.2. Higher-order dynamical system features

In this paper, the covariance is used as a measure of dependence between different regions of the brain. Covariance, however, can only detect second-order dependence between the signals. Other more-powerful measures are needed if one wants to exploit higher-order dynamical system (DS) features of the brain.

Various measures have been explored in the literature. For instance, phase synchronization has also been employed as DS features for classifying BCI signals [9, 10]. Another example is the mutual information [13], but its use in BCI context remains unexplored. In all these cases, however, it is not yet clear how spatial and temporal filters can be learned automatically from the data.

6. CONCLUSION

In this paper, we exploited the collective dynamics of the cortex as features for BCI. We also proposed a framework for learning the optimal spatial and temporal filters during the extraction of these features. For 4 of the 5 subjects tested, our automated approach reduces classification errors to less than 10%. This performance is comparable to that of the CSP features obtained with manually tuned parameters. Further comparisons with other filter learning approaches also show the competitive performance of our method. Our results suggest that the dynamical system features combined with filter learning approach are very promising for BCI. More investigation is needed to fully demonstrate its advantage.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842–1857, 1999.

[3] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and Clinical Neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[4] 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 Information Processing Systems*, vol. 18, pp. 315–322, MIT Press, Cambridge, Mass, USA, 2006.

[5] K.-R. Müller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, 2003.

[6] T. N. Lal, M. Schröder, N. J. Hill, et al., "A brain computer interface with online feedback based on magnetoencephalography," in *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, pp. 465–472, Bonn, Germany, August 2005.

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

[8] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 993–1002, 2004.

[9] E. Gysels and P. Celka, "Phase synchronization for the recognition of mental tasks in a brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 4, pp. 406–415, 2004.

[10] L. Song, E. Gysels, and E. Gordon, "Phase synchrony rate for the recognition of motor imagery in brain-computer interface," in *Advances in Neural Information Processing Systems*, vol. 18, pp. 1265–1272, MIT Press, Cambridge, Mass, USA, 2006.

[11] L. Song and J. Epps, "Improving separability of EEG signals during motor imagery with an efficient circular Laplacian," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 2, pp. 1048–1051, Toulouse, France, May 2006.

[12] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[13] A. Pikovsky, M. Rosenblum, J. Kurths, and B. Chirikov, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2003.

[14] R. Srinivasan, P. L. Nunez, and R. B. Silberstein, "Spatial filtering and neocortical dynamics: estimates of EEG coherence," *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 7, pp. 814–826, 1998.

[15] X. Gao, Z. Zhang, B. Hong, and S. Gao, http://ida.first.fraunhofer.de/projects/bci/competition_iii/results/berlin_IVa/YijunWang_desc.pdf

[16] L. Song and J. Epps, "Classifying EEG for brain-computer interfaces: learning optimal filters for dynamical system features," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 857–864, Pittsburgh, Pa, USA, June 2006.