Adaptive Spatio-Temporal Filtering for Movement Related Potentials in EEG-Based Brain–Computer Interfaces

Jun Lu, Kan Xie, and Dennis J. McFarland

***Abstract*—Movement related potentials (MRPs) are used as features in many brain-computer interfaces (BCIs) based on electroencephalogram (EEG). MRP feature extraction is challenging since EEG is noisy and varies between subjects. Previous studies used spatial and spatio-temporal filtering methods to deal with these problems. However, they did not optimize temporal information or may have been susceptible to overfitting when training data are limited and the feature space is of high dimension. Furthermore, most of these studies manually select data windows and low-pass frequencies. We propose an adaptive spatio-temporal (AST) filtering method to model MRPs more accurately in lower dimensional space. AST automatically optimizes all parameters by employing a Gaussian kernel to construct a low-pass time-frequency filter and a linear ridge regression (LRR) algorithm to compute a spatial filter. Optimal parameters are simultaneously sought by minimizing leave-one-out cross-validation error through gradient descent. Using four BCI datasets from 12 individuals, we compare the performances of AST filter to two popular methods: the discriminant spatial pattern filter and regularized spatio-temporal filter. The results demonstrate that our AST filter can make more accurate predictions and is computationally feasible.**

***Index Terms*—Brain–computer interfaces (BCIs), electroencephalogram (EEG), movement related potentials (MRPs).**

I. Introduction

THE AIM of brain–computer interface (BCI) research is to build new communication channels that directly translate brain signals into control commands for output devices such as computers or neuroprosthesis [1]. Among different techniques for the noninvasive measurement of the human brain, the electroencephalogram (EEG) is commercially affordable and has excellent temporal resolution that enables BCIs capable of real time interactions [2]. Over the past decade, various EEG-based BCIs have been developed to help people who have damage in their peripheral pathways to recover their communication abilities [3]–[5]. So far, a lot of these BCIs are based on event-related desyncronization (ERD) [6]–[8]

Generally, these BCI systems use either phase-locked or oscillatory features [9]. Phase locked features, such as the movement related potential (MRP) [10], the error potential [11] and the evoked response P300 [12], use the signal amplitude directly for prediction. Oscillatory features, such as event-related desyncronization (ERD) [6]–[8], use the power or coherence of a signal in particular frequency bands for prediction. Here, we focus on MRP feature extraction and prediction in EEG-based BCIs, which is a challenging problem for signal processing and machine learning, because the MRPs of EEG are subject specific and buried in a large amount of noise such as task-unrelated neural activities (e.g., the visual α rhythm) and extra-neural artifacts (e.g., muscle activity and eye blinks). So far, some well-known feature extraction methods have been applied to the predictor to enhance the signal-to-noise ratio (SNR) of phase locked components. These methods can be unsupervised, such as principal component analysis [13], independent component analysis [14]–[17] and synchronous responses projection [18], or supervised, such as discriminative spatial pattern (DSP) filtering [19] and its variant local DSP filtering [20]. After feature extraction, a proper prediction algorithm is required. Researchers have considered a number of approaches to this prediction problem, including linear methods [21], nonlinear methods [22], neural networks [23] and a combination of classifiers [24]. Most current BCI designs pair highly complex feature extraction with a relatively simple linear classifier [9] since there is no clear evidence that one method is best for the prediction of phase locked components. This arrangement is probably due to a preference for simplicity and the belief that linear classification would be sufficient after adequate feature extraction [25], [26]. Furthermore, it has been demonstrated that if both the feature extraction and the classifier are linear, instead of seeking them separately, we can equivalently learn a general projection directly (i.e., a spatio-temporal filter). This unified discriminative approach might provide a better overall performance [2], [9]. But this technique needs to estimate the covariance matrix accurately in high dimensional space. This is a difficult problem when the size of the training set is limited, as in many BCI applications. To resolve this problem, regularized frameworks with empirical risk minimization have been proposed to control the complexity of spatio-temporal filter and

Manuscript received May 08, 2013; revised December 13, 2013 and March 10, 2014; accepted March 30, 2014. Date of publication April 07, 2014; date of current version July 03, 2014. This work was supported in part by National Natural Science Foundation of China under Grant 61304140, Grant 61273192, Grant 61333013, and Grant 61271210.

J. Lu and K. Xie are with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: lujun.rylj@gmail.com; kanxiegdut@gmail.com).

D. J. McFarland is with Laboratory of Neural Injury and Repair, Wadsworth Center, New York State Department of Health, Albany, NY 12201 USA (e-mail: mcfarlan@wadsworth.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNSRE.2014.2315717

avoid overfitting [2], [9], [27]. In those studies, the EEG signals from manually selected time intervals are averaged [27] or low pass filtered [2] as a simple preprocessing to reduce the feature dimension. This preprocessing is equivalent to a low-pass time-frequency filter. Ideally, the time-frequency filter and subsequent regularized linear prediction model should be optimized by the same criteria (e.g., maximizing the discriminant ability or minimizing the empirical error) and directly driven by the raw EEG data [28].

In this paper, we propose an adaptive spatio-temporal (AST) filtering approach for MRP prediction. Since MRPs behave as slowly changing potentials in specific time intervals, we use a Gaussian kernel to construct a low-pass time-frequency filter that has only two parameters, the kernel center and radius. We compute the spatial filter with the linear ridge regression (LRR) algorithm [29] which can address not only the classification problem but also regression problems.[1] As an adaptive approach, for different EEG datasets, the optimal parameters of the spatiotemporal filter, including the center and radius of Gaussian kernel and the regularization coefficient of LRR, are automatically estimated by minimizing the error of leave-one-out (LOO) cross validation (CV). CV is a popular approach to model selection [32]. The n-fold CV splits the data into n parts, and uses each alternatively to train and to validate the model. The final performance is the mean of the performances on the n different validation sets. CV maximizes the total number of validation trials and potentially helps to protect against overfitting. When n equals the number of trials, n-fold CV is the LOO CV. To accelerate this package optimization procedure, we design a gradient descent method that employs the closed form of LRR LOO error [33] and explicitly compute the derivatives about all the parameters. This technique is inspired by the study of Bo *et al.* [34] which was originally proposed to adjust the feature scaling for kernel Fisher discriminate analysis (FC-KFDA).

Our novel contributions in this paper are summarized as follows. First, the AST filtering automatically optimizes spatio- and temporal-filters in conjunction for MRP to adapt to variations between subjects. This avoids significant human intervention in parameter settings which occurs in the popular algorithms DSP filtering [19] and regularized spatio-temporal (RST) filtering [27]. Second, AST employs a Gaussian kernel to model the slow shift of the MRP and thus simplifies the temporal filtering model. It is a kind of parameter regularization with prior information of neurophysiology (i.e., that the MRP is a low frequency signal). The complexity of this temporal filter is much lower compared to a filter based on weighting all of the samples in time. Third, the AST filtering uses cross-validation and is computationally efficient due to the technique of analytically computing the derivative of parameters based on the close form of LOO error. Although the AST is somewhat similar in spirit to FC-KFDA (i.e., the idea of unifying the feature extraction and prediction as a two layer learning system and the technique of optimization), their mechanisms are much different. AST focuses on filtering the EEG matrix of each trial and the Gaussian kernel is used to perform a linear smoothing while FC-KFDA adjusts the distance between the feature vectors of each pair of samples and the Gaussian kernel is used as a nonlinear distance metric.

[1]With the development of BCI, regression is preferable to classification because it is better suited to controlling continuous movements of cursor or neuroprosthesis in real time and it generalizes more readily to novel target locations [30], [31]. For the moment, this paper focus on the binary classification problem of MRPs.

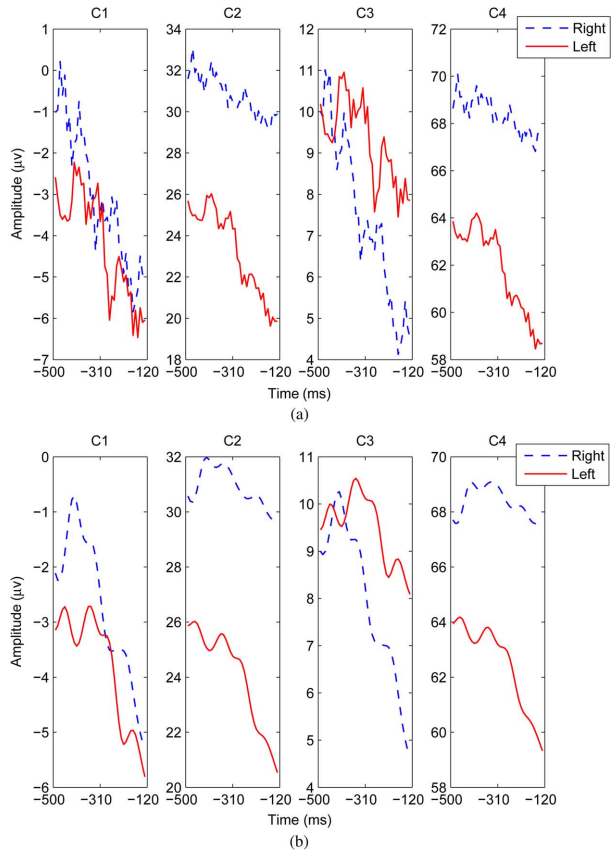


Fig. 1. Averaged potentials in channels C1, C2, C3, and C4 for left (red solid line) and right (blue dash line) finger movements. (a) Before low-pass filtering. (b) After low-pass filtering.

The rest of the paper is organized as follows. Section II introduces the neurophysiological background of MRP. Section III briefly describes the discriminant spatial pattern (DSP) [19] filter as well as the regularized spatio-temporal (RST) filter [27], and then proposes the AST filter. Section IV describes the data used for comparing these methods, the evaluation procedure, the results and discussion. Finally, Section V presents the conclusion.

II. Neurophysiological Background

The MRP is a slow negative shift that precedes voluntary movement. This negative shift starts to be more pronounced about 500 ms before the onset of the movement [35]. For finger and hand movements, the MRP is characterized by a contralateral dominance, which corresponds to left or right sensorimotor cortex [36]. Fig. 1(a) shows the EEG potential averaged over all trials from channels C1, C2, C3, and C4, between -500 ms to -120 ms before the keystroke of the finger movement experiment of Dataset I, BCI competition I [37], [38]. The increasingly negative waveforms of the channels over motor cortex illustrate the decreasing nature of MRPs. At channel C4 (over right

motor cortex) the potentials associated with left finger movement decrease more rapidly than that at channel C3 (over left motor cortex) and vice versa. This contra-lateral dominance is more apparent in some channels such as C4 and C2; is more visible if the potentials are low-pass filtered (0–7 Hz), as shown in Fig. 1(b) and is more obvious between -310 ms and -120 ms, which is closer to the actual keystroke than at -500 ms. Thus, selection of the parameters characterizing the spatial, frequency and time characteristics of the signal are important to MRP feature extraction. These parameters should be adaptively selected since the MRPs are subject specific.

III. Methods

A. Discriminative Spatial Pattern (DSP) Filter

Liao *et al.* [19] proposed the DSP filtering algorithm to extract the MRP features. The goal of DSP filtering is to transform the raw EEG matrix of each trial into temporal sequences such that within-class variance is minimized and between-class separation is maximized. As a spatial filtering method, DSP cannot optimize temporal information. An approximation is to low-pass filter the EEG data first and use sliding time window. For this approach, the cut-off frequency of the signal and the width of the sliding window are set manually. The start point of sliding time window and the number of spatial filters are alternatively chosen with a CV on the corresponding training data. This alternative parameter selection does not guarantee a local optimal solution. The DSP algorithm is formulated as follows.

Let $\mathbf{X}_j^i \in \mathbb{R}^{C\times T}$ denote the low-passed EEG signals of trial i from class j, C is the number of channels and T is the number of samples in time. Thus, the within-class variance matrix $\mathbf{S}_w$ and the between-class variance matrix $\mathbf{S}_b$ can be represented as

$$\mathbf{S}_w = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\left(\mathbf{X}_j^i - \mathbf{M}_j\right)\left(\mathbf{X}_j^i - \mathbf{M}_j\right)^{\mathrm{T}}$$

$$\mathbf{S}_b = \sum_{j=1}^{K} n_j(\mathbf{M}_j - \mathbf{M})(\mathbf{M}_j - \mathbf{M})^{\mathrm{T}}$$

where $\mathbf{M}_j = (1/n_j)\sum_{i=1}^{n_j}\mathbf{X}_j^i$ is the center of class j, $\mathbf{M} = (1/n)\sum_{h=1}^{n}\mathbf{X}^h$ is the center of all the training trials, K is the number of different classes (here $K=2$), n_j is the number of trials belongs to class j and n is total number of all the trials. If we assume $\mathbf{w}\in\mathbb{R}^C$ is the spatial filter, the MRP component extracted by DSP can be represented as

$$\mathbf{v} = \mathbf{w}^{\mathrm{T}}\mathbf{X}. \tag{1}$$

Therefore, the projected variances in feature space can be calculated as

$$\mathbf{S}'_w = \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w}$$
$$\mathbf{S}'_b = \mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}.$$

The objective is to maximize the ratio of $\mathbf{S}'_b$ and $\mathbf{S}'_w$, i.e., $\max_{\mathbf{w}} J = \mathbf{S}'_b/\mathbf{S}'_w$, where the optimal spatial filter $\mathbf{w}^*$ can be found by solving the generalized eigenvalue problem $(\mathbf{S}_w^{-1}\mathbf{S}_b)\mathbf{w} = \beta\mathbf{w}$. To improve generalization of the model based on estimates with noisy EEG signals, one may use the regularization as $\mathbf{S}_w \leftarrow \mathbf{S}_w + \lambda\mathbf{I}$ to get a relative robust solution. Then, the average amplitude is extracted as the feature

$$z = \mathbf{w}^{*\mathrm{T}}(\mathbf{X} - \mathbf{M})\mathbf{1} \tag{2}$$

where $\mathbf{1}$ is a column vector of T ones. Finally, these features are used by a classifier such as Fisher discriminant analysis (FDA) or a support vector machine (SVM) to predict the trial labels.

B. Regularized Spatio-Temporal (RST) Filter

In order to improve the overall performance of feature extraction and classification, various spatio-temporal filters for classifying EEG with MRP and P300 features have been presented. These filters are build on the frameworks of regularized FDA [27], [37], SVM [37], or regularized linear logistic regression [2], [9]. Most of them perform low-pass filtering or average the potentials of each time window to reduce the feature dimension, then estimate the projection of the feature vector. These methods ignore the spatial-temporal structure [24], [27], [37]. The RST filter based on FDA is formulated as follows [27], [37].

Let $\mathbf{X}_j^i \in \mathbb{R}^{C\times T}$ denotes the EEG signals of trial i from class j after the low-pass filtering or the averaging across time intervals. Then $\mathbf{X}_j^i$ is concatenated into features vector $\mathbf{x}_j^i \in \mathbb{R}^{C\cdot T}$. The goal of the RST algorithm is to seek the spatio-temporal filter $\mathbf{u}\in\mathbb{R}^{C\cdot T}$ that compress these feature vectors into one dimension where the within-class variance is minimized and the between-class separation is maximized. Similar to the DSP filter, the optimal spatial filter $\mathbf{u}^*$ can be found by solving the generalized eigenvalue problem $(\mathbf{S}_w^{-1}\mathbf{S}_b)\mathbf{u} = \beta\mathbf{u}$. The within-class scatter matrix $\mathbf{S}_w$ and the between-class scatter matrix $\mathbf{S}_b$ are computed as

$$\mathbf{S}_w = \sum_{j=1}^{K}\sum_{i=1}^{n_j}\left(\mathbf{x}_j^i - \mathbf{m}_j\right)\left(\mathbf{x}_j^i - \mathbf{m}_j\right)^{\mathrm{T}}$$

$$\mathbf{S}_b = \sum_{j=1}^{K} n_j(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^{\mathrm{T}}$$

where $\mathbf{m}_j = (1/n_j)\sum_{i=1}^{n_j}\mathbf{x}_j^i$ is the center of class j, $\mathbf{m} = (1/n)\sum_{h=1}^{n}\mathbf{x}^h$ is the center of all the training trials. Then the class label of EEG signal $\mathbf{X}_j^i$ can be predicted by $\hat{y} = \mathbf{sign}(\mathbf{u}^{*\mathrm{T}}\mathbf{x} - b)$, where $b = \mathbf{u}^{\mathrm{T}}\mathbf{m}$ is the bias. To improve generalization of the model based on estimates with noisy EEG signals, one may used the regularization as $\mathbf{S}_w \leftarrow \mathbf{S}_w + \lambda\mathbf{I}$. The regularization coefficient λ is determined by the CV in an *ad hoc* manner. Note that FDA is equivalent to least square regression for a binary classification problem [39]. The optimization of this RST filter can be also expressed as

$$\min_{\mathbf{u}} \frac{1}{2}\left(\|\mathbf{X}^{\mathrm{T}}\mathbf{u} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{u}\|_2^2\right)$$

where $\mathbf{X} = [\mathbf{x}^1, \ldots \mathbf{x}^h, \ldots, \mathbf{x}^n] \in \mathbb{R}^{(C\cdot T)\times n}$ and label vector $\mathbf{y}\in\mathbb{R}^n$ have been centered, the l_2-norm of spatio-temporal filter are constrained. Other sophisticated regularization techniques with the l_1-norm and trace-norm have also been discussed and compared in the literatures [2], [9], [37]. However, the classification accuracies of these regularization techniques

are similar for presented data sets. For simplicity, we use the l_2-norm regularization in the following with emphasis on the optimization and processing of the spatio-temporal filter.

C. Adaptive Spatio-Temporal (AST) Filter

In order to further improve the overall performance of MRP prediction, AST filter integrates the optimizations of preprocessing (i.e., the low-pass filtering and time interval selection), feature extraction, as well as classification. Specially, AST filter captures the temporal structure of MRPs with a Gaussian kernel that smoothly models the instantaneous voltage of the EEG signal with its temporal neighbors. This Gaussian kernel constructs a low-pass filter centering at an optimal time point. Next, a spatial filter is built on the framework of the LRR algorithm. Unlike DSP filter, AST filter simultaneously adjusts the parameters of the temporal filter and the spatial filter under the unified framework of minimizing the LOO error of the training set. Unlike RST filter, AST filter is directly driven by the raw EEG data without the imprecise and time-consuming selection of time interval, cut-off frequency and regularization coefficient.

First, the temporal filtering model based on a Gaussian kernel is formulated as

$$(\mathbf{X})_{ik} = \sum_{j=1}^{T} \frac{1}{Z} \omega_j(\tau, \theta) \mathbf{v}_{ik}(j) \tag{3}$$

$$\omega_j(\tau, \theta) = \exp\left[-\theta(\tau - j)^2\right]$$
$$s.t.\ \theta > 0 \quad \text{and} \quad T \geq \tau \geq 1 \tag{4}$$

where $(\mathbf{X})_{ik}$ is the MRP feature extracted from channel i at trial k; $Z = \sum_{j=1}^{T} \omega_j(\tau, \theta)$ is the normalization item; $\mathbf{v}_{ik}(j)$ is the voltage of brain signal at time point j. The kernel parameters τ and θ control the center and the radius of the filter. Given τ, if θ is small, the value of $\omega_j(\tau, \theta)$ is less sensitive to the distance $(\tau - j)^2$, thus the AST filtering tends to average the amplitudes of all the sample points in time. Conversely, if θ is large, the AST filter tends to focus more on samples close to the filter center τ.

This temporal filtering model based on a Gaussian kernel actually performs a low-pass filtering at moment τ. Given a EEG signal $x(t)$ and a Gaussian function $g(t) = \exp(-\theta t^2)$, then the filtered signal $\tilde{x}(t)$ is obtained by the convolution operation as $\tilde{x}(t) = g(t) * x(t)$. Because the Fourier transform of $g(t)$ is $\hat{g}(f) = \sqrt{\pi/\theta} \exp(-(\pi^2 f^2/\theta))$ [40], where f is the frequency, thus $g(t) * x(t)$ is a low-pass filtering and θ controls the band width. Since $\tilde{x}(t) = g(t) * x(t) = \sum_{j=1}^{T} g(t - j)x(j)$, when $t = \tau$

$$\tilde{x}(\tau) = \sum_{j=1}^{T} \exp\left[-\theta(\tau - j)^2\right] x(j). \tag{5}$$

Equation (5) indicates that, at moment τ, the low-pass filtered EEG signal can be expressed as the inner product of the original EEG signal and the Gaussian kernel. By adding the normalization item, (5) becomes (3). As a result, the temporal filtering model depicts how slow the MRP shifts and when the MRP occurs. We can adapt θ and τ for the datasets recorded from different subjects or experiments.

Secondly, the spatial filtering (prediction) model is built on the framework of LRR. LRR minimizes the penalized sum-of-squares error function by incorporating the l_2-norm regularization to control the model complexity and improve the generalization performance. Here, the spatial filtering model is formulated as

$$\min_{\boldsymbol{\alpha}, \theta} \frac{1}{2} \left(\|\tilde{\mathbf{X}}^{\mathrm{T}} \boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_2^2 \right) \tag{6}$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{C \times n}$ is the centered feature matrix, i.e.,

$$(\tilde{\mathbf{X}})_{i,k} = (\mathbf{X})_{i,k} - \frac{1}{n} \sum_{k=1}^{n} (\mathbf{X})_{i,k} \tag{7}$$

n is the number of trials; $\boldsymbol{\alpha} \in \mathbb{R}^C$ is the linear projection of LRR and the weight of spatial filter for all the MRP features, $\mathbf{y} \in \mathbb{R}^n$ is the centered label vector and $\lambda \geq 0$ is the regularization coefficient controlling the bias-variance trade-off. Given τ, θ and λ, the solution is

$$\boldsymbol{\alpha} = \left(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \lambda \mathbf{I}\right)^{-1} \tilde{\mathbf{X}}\mathbf{y} \tag{8}$$

where $\mathbf{I}$ denotes the $C \times C$ identity matrix.

Note that (3) and (6) indicate that τ, θ and λ have essential roles in the AST filtering model and should be derived from the EEG training data. The LOO error is considered to be an almost unbiased estimator of the generalization error [41] and has been used frequently in model selection [42]–[44]. The development of closed-form solutions for performing LOO CV in certain learning algorithms such as LRR and KFDA [45] significantly reduces the computational complexities. Thus, the present study uses the LOO error as the optimization criterion in tuning the parameters of AST filtering. Specifically, we explicitly compute the derivatives of LOO error with respect to τ, θ and λ, and then estimate them by a gradient descent method. This procedure is inspired by the studies of feature scaling in kernel learning algorithms [34], [46]. The difference is that these authors focus on adjusting the kernel matrices, whereas we extend the method to control the spatio-temporal filtering model. In accord with [33], the closed form of LOO error for AST filtering is

$$\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y} \odot \mathbf{e} \tag{9}$$

where $\mathbf{r} \in \mathbb{R}^n$ is the residual error vector, the hat matrix $\mathbf{H} = \tilde{\mathbf{X}}^{\mathrm{T}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \lambda \mathbf{I})^{-1}\tilde{\mathbf{X}}$, $\odot$ denotes an element-wise division, $\mathbf{e} = \mathbf{1} - diag(\mathbf{H})$, $\mathrm{diag}(\mathbf{H})$ denotes the diagonal elements of $\mathbf{H}$, $\mathbf{1}$ is a column vector of N ones. The aim of AST filtering is to minimize the LOO error

$$\min_{\tau, \theta, \lambda} J = \frac{1}{2n} \|\mathbf{r}\|_2^2. \tag{10}$$

Then, according to the chain rule, the derivative of J with respect to θ can be expressed as

$$\frac{\partial J}{\partial \theta} = \frac{1}{n} \mathbf{r}^{\mathrm{T}} \frac{\partial \mathbf{r}}{\partial \theta}. \tag{11}$$

And based on (9), the derivative of $\mathbf{r}$ with respect to θ is given by [34]

$$\frac{\partial \boldsymbol{r}}{\partial \theta} = (\mathbf{I} - \mathbf{H})\mathbf{y} \odot \mathbf{e} \odot \mathbf{e} \otimes diag\left(\frac{\partial \mathbf{H}}{\partial \theta}\right) - \frac{\partial \mathbf{H}}{\partial \theta}\mathbf{y} \odot \mathbf{e} \tag{12}$$

where $\otimes$ denotes an element-wise product. Let $\mathbf{G} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}} + \lambda\mathbf{I}$, then the derivative of $\mathbf{H}$ with respect to θ is computed as

$$\frac{\partial \mathbf{H}}{\partial \theta} = \frac{\partial \tilde{\mathbf{X}}^{\mathrm{T}}}{\partial \theta}\mathbf{G}^{-1}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}^{\mathrm{T}}\frac{\partial \mathbf{G}^{-1}}{\partial \theta}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{G}^{-1}\frac{\partial \tilde{\mathbf{X}}}{\partial \theta}. \quad (13)$$

Based on [47], the derivative of $\mathbf{G}^{-1}$ with respect to θ is

$$\begin{aligned}\frac{\partial \mathbf{G}^{-1}}{\partial \theta} &= -\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \theta}\mathbf{G}^{-1} \\ &= -\mathbf{G}^{-1}\left(\frac{\partial \tilde{\mathbf{X}}}{\partial \theta}\tilde{\mathbf{X}}^{\mathrm{T}} + \tilde{\mathbf{X}}\frac{\partial \tilde{\mathbf{X}}^{\mathrm{T}}}{\partial \theta}\right)\mathbf{G}^{-1}. \quad (14)\end{aligned}$$

Based on (7), the derivative of $\tilde{\mathbf{X}}^{\mathrm{T}}$ with respect to θ is

$$\frac{\partial(\tilde{\mathbf{X}})_{i,k}}{\partial \theta} = \frac{\partial(\mathbf{X})_{i,k}}{\partial \theta} - \frac{1}{n}\sum_{k=1}^{n}\frac{\partial(\mathbf{X})_{i,k}}{\partial \theta}. \quad (15)$$

According to temporal filtering model (3) and (4), the derivative of $(\mathbf{X})_{i,k}$ with respect to θ is

$$\frac{\partial(\mathbf{X})_{i,k}}{\partial \theta} = \sum_{j=1}^{T}\frac{\partial\left(\frac{\omega_j}{Z}\right)}{\partial \theta}\mathbf{v}_{ik}(j) \quad (16)$$

where

$$\frac{\partial\left(\frac{\omega_j}{Z}\right)}{\partial \theta} = (Z^{-1})\frac{\partial \omega_j}{\partial \theta} - \omega_j(Z^{-2})\left(\sum_{j=1}^{T}\frac{\partial \omega_j}{\partial \theta}\right). \quad (17)$$

Furthermore, according to the kernel function (4)

$$\frac{\partial \omega_j}{\partial \theta} = -\omega_j(\tau - j)^2. \quad (18)$$

Combining (11)–(18) yields the derivative of the LOO error with respect to θ. At same time, according to (4), the derivative of ω_j with respect to τ is

$$\frac{\partial \omega_j}{\partial \tau} = 2\omega_j\theta(j - \tau). \quad (19)$$

The derivative of $\mathbf{H}$ with respect to λ is

$$\frac{\partial \mathbf{H}}{\partial \lambda} = -\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{G}^{-1}\mathbf{G}^{-1}\tilde{\mathbf{X}}. \quad (20)$$

Likewise, the derivative of J with respect to τ and λ can be computed in a similar way.

Since both θ and λ are positive, the parameterizations $(\theta, \lambda) = (\exp(\xi), \exp(\psi))$ are used to avoid these constraints, and then $\partial\theta/\partial\xi = \theta$, $\partial\lambda/\partial\psi = \lambda$. In addition $T \geq \tau \geq 1$, a sigmoid function is employed to parameterize τ as: $\tau = tanh(\gamma)(T-1)/2 + (T+1)/2$, and then $\partial\tau/\partial\gamma = sech^2(\gamma)(T-1)/2$. The gradient decent method is implemented with the Quasi-Newton BFGS algorithm. The step size is determined by the cubic polynomial line search procedure. Table I lists the pseudo-code of the AST filtering algorithm. The computational complexity of AST is estimated as

$$L \times O(T \times C \times N + C^3) + O(C^3) \quad (21)$$

TABLE I
AST FILTERING ALGORITHM

(1) **Initialization**: given multichannel EEG dataset $\{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$, set initial parameters τ_0, θ_0 and λ_0, iteration number L and stop criterion δ.
(2) **for** $l = 1 : L$
  (3) Filter the brain signal for each channel with respect to the dataset $\{(\mathbf{X}_i, y_i)\}_i^n$, kernel center τ_{l-1}, kernel radius θ_{l-1} by Eq.(3), (4);
  (4) Get the feature matrix $\tilde{\mathbf{X}}$ by Eq.(7);
  (5) Calculate the LOO error $J(\tau_{l-1}, \theta_{l-1}, \lambda_{l-1})$ by Eq.(9), (10);
  (6) Calculate $\frac{\partial J}{\partial \tau_{l-1}}$, $\frac{\partial J}{\partial \theta_{l-1}}$ and $\frac{\partial J}{\partial \lambda_{l-1}}$ by Eq.(11)−(20);
  (7) Update $(\tau_{l-1}, \theta_{l-1}, \lambda_{l-1})$ to $(\tau_l, \theta_l, \lambda_l)$;
  (8) Calculate $J(\tau_l, \theta_l, \lambda_l)$ by Eq.(9), (10);
  (9) If $|J(\tau_l, \theta_l, \lambda_l) - J(\tau_{l-1}, \theta_{l-1}, \lambda_{l-1})| < \delta$, **break**, **end**;
(10) **end**;
(11) Calculate $\boldsymbol{\alpha}$ by Eq.(8) with the optimized τ, θ and λ, then output them.

where L is the iteration number, $O(T \times C \times N + C^3)$ comes from the computation of the temporal filtering, feature extraction and the derivative of J with respect to the parameters using all the time points, channels and trials, $O(C^3)$ in the first item is the computational complexity of inverting matrix $\mathbf{G}$, the second $O(C^3)$ is the computational complexity of calculating $\boldsymbol{\alpha}$ using optimized parameters.

IV. EVALUATION

A. EEG Datasets

MRP is a negative shift in the electric potential that appears over specific areas of sensorimotor cortex prior to the onset of voluntary movement, reaching a negative peak approximately 100 ms after movement onset. MRP can also be produced by imagined movement. Here, we use four datasets of EEG associated with finger movements and motor imageries from BCI competitions and our own lab to evaluate the proposed method and compare it with DSP filtering and RST filtering. In all cases the predictions are based on single-trial data.

1) Dataset I, BCI Competition I: The objective of Dataset I, from BCI competition I, was to predict the laterality of upcoming finger movements (left versus right hand) before a keystroke [37]. The EEG signals were measured from subject S1 with 27 channels at 1000 Hz using a band-pass filter from 0.05 to 200 Hz. Epochs 1500 ms long were cut out of the continuous raw signals, each ending at 120 ms before the respective keystroke. 513 trials (100 trials for test and 413 trials for training) were provided for the competition evaluation in all. We down sampled the data to 100 Hz to reduce computational burden.

2) Dataset IV, BCI Competition II: The objective of Data set IV, from BCI competition II, was to predict the laterality of upcoming finger movements (left versus right hand) before a key press [37]. The EEG signals were measured from subject S2 with 28 channels at 1000 Hz using a band-pass filter from 0.05 to 200 Hz. Epochs 500 ms long were cut out of the continuous raw signals, each ending at 130 ms before the respective key press. 416 trials (100 trials for test and 316 trials for training) were provided for the competition evaluation in all. We down sampled the data to 100 Hz to reduce computational burden. In addition, we changed the class labels from $\{0, 1\}$ to $\{-1, 1\}$ to better compare prediction performance with the other data sets in this study.

3) Dataset IVa, BCI Competition III: The goal of Data set IVa, from BCI competition III, was to classify right hand and

foot motor imageries [48]. The EEG signals were recorded in five subjects (S3–S7) with 118 channels at 1000 Hz using a band-pass filter from 0.05 to 200 Hz. For each trial, a visual cue lasting 3.5 s indicated that the subjects should perform motor imageries. A total of 280 trials were available for each subject, and the training set sizes were 168, 224, 84, 56, and 28 for subjects S3–S7, respectively. The test sets consisted of the remaining trials. We down sampled the data to 100 Hz to reduce computational burden. In addition, we changed the class labels from $\{1, 2\}$ to $\{-1, 1\}$ to better compare prediction performance with the other data sets in this study.

4) Dataset of Cursor Movement Control: Our own data set contains EEG signals from five subjects (S8–S12) who performed motor imageries of left hand and right hand (or a hand and a foot) to control vertical cursor movement toward one of two targets located at different heights along the right edge of the video screen. The EEG was recorded at 160 Hz from 64 channels covering the whole scalp followed the international 10/20 system. A training set and a test set were available for each subject. Each set contained 50–60 trials for each target.

B. Preprocessing

1) Channel Selection: To minimize artifacts and reduce computation burden, we restrict our analysis to channels over sensorimotor cortex where MRPs are generated. For dataset I BCI competition I, dataset IV BCI competition II and the dataset of cursor movement control, we used channel FC1–FC4, C1–C6, CP1–CP4, FCZ, CZ, and CPZ; For dataset IVa BCI competition III, we used channel FC1–FC4, CFC1–CFC6, C1–C6, CCP1–CCP6, CP1–CP4, FCZ, CZ, and CPZ.

2) Selection of Frequency Band and Time Segment: For the proposed AST filtering: we used the normalized EEG data without selection of frequency band or time segment. The normalization removed the mean of the EEG data on a channel and trial-wise basis.

For DSP filtering and RST filtering: the data for each trial was low-pass filtered with a fifth-order Butterworth filter. The candidate frequency bands were {0–3 Hz, 0–5 Hz, 0–7 Hz, 0–10 Hz, 0–20 Hz}. The candidate time segments were 200 ms wide windows, with 190 ms overlap moving from the ending point to the start point of each trial for dataset I BCI competition I and dataset IV BCI competition II [19], with no overlap moving from the start point to the ending point of each trial for dataset IVa BCI competition III and the dataset of cursor movement control. The candidate regularization coefficients were $\{1, 10, 10^2, 10^3, 10^4, 10^5\}$.

DSP filtering performs feature extraction, the spatial filters with the largest eigenvalues are chosen, and then FDA is employed as the classifier. The optimal combination of frequency band, time segment, regularization coefficient and number of spatial filters were determined by five-fold CV.

RST filtering often needs to reduce the feature dimension. After frequency filtering and selection of a time segment, the means of consecutive five-tuple of data points for each channel were calculated as the features [37]. The optimal combination of frequency band, time segment and regularization coefficient were determined by five-fold CV.

TABLE II
PARAMETERS OF DSP AND RST FILTERING DETERMINED BY FIVE-FOLD CROSS VALIDATION ON THE TRAINING SETS FOR EACH SUBJECT RESPECTIVELY (FB: FREQUENCY BAND, TS: TIME SEGMENT, RC: REGULARIZATION COEFFICIENT, NU: NUMBER OF SPATIAL FILTERS ACCORDING TO THE LARGEST EIGENVALUES)

Subject	Method	FB	TS	RC	NU
S1	DSP	0−20Hz	-320−-120ms	10^4	7
	RST	0−20Hz	-320−-120ms	10^3	_
S2	DSP	0−20Hz	-330−-130ms	10^5	9
	RST	0−20Hz	-330−-130ms	10	_
S3	DSP	0−7Hz	1.0−1.2s	10^4	8
	RST	0−3Hz	1.2−1.4s	10^3	_
S4	DSP	0−10Hz	0−0.2s	10^2	2
	RST	0−3Hz	0.4−0.6s	10	_
S5	DSP	0−3Hz	1.4−1.6s	10^4	2
	RST	0−10Hz	1.0−1.2s	1	_
S6	DSP	0−3Hz	1.4−1.6s	1	8
	RST	0−5Hz	1.2−1.4s	10^5	_
S7	DSP	0−3Hz	2.6−2.8s	10	8
	RST	0−5Hz	1.4−1.6s	10^3	_
S8	DSP	0−20Hz	0−0.2s	10^5	6
	RST	0−3Hz	0.4−0.6s	10^3	_
S9	DSP	0−7Hz	0−0.2s	10^4	3
	RST	0−5Hz	0−0.2s	10^2	_
S10	DSP	0−5Hz	0.2−0.4s	10^4	5
	RST	0−3Hz	0.4−0.6s	10^2	_
S11	DSP	0−7Hz	0.4−0.6s	10^5	3
	RST	0−7Hz	0.6−0.8s	10^3	_
S12	DSP	0−3Hz	0−0.2s	10^2	5
	RST	0−7Hz	0−0.2s	1	_

Table II lists the results of parameter selection. As can be seen in Table II, for dataset I BCI competition I (subject S1) and dataset IV BCI competition II (subject S2), the time segments close to the onset of finger movements contained more discriminating information which is in accord with related works [19], [37]. For dataset IVa BCI competition III (S3–S7) and dataset of cursor movement control (S8–S12), the experiment required the subjects to perform the motor imageries within a period about 3.5 s when the target cues were presented, so the exact onset of subjects' motor imageries for each trial was not clear. Thus the selected time segments differ from subject to subject. Moreover, the frequency bands should be adapted since they varied between individuals.

C. Results and Discussion

1) Comparison of Classification Performance: Table III reports the classification accuracies obtained on the test sets. The results show that the AST filtering method produced the best results with these data, as it reached the greatest mean accuracy. A one-way ANOVA with repeated measures indicated that the five filtering methods differed significantly in classification accuracy ($\mathrm{d.f.} = 4$, $\mathrm{F} = 5.26$, $p = 1.2 \times 10^{-3}$). Additional two-group ANOVAs showed that the AST filtering was significantly better than each of the other methods in classification accuracy ($p \leq 3.74 \times 10^{-2}$). Without the selection of time segments and frequency band, DSP filtering cannot extract temporal information, RST filtering suffers from the overfitting of high dimension feature. Both of them classified the MRP data poorly (<66% on average). AST filtering integrates the signal processing with prediction and automatically estimates

TABLE III
CLASSIFICATION ACCURACIES (%) OBTAINED FOR EACH SUBJECT WITH DSP FILTERING, RST FILTERING AND AST FILTERING. DSP* AND RST* RESPECTIVELY DENOTE DSP FILTERING AND RST FILTERING WITHOUT THE SELECTION OF FREQUENCY BAND AND TIME SEGMENT. HIGHEST ACCURACIES FOR EACH SUBJECT ARE IN BOLD ITALICS

Sub.	DSP	RST	DSP*	RST*	AST
S1	96.00	***97.00***	63.00	92.00	***97.00***
S2	74.00	70.00	66.00	***79.00***	***79.00***
S3	81.25	***82.14***	51.79	60.71	79.46
S4	89.29	83.93	53.57	75.00	***94.64***
S5	66.33	60.20	56.12	52.04	***67.86***
S6	69.20	55.36	51.34	53.13	***79.46***
S7	51.19	50.79	***54.76***	52.38	53.57
S8	***73.08***	71.15	62.50	61.54	***73.08***
S9	73.08	70.19	53.85	72.12	***77.88***
S10	70.83	71.88	58.33	62.50	***78.13***
S11	66.35	70.19	58.65	63.46	***72.12***
S12	63.00	***69.00***	61.00	57.00	63.00
Mean	72.80	70.99	57.58	65.07	***76.27***
Std.	11.86	12.59	4.77	12.24	12.02

the optimal parameters in a continuous space by the gradient decent method, which is more effective and convenient. Note that, when performing motor imagery, not all the subjects showed MRP with strong discriminant ability, such as subjects S5, S7, and S12. All the presented filtering methods had poor classification accuracies for subjects S5, S7, and S12 ($<70\%$).

Furthermore, we compared the results of AST filtering with the results of the winners of the three public BCI competition datasets. The BCI winners reported that they combined the MRP and ERD feature for EEG classifications. Since MRP and ERD relate to different aspects of limb movements, combining them may improve the classification accuracy. Thus, in order to have a fair comparison, we also combined them. Specifically, the MRP feature was extracted by our proposed AST filtering and the ERD feature was extracted by common spatial pattern (CSP) filtering. The time segment for CSP filtering was from 0.5 to 2.5 s after the cue instructing the subject to perform motor imagery and the channels, frequency bands and the spatial filters for CSP filtering were selected by cross validation. FDA was used to combine the MRP feature and ERD feature. Note that, the training sets of subjects S6 and S7 are small. The winners of dataset IVa, BCI competition III used the ERD feature and semi-supervised approach to enlarge the training sets of subjects S6 and S7 with the unlabeled data. Thus, for S6 and S7, we also used the ERD feature and semi-supervised approach. Specifically, the semi-supervised graph-based method [49] was applied to classify the unlabeled trials based on ERD features, then the prediction confidence (PC) of each unlabeled trial was computed as: $PC_i = \max\{\exp(F_{i,c})/\sum_c \exp(F_{i,c})|c = 1, 2\}$ where $F_{i,c} \geq 0$ denotes the classification of trial i for class c [49]. After that, the unlabeled trials with $PC > 0.7$ were move into the training set. Based on the extended training set, the MRP feature was extracted by AST filtering. At last, the MRP feature and ERD feature were combined by FDA to classify the rest unlabeled trials with $PC \leq 0.7$. The classification accuracies obtained by our methods and the BCI winners are list in Table IV. It is shown that on average our results (94.19%) and the reported best results (93.40%) are very close ($p = 0.50$, two-group ANOVAs)

TABLE IV
CLASSIFICATION ACCURACIES (%) OBTAINED FOR EACH SUBJECT BY AST-CSP COMBINING METHOD AND BCI COMPETITION WINNERS

Sub.	Our results	Scores of BCI winners
S1	99.00	96.00
S2	89.00	84.00
S3	94.64	95.54
S4	100.00	100.00
S5	83.67	80.61
S6	97.77	100.00
S7	95.24	97.62
Mean	94.19	93.40
Std.	5.90	7.83

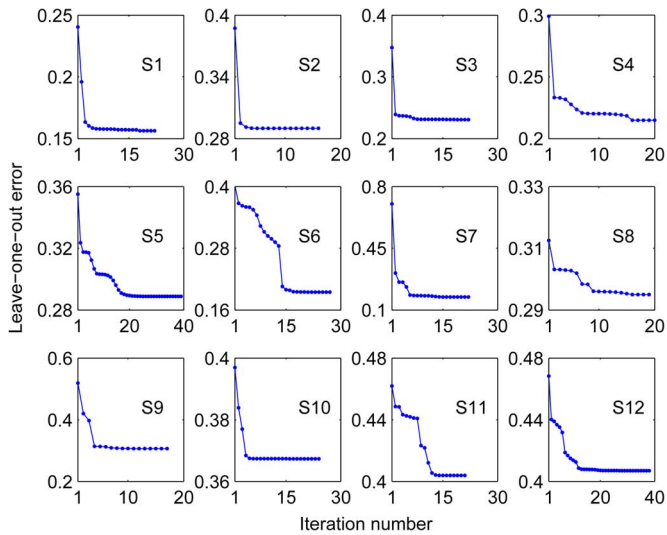


Fig. 2. Convergence of the gradient-based optimization procedure of the AST filtering algorithm. Graph shows LOO errors computed over the iterations. Each curve corresponds to one subject.

2) Convergence of AST Filtering: Fig. 2 displays the convergence of the gradient-based optimization procedure of AST filtering for each subject. Since the LOO error is nonconvex with respect to the parameters, multiple optima exist. In practice, we randomly initialized the parameters in following ranges: regularization coefficient $\lambda \in [1, \exp(4)]$, kernel radius factor $\theta \in [\exp(-8), \exp(-1)]$, and $\gamma \in [-4, 4]$, i.e., kernel center $\tau \in [1, T]$, then select the initial parameters which lead to the lowest LOO error. The iterations were stopped when the decrease of the LOO error was less than 10^{-6}. Fig. 2 shows that the LOO errors decreased markedly in the first 20 iterations, then rapidly converged to a final stable value. The computation of the AST algorithm was performed using MATLAB running on Windows 7 Professional SP1 64 bit with Intel Core i7-2640M CPU 2.8 GHz. Given a group of initial parameters, the time of AST optimization for each subject ranged between 0.48 and 1.81 s. Of these, the slowest was subject S1 with 17 channels and 413 training trials. Thus, the AST filtering model can be updated between runs or sessions (about 400 trials) within 2 s for the BCI systems based on MRP feature, if the number of channels is no more than 17.

3) The MRPs Extracted by AST Filtering: Fig. 3 shows the average over all trials of MRPs extracted by AST filtering algorithm for subjects S1 and S4. S1 and S4 are from datasets of finger movement and motor imagery, respectively.

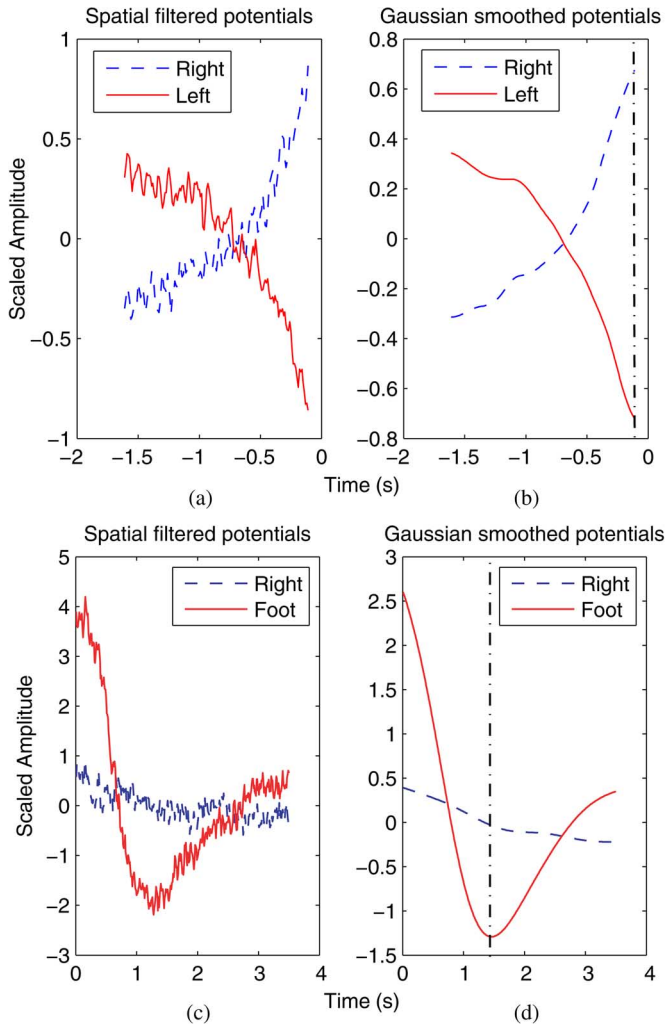


Fig. 3. Averaged potentials extracted by AST filtering algorithm according to left finger movement/foot motor imagery (red solid line) and right finger movements/right hand motor imagery (blue dash line). (a) and (b): Subject 1. (c) and (d): Subject 4.

Fig. 3(a) and (c) demonstrate the average potentials extracted by the spatial filters of AST filtering. These average potentials, corresponding to different mental tasks, shift gradually with different trends. The potentials extracted by the spatial filter are smoothed by Gaussian kernels. This is the AST temporal filtering. The average results are shown in Fig. 3(b) and (d) in which the trends of average potentials are more clear. The Gaussian kernel smoothing actually performs a low-pass filtering that can match the slow shift of MRPs by adjusting the radius parameter. The black dotted lines indicate the centers of Gaussian kernels that are located at the most discriminating time points. The most discriminating time point corresponds to the lowest LOO error and is not necessarily the point that at which the difference between the average trials is greatest. As shown in Fig. 3(c) and (d), at time 0, there is a clear amplitude difference between the MRPs of the imagined right-hand and foot movements. This is due to the normalization. Since the subsequent negative voltage shift in the foot imagery condition created a net negative value for the trial mean, removal of the trial mean created the positive values at time 0 in the foot imagery condition.

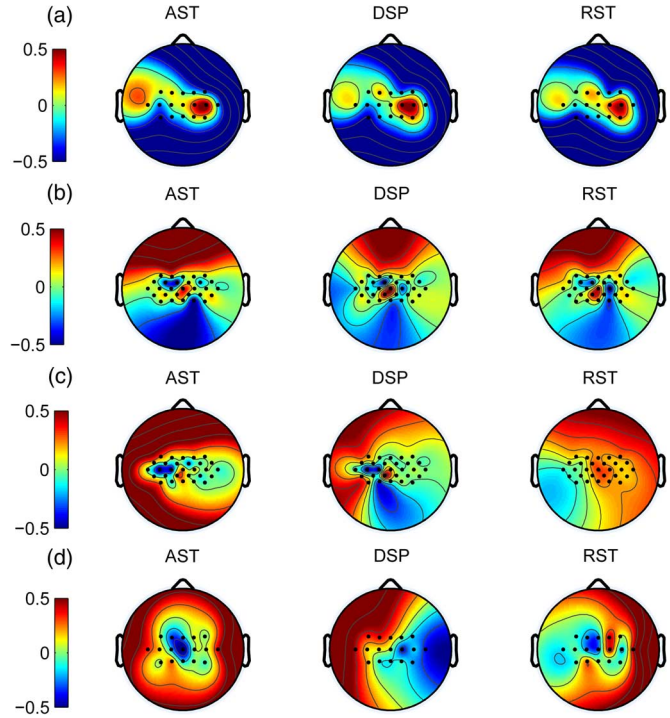


Fig. 4. Scalp maps from representative subjects of normalized spatial filters produced by AST, DSP, and RST filtering. For DSP and RST filtering, the scalp maps correspond to the largest eigenvalues. (a) Subject S2. (b) Subject S4. (c) Subject S6. (d) Subject S11.

4) Comparison of Spatial Filters: Fig. 4 shows some examples of spatial filters obtained by different filtering methods for different subjects. Note that since RST algorithm does not separately estimate the spatial filter and temporal filter, but optimizes a projection vector called temporal-spatial filter, we sum the elements for each channel to represent the spatial filter weight for that channel. The spatial filters of RST, DSP, and AST were normalized to keep the sum of the squares of the elements equal to 1. When a weight is close to 0, then it means that the channel does not contribute much to spatial filtering. In contrast, a weight with high absolute value denotes an important role of the channel. For the AST filtering as compared to the DSP filtering and RST filtering, the absolute values of filter weights are more sharply focused over motor areas around C3, Cz, and C4 corresponding to different limb movements/motor imageries.

5) Comparison of Temporal Filters: Fig. 5 shows the temporal filters produced by RST and AST filtering algorithms. Fig. 5 also shows the temporal filters used by DSP as described in (2). Note that for RST filtering, consecutive five-tuple of data points were averaged beforehand. Since RST filtering does not separately optimize the spatial filter and temporal filter, but optimizes a projection vector called temporal-spatial filter, we used the average of all channels at a given time point as the temporal filter weight for that time point. The temporal filters of DSP, RST, and AST filtering were normalized so that the sum of the squared filter elements was 1. Fig. 5(a) shows that for subject S2, the temporal filters of RST, DSP, and AST filtering have greater absolute values close to the onsets of finger movements, which is in accord with the previous studies [19], [37]. Fig. 5(b)–(d) shows this for subjects S4, S6, and S11, respectively. For these subjects, the classification accuracies of DSP and RST filtering

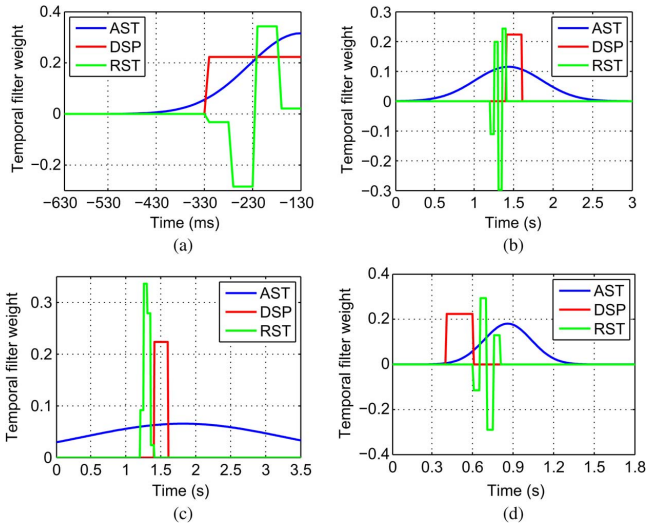


Fig. 5. Optimal temporal filters of RST, DSP, and AST filtering for the representative subjects. (a) Subject S2, for AST filtering, $\tau = 50, \theta = \exp(-5.44)$. (b) Subject S4, for AST filtering, $\tau = 143.55, \theta = \exp(-8.18)$. (c) Subject S6, for AST filtering, $\tau = 182.38, \theta = \exp(-10.63)$. (d) Subject S11, for AST filtering, $\tau = 86.05, \theta = \exp(-6.39)$.

were poor. It is probably because that the timings of the imagined movements varied in large ranges, the 200 ms window was sensitive to this variation, while the AST filtering adaptively adjusted the Gaussian filter with wider radiuses which are more robust. Generally, the temporal filters of DSP are not optimized to capture the maximum information in the signal. The temporal filters of RST have a larger number of parameters which may lead to overfitting of the limited training set. The temporal filters of AST are adaptive smoothers to extract the MRP components.

6) Why AST Filtering Works Better: Since the MRP has a slowly evolving activity, AST, DSP, and RST each use low frequency amplitude as signal features, but they do this in different ways. First, DSP only performs feature extraction in conjunction with a specific classifier, such as FDA or SVM [19]. Since the objectives of DSP and the classifier are different, it cannot theoretically ensure that DSP and the classifier will work well together. AST and RST integrate feature extraction and classification in predictive models with unified objectives. In this aspect, AST and RST are superior to DSP. Second, both DSP and RST need additional low-pass filtering and selecting of the time segment [19], [27], [37]. Since the parameters of this preprocessing are obtained by grid search they need significant human intervention and may be inaccurate due to the limited resolution of the grids. (If the resolution of the grids is high, the computation burden could increase dramatically). AST learns these parameters smoothly in an automatic manner which is more suitable for building a practical and feasible BCI. Third, since the EEG matrix structure is collapsed to make vector inputs for RST, the temporal and spatial correlations of the EEG signals are lost [27], [37]. For example, if an EEG matrix recorded by m channels and n time points (e.g., $m = 17$, $n = 20$) is represented as a mn-dimensional vector, it suggests that the EEG is specified by mn (e.g., 340) independent variables. However, only a few latent variables would be necessary to model the variance in the EEG that predicts the criterion. The degrees of freedom of the MRP filtering model should be far less than mn. When the training size is small, such as with subjects S5, S6, and S7 in dataset IV, BCI competition II, RST may be overfitting the data with a large number of redundant and irrelevant variables. In contrast, AST retains the EEG matrix structure while the temporal and spatial filter parameters are specified with just $2 + m$ degrees of freedom. We note that although AST avoids the overfitting issue by constraining the parameter space, as a double-edged sword, over-restricting the parameter space may cause underfitting to the data as well. Therefore, when the data set is large enough, a sophisticated learning model with more parameters may perform better.

7) The Online Evaluation of AST Filtering: At present, the AST filtering algorithm is tested offline without intervention and feedback. However, in a close loop BCI system, man and machine adapt to each other and the mental state of the subject will be more complex in the online situation. Since the online evaluating is more close to BCI applications, we plan to evaluate the performance of AST filtering online and revised it in a future study.

8) The Inter-Session Variability Problem: In this study, the experiment data consisted of several sessions. These sessions were conducted on the same day with some minutes break in between [37], [48]. The AST filtering model can generalize well from training set to test set which means that AST may help to overcome the intersession variability by filtering out the variable noise and extract the underlying MRP features. However, if the interval between sessions were long, such as several days or weeks, the inter-session variability of the MRP may be great (due to the differences in subjects' mental status or recoding environments). Then the marginal distribution of extracted features will change which is called as "covariate shift" [50], [51]. Present AST filtering based on ordinary LRR and cross validation does not track the "covariate shift." In the future, AST filtering can be improved by transfer learning techniques [52], [53] to address the inter-session variability problem.

V. Conclusion

In this paper, we proposed the AST filtering algorithm for MRPs extraction and classification. In order to better model MRPs and reduce the danger of overfitting, the AST algorithm constructs a low-pass temporal filter by using a Gaussian kernel with only two parameters (i.e., the center and the radius) and builds the spatial filter based on LRR model with a regularization item. Moreover, the AST algorithm uses the closed form of LOO error with LRR to simultaneously optimize the Gaussian kernel parameters, the spatial filter and the regularization coefficient with the efficient gradient descend method. Thus, the AST algorithm integrates MRP signal processing and prediction. This approach does not need the time consuming and imprecise preselecting of cutoff frequencies and time segments, but directly classifies the raw EEG signal. We compared the AST filter with DSP filtering and RST filtering on four BCI datasets from 12 subjects. Results showed that the proposed AST filtering can extract physiological meaningful information of MRP and outperform DSP filtering as well as RST filtering in average classification accuracy.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers who have given many valuable comments. They

would also like to acknowledge the Berlin BCI group, Germany for sharing their dataset.

REFERENCES

[1] J. Wolpaw and E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice*. New York: Oxford Univ. Press, 2012.

[2] R. Tomioka and K.-R. Müller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface," *Neuroimage*, vol. 49, no. 1, pp. 415–432, 2010.

[3] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain-bomputer interface: Fast acquisition of effective performance in untrained subjects," *Neuroimage*, vol. 37, no. 2, pp. 539–550, 2007.

[4] R. Ortner, B. Z. Allison, G. Korisek, H. Gaggl, and G. Pfurtscheller, "An SSVEP BCI to control a hand orthosis for persons with tetraplegia," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 1, pp. 1–5, Feb. 2011.

[5] T. Yu, Y. Li, J. Long, and Z. Gu, "Surfing the internet with a BCI mouse," *J. Neural Eng.*, vol. 9, no. 3, p. 036012, 2012.

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

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

[8] G. Pfurtscheller, G. Muller-Putz, A. Schlogl, B. Graimann, R. Scherer, R. Leeb, C. Brunner, C. Keinrath, F. Lee, and G. Townsend *et al.*, "15 years of BCI research at Graz University of Technology: Current projects," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 205–210, Jun. 2006.

[9] J. Farquhar, "A linear feature space for simultaneous learning of spatio-spectral filters in BCI," *Neural Netw.*, vol. 22, no. 9, pp. 1278–1285, 2009.

[10] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Muller, V. Kunzmann, F. Losch, and G. Curio, "The Berlin brain-computer interface: EEG-based communication without subject training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 147–152, Jun. 2006.

[11] G. Schalk, J. R. Wolpaw, D. J. McFarland, and G. Pfurtscheller, "EEG-based communication: Presence of an error potential," *Clin. Neurophysiol.*, vol. 111, no. 12, pp. 2138–2144, 2000.

[12] L. A. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. 6, pp. 510–523, 1988.

[13] J. Dien, K. M. Spencer, and E. Donchin, "Localization of the event-related potential novelty response as defined by principal components analysis," *Cognitive Brain Res.*, vol. 17, no. 3, pp. 637–650, 2003.

[14] S. Lemm, G. Curio, Y. Hlushchuk, and K.-R. Muller, "Enhancing the signal-to-noise ratio of ICA-based extracted ERPs," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 4, pp. 601–607, Apr. 2006.

[15] B. G. Edwards, V. D. Calhoun, and K. A. Kiehl, "Joint ICA of ERP and fMRI during error-monitoring," *Neuroimage*, vol. 59, no. 2, pp. 1896–1903, 2012.

[16] Z. He, S. Xie, S. Ding, and A. Cichocki, "Convolutive blind source separation in frequency domain based on sparse representation," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 5, pp. 1551–1563, Jul. 2007.

[17] G. Zhou, S. Xie, Z. Yang, and J. Zhang, "Nonorthogonal approximate joint diagonalization with well-conditioned diagonalizers," *IEEE Trans. Neural Netw.*, vol. 20, no. 11, pp. 1810–1819, Nov. 2009.

[18] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.

[19] X. Liao, D. Yao, D. Wu, and C. Li, "Combining spatial filters for the classification of single-trial EEG in a finger movement task," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 821–831, May 2007.

[20] H. Wang and J. Xu, "Local discriminative spatial patterns for movement-related potentials-based EEG classification," *Biomed. Signal Process. Control*, vol. 6, no. 4, pp. 427–431, 2011.

[21] D. J. Krusienski, E. W. Sellers, F. Cabestaing, S. Bayoudh, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "A comparison of classification techniques for the P300 speller," *J. Neural Eng.*, vol. 3, no. 4, p. 299, 2006.

[22] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "BCI competition 2003-data set IIb: Support vector machines for the P300 speller paradigm," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1073–1076, Jun. 2004.

[23] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[24] A. Rakotomamonjy and V. Guigue, "BCI competition III: Dataset II-ensemble of SVMs for BCI P300 speller," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1147–1154, Mar. 2008.

[25] K.-R. Muller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 165–169, Jun. 2003.

[26] J. Mak, Y. Arbel, J. Minett, L. McCane, B. Yuksel, D. Ryan, D. Thompson, L. Bianchi, and D. Erdogmus, "Optimizing the P300-based brain-computer interface: Current status, limitations and future directions," *J. Neural Eng.*, vol. 8, no. 2, p. 025003, 2011.

[27] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP componentsła tutorial," *Neuroimage*, vol. 56, no. 2, pp. 814–825, 2011.

[28] J. Farquhar and N. J. Hill, "Interactions between pre-processing and classification methods for event-related-potential classification," *Neuroinformatics*, pp. 1–18, 2012.

[29] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[30] D. J. McFarland and J. R. Wolpaw, "Sensorimotor rhythm-based brain-computer interface (BCI): Feature selection by regression improves performance," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, no. 3, pp. 372–379, Sep. 2005.

[31] J. Fruitet, D. J. Mcfarland, and J. R. Wolpaw, "A comparison of regression techniques for a two-dimensional sensorimotor rhythm-based brain-computer interface," *J. Neural Eng.*, vol. 7, no. 1, p. 016003, 2010.

[32] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Stat. Soc. Ser. B Methodol.*, pp. 111–147, 1974.

[33] R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*. New York: Chapman Hall, 1982, vol. 5.

[34] L. Bo, L. Wang, and L. Jiao, "Feature scaling for kernel fisher discriminant analysis using leave-one-out cross validation," *Neural Comput.*, vol. 18, no. 4, pp. 961–978, 2006.

[35] R. Cui, D. Huter, W. Lang, and L. Deecke, "Neuroimage of voluntary movement: Topography of the bereitschaftspotential, a 64-channel DC current source density study," *Neuroimage*, vol. 9, no. 1, pp. 124–134, 1999.

[36] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Combining features for BCI," in *Adv. Neural Inf. Process. Syst.*, 2002, pp. 1115–1122.

[37] B. Blankertz, G. Curio, and K.-R. Muller, "Classifying single trial EEG: Towards brain computer interfacing," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 157–164, 2002.

[38] G. Pires, U. Nunes, and M. Castelo-Branco, "Single-trial EEG classification of movement related potential," in *Proc. IEEE 10th Int. Conf. Rehabil. Robot.*, 2007, pp. 569–574.

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

[40] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulars, Graphs, and Mathematical Tables*. Mineola, NY: Dover, 1964, vol. 55.

[41] A. Luntz and V. Brailovsky, "On estimation of characters obtained in statistical procedure of recognition," *Technicheskaya Kibernetica*, vol. 3, no. 6, 1969.

[42] F. Ojeda, J. A. Suykens, and B. De Moor, "Low rank updated LS-SVM classifiers for fast variable selection," *Neural Netw.*, vol. 21, no. 2, pp. 437–449, 2008.

[43] S. Arlot and A. Celisse, "Segmentation of the mean of heteroscedastic data via cross-validation," *Stat. Comput.*, vol. 21, no. 4, pp. 613–632, 2011.

[44] J. Yuan, X. Liu, and C.-L. Liu, "Leave-one-out manifold regularization," *Exp. Syst. Appl.*, vol. 39, no. 5, pp. 5317–5324, 2012.

[45] G. C. Cawley and N. L. Talbot, "Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers," *Pattern Recognit.*, vol. 36, no. 11, pp. 2585–2592, 2003.

[46] G. C. Cawley and N. L. Talbot, "Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters," *J. Mach. Learn. Res.*, vol. 8, pp. 841–861, 2007.

[47] S. M. Selby, *Standard Mathematical Tables*. Boca Raton, FL: CRC, 1970.

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

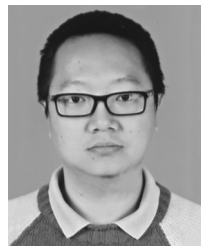
[49] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[50] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama, "Application of covariate shift adaptation techniques in brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 6, pp. 1318–1324, Jun. 2010.

[51] A. Satti, C. Guan, D. Coyle, and G. Prasad, "A covariate shift minimisation method to alleviate non-stationarity effects for an adaptive brain-computer interface," in *Proc. 20th IEEE Int. Conf. Pattern Recognit.*, 2010, pp. 105–108.
[52] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[53] M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Neural Inf. Process. Syst.*, 2007, vol. 7, pp. 1433–1440.

Jun Lu received the Ph.D. degree in electronic engineering from South China University of Technology, Guangzhou, China, in 2008.

In 2009 and 2010, he was a post doctor at Center for Brain Computer Interfaces and Brain Information Processing, South China University of Technology, Guangzhou, China, where he has worked on EEG signal processing and brain–computer interfaces. Since 2011, he has been a Researcher in the School of Automation, Guangdong University of Technology, Guangzhou, China. In 2012, he was a Visiting Scientist at Laboratory of Neural Injury and Repair, Wadsworth Center, New York State Department of Health, Albany, NY, USA. His research interests include feature learning and brain–computer interfaces.



Kan Xie received the M.S. degree in software engineering from the South China University of Technology, Guangzhou, China, in 2009. Currently, he is pursuing the Ph.D. degree in intelligent signal and information processing at the Guangdong University of Technology, Guangzhou, China.

His research interests include machine learning, nonnegative signal processing, blind signal processing, and biomedical signal processing.

Dennis J. McFarland received the Ph.D. degree from the University of Kentucky, Lexington, KY, USA, in 1978.

He is currently a Research Scientist at the Wadsworth Center, New York State Department of Health, Albany, NY, USA. His research interests include development of EEG-based communication and analysis of auditory processing.