33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Optimal Design of a Bank of Spatio-Temporal Filters for EEG Signal Classification

Hiroshi Higashi and Toshihisa Tanaka

***Abstract*— The spatial weights for electrodes called common spatial pattern (CSP) are known to be effective in EEG signal classification for motor imagery based brain computer interfaces (MI-BCI). To achieve accurate classification in CSP, the frequency filter should be properly designed. To this end, several methods for designing the filter have been proposed. However, the existing methods cannot consider plural brain activities described with different frequency bands and different spatial patterns such as activities of mu and beta rhythms. In order to efficiently extract these brain activities, we propose a method to design plural filters and spatial weights which extract desired brain activity. The proposed method designs finite impulse response (FIR) filters and the associated spatial weights by optimization of an objective function which is a natural extension of CSP. Moreover, we show by a classification experiment that the bank of FIR filters which are designed by introducing an orthogonality into the objective function can extract good discriminative features. Moreover, the experiment result suggests that the proposed method can automatically detect and extract brain activities related to motor imagery.**

I. INTRODUCTION

Electroencephalography (EEG)-based brain computer interface (BCI) is a challenging application in biomedical engineering. Since changes of energies of rhythmically oscillating components related movements of limbs are observed in an EEG signal, so-called motor imagery based BCI (MI-BCI) is a promising paradigm of BCI [1], [2].

An efficient method for extracting the brain activity for MI-BCI is the common spatial pattern (CSP) [3], which uses spatial weights that extract the most discriminative information. The spatial weights minimize the variance ratio of the spatio-filtered signals for two classes in a learning dataset. Although together with the spatial weights, an input EEG signal is successfully classified, for the effective implementation, the observed signal should be bandpass-filtered to extract frequency components associated with motor imagery activities [3]. For the bandpass filtering, a passband of 7–30Hz is typically chosen in MI-BCI [1], [3]–[6]. However, the optimum frequency band for classification is highly dependent on users and measurement environments [3]. Therefore, it is crucial to find an optimal filter for classification in MI-BCI. Recently, several approaches to this problem such as common spatio-spectral pattern (CSSP) [4], common sparse spectral spatial pattern (CSSSP) [5], and spectrally weighted CSP (SPEC-CSP) [6] have been proposed.

The brain activities observed in the bands of 8–13 Hz (called mu rhythm which is generated in a sensorimotor area) and 13–30 Hz (called beta rhythm which is mainly generated in a sinciput area) are used as features in MI-BCI [2], because the band of 7–30 Hz is generally specified in the CSP procedure. However, because CSP, CSSSP, and SPEC-CSP use only one frequency filter, these methods cannot efficiently extract plural brain activities occurring in different frequency bands and different spatial patterns such as mu and beta rhythms. Moreover, as well as frequency bands, the spatial patterns of the brain activities highly depend on users and measurement environment [1].

This paper provides a new method to automatically design a bank of frequency/spatial filters using learning datasets. In the proposed method, we introduce into the CSP cost function the additional parameters representing coefficients of an finite impulse response (FIR) filter. That is, the proposed cost function provides optimization of parameters including spatial weights and the FIR filter coefficients. Moreover, to extract multiple FIR filters, we add to the optimization problem the constraint that the vectors whose elements are coefficients of filters are orthogonal to each other. An optimization procedure based on the alternating least square (ALS) is used to solve the proposed optimization problem, where the optimization problem is divided into several subproblems. Each subproblem is reduced to generalized eigenvalue problem, which can be solved by a well-established optimization method. In the experimental section, it has been shown that features extracted by a bank of spatio-temporal filters given by the proposed method are effective and superior for classification in MI-BCI.

II. COMMON SPATIAL PATTERN (CSP)

We first review a basic CSP method [3]. Let $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ be a matrix representing observed signals, where M is the number of channels and N is the number of samples. The CSP finds a spatial weight vector, $\boldsymbol{w} \in \mathbb{R}^M$, in such a way that the variance of a signal extracted by linear combination of $\boldsymbol{X}$ and $\boldsymbol{w}$ is minimized in a class [3]. In BCI application, we do not directly use $\boldsymbol{X}$, but use the filtered signal described as $\hat{\boldsymbol{X}} = \mathcal{H}(\boldsymbol{X})$ in CSP, where $\mathcal{H}$ is a bandpass filter which enhances brain activity of motor imagery. Denote the components (vectors) of $\hat{\boldsymbol{X}}$ by $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_N]$, where $\hat{\boldsymbol{x}}_n \in \mathbb{R}^M$ and n is the time index. The time average of the observed signal is given by $\boldsymbol{\mu} = N^{-1}\sum_{n=1}^{N}\hat{\boldsymbol{x}}_n$. Then, the time variance of the extracted signal of $\hat{\boldsymbol{X}}$ is given by

H. Higashi and T. Tanaka are with the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan higashi@sip.tuat.ac.jp, tanakat@cc.tuat.ac.jp

H. Higashi and T. Tanaka are also with the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Saitama, Japan.

This work is supported in part by KAKENHI, Grant-in-Aid for Scientific Research (B), 21360179.

$\sigma^2(\boldsymbol{X},\boldsymbol{w}) = N^{-1}\sum_{n=1}^{N}|\boldsymbol{w}^T(\hat{\boldsymbol{x}}_n-\boldsymbol{\mu})|^2$, where $\cdot^T$ denotes the transpose of a vector or matrix. We assume that sets of the learning data are represented as $\mathcal{C}_1$ and $\mathcal{C}_2$, where $\mathcal{C}_d$ contains the signals belonging to class d, d represents a class label chosen in $\{1,2\}$, and $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. CSP finds the weight vector that minimizes the intra-class variance in $\mathcal{C}_c$ under the normalization of samples, where c is a class label. More specifically, for class, c, fixed, CSP finds $\boldsymbol{w}_c$ by solving the following optimization problem [3];

$$\begin{aligned}\min_{\boldsymbol{w}} \quad & E_{\boldsymbol{X}\in\mathcal{C}_c}[\sigma^2(\boldsymbol{X},\boldsymbol{w})],\\ \text{subject to} \quad & \sum_{d=1,2} E_{\boldsymbol{X}\in\mathcal{C}_d}[\sigma^2(\boldsymbol{X},\boldsymbol{w})] = 1,\end{aligned} \tag{1}$$

where $E_{\boldsymbol{X}\in\mathcal{C}_d}[\cdot]$ denotes the expectation over $\mathcal{C}_d$. Then, (1) can be rewritten as

$$\min_{\boldsymbol{w}} \quad \boldsymbol{w}^T\boldsymbol{\Sigma}_c\boldsymbol{w}, \quad \text{subject to} \quad \boldsymbol{w}^T(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\boldsymbol{w}=1, \tag{2}$$

where $\boldsymbol{\Sigma}_d$ are defined as $\boldsymbol{\Sigma}_d = E_{\boldsymbol{X}\in\mathcal{C}_d}[N^{-1}\sum_{n=1}^{N}(\hat{\boldsymbol{x}}_n-\boldsymbol{\mu})(\hat{\boldsymbol{x}}_n-\boldsymbol{\mu})^T]$, for $d=1,2$. The solution of (2) is given by the generalized eigenvector corresponding to the smallest generalized eigenvalue of the generalized eigenvalue problem described as

$$\boldsymbol{\Sigma}_c\boldsymbol{w} = \lambda(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\boldsymbol{w}. \tag{3}$$

Though the solution of (2) is given by the eigenvector corresponding to the smallest eigenvalue in (3), we can use the other eigenvectors for feature extraction [4]–[6]. The M eigenvectors can be obtained by solving (3) as $\hat{\boldsymbol{w}}_1,\ldots,\hat{\boldsymbol{w}}_M$, where $\hat{\boldsymbol{w}}_i$ is the eigenvector corresponding to the i-th largest eigenvalue of (3). We assume that the $2r$ eigenvectors are used to classify an unlabeled data, $\boldsymbol{X}$. Then we obtain the feature vector, $\boldsymbol{y}\in\mathbb{R}^{2r}$, from $\boldsymbol{X}$ defined as

$$\begin{aligned}\boldsymbol{y} =& [\sigma^2(\boldsymbol{X},\hat{\boldsymbol{w}}_1),\ldots,\sigma^2(\boldsymbol{X},\hat{\boldsymbol{w}}_r),\\ & \sigma^2(\boldsymbol{X},\hat{\boldsymbol{w}}_{M-r+1}),\ldots,\sigma^2(\boldsymbol{X},\hat{\boldsymbol{w}}_M)]^T.\end{aligned} \tag{4}$$

For classification, $\boldsymbol{y}$ is input to a classifier.

III. FIR FILTER DESIGN METHOD

As mentioned earlier, the classification accuracy depends on the choice of the bandpass filter $\mathcal{H}$ [3]. In this section, we present the method for designing a bank of spatio-temporal filters. We develop a cost function to design the bank of filters and an optimization procedure of the cost function using an alternating optimization.

A. Feature Extraction

Let θ_p, $p=1,\ldots,P$ be the filter coefficients of an FIR filter where P is the order of the filter, and the filtered signal of $\boldsymbol{X}$ denoted as $\hat{\boldsymbol{X}}=[\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_K]$, can be defined as $\hat{\boldsymbol{x}}_n = \sum_{p=1}^{P}\theta_p\boldsymbol{x}_{n+P-p}$ for $n=1,\ldots,K$, where $K=N-P+1$. The variance of the filtered signal with the spatial weight is

$$\alpha(\boldsymbol{X},\boldsymbol{w},\boldsymbol{\theta}) = \frac{1}{K}\sum_{n=1}^{K}\left|\boldsymbol{w}^T\sum_{p=1}^{P}\theta_p\boldsymbol{x}_{n+P-p}-\boldsymbol{\mu}\right|^2, \tag{5}$$

where a vector of $\boldsymbol{w}$ acts as spatial weights, $\boldsymbol{\theta}$ is the vector of the filter coefficients defined as $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_P]^T$, and $\boldsymbol{\mu}$ is the time average given by $\boldsymbol{\mu} = K^{-1}\sum_{n=1}^{K}\boldsymbol{w}^T\sum_{p=1}^{P}\theta_p\boldsymbol{x}_{n+P-p}$. We define $\boldsymbol{A}_n$, $n=1,\ldots,K$, whose elements are from $\boldsymbol{X}$ defined as

$$[\boldsymbol{A}_n]_{m,p} = [\boldsymbol{X}]_{m,n+P-p}, \tag{6}$$

where $m=1,\ldots,M$, $p=1,\ldots,P$ and $[\cdot]_{i,j}$ denotes the entry in ith row and jth column of a matrix. Therefore, (5) can be modified to

$$\alpha(\boldsymbol{X},\boldsymbol{w},\boldsymbol{\theta}) = \frac{1}{K}\sum_{n=1}^{K}\left|\boldsymbol{w}^T\hat{\boldsymbol{A}}_n\boldsymbol{\theta}\right|^2, \tag{7}$$

where $\hat{\boldsymbol{A}}_n$ is defined as $\hat{\boldsymbol{A}}_n = \boldsymbol{A}_n - K^{-1}\sum_{m=1}^{K}\boldsymbol{A}_m$.

B. Design Criteria and Optimization

For the feature value defined in (7), we design F frequency filters and the associated spatial weights. Let $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ for $i=1,\ldots,F$ be the spatial weights and the coefficients of filters, respectively. The underlying idea behind the proposed method is to find optimal value of both $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ by maximization of $\alpha(\boldsymbol{X},\boldsymbol{w}_i,\boldsymbol{\theta}_i)$ with respect to $\boldsymbol{X}\in\mathcal{C}_c$ under the normalization of samples. Additionally, to obtain the different filter coefficients in each $\boldsymbol{\theta}_i$, we introduce an additional constraint that $\boldsymbol{\theta}_i$, $i=1,\ldots,F$ are mutually orthogonal. Therefore, we formulate the following maximization problem;

$$\begin{aligned}\max_{\boldsymbol{w}_i,\boldsymbol{\theta}_i,i=1,\ldots,F} \quad & \sum_{i=1}^{F}\hat{J}(\boldsymbol{w}_i,\boldsymbol{\theta}_i),\\ \text{subject to} \quad & \frac{\boldsymbol{\theta}_k^T\boldsymbol{\theta}_j}{\|\boldsymbol{\theta}_k\|\,\|\boldsymbol{\theta}_j\|} = \delta_{kj}, \quad k,j=1,\ldots,F,\end{aligned} \tag{8}$$

where

$$\hat{J}(\boldsymbol{w},\boldsymbol{\theta}) = \frac{\sum_{d=1,2}E_{\boldsymbol{X}\in\mathcal{C}_d}[\alpha(\boldsymbol{X},\boldsymbol{w},\boldsymbol{\theta})]}{E_{\boldsymbol{X}\in\mathcal{C}_c}[\alpha(\boldsymbol{X},\boldsymbol{w},\boldsymbol{\theta})]}, \tag{9}$$

c is an optional class label, and δ_{ij} is the Kronecker delta defined as 1 for $i=j$ and 0 otherwise. In (8), the cost function can be divided with respect to a filter index, i, however $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ should be update when other optimization parameters, $\boldsymbol{w}_j$ and $\boldsymbol{\theta}_j$, are changed due to the orthogonal constraint where $j\in\{k|k=1,\ldots,F,k\neq i\}$. Then, we optimize $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ in an order of 1 to F, that is, the filters, $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{i-1}$, are already fixed in the optimization step for $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$. By sequential optimization, we can represent (8) with respect to each i as

$$\begin{aligned}\max_{\boldsymbol{w}_i,\boldsymbol{\theta}_i} \quad & \hat{J}(\boldsymbol{w}_i,\boldsymbol{\theta}_i)\\ \text{subject to} \quad & \boldsymbol{\theta}_i^T\boldsymbol{\theta}_j = 0, \quad j=1,\ldots,i-1.\end{aligned} \tag{10}$$

In (10), since we cannot seek for $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ simultaneously, we adopt alternating optimization procedure based on ALS. Two subproblems that separately find $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ are obtained as follows.

The first subproblem is to optimize $\boldsymbol{w}_i$ while fixing $\boldsymbol{\theta}_i$. Define $\boldsymbol{R}_d(\boldsymbol{\theta}) = E_{\boldsymbol{X}\in\mathcal{C}_d}\left[K^{-1}\sum_{n=1}^{K}\hat{\boldsymbol{A}}_n\boldsymbol{\theta}\boldsymbol{\theta}^T\hat{\boldsymbol{A}}_n^T\right]$ for $d=1,2$. Then (8) can be written as

$$\max_{\boldsymbol{w}_i} \quad \hat{J}_1(\boldsymbol{w}_i|\boldsymbol{\theta}_i) = \frac{\boldsymbol{w}_i^T(\boldsymbol{R}_1(\boldsymbol{\theta}_i)+\boldsymbol{R}_2(\boldsymbol{\theta}_i))\boldsymbol{w}_i}{\boldsymbol{w}_i^T\boldsymbol{R}_c(\boldsymbol{\theta}_i)\boldsymbol{w}_i}. \tag{11}$$

The solution of (11) is given by the generalized eigenvector corresponding to the largest generalized eigenvalue of the generalized eigenvalue problem described as

$$(\boldsymbol{R}_1(\boldsymbol{\theta}_i) + \boldsymbol{R}_2(\boldsymbol{\theta}_i))\boldsymbol{w}_i = \lambda \boldsymbol{R}_c(\boldsymbol{\theta}_i)\boldsymbol{w}_i. \tag{12}$$

We normalize $\boldsymbol{w}_i$ to unit norm.

The second subproblem is to optimize $\boldsymbol{\theta}_i$ while fixing $\boldsymbol{w}_i$. Define $\boldsymbol{Q}_d(\boldsymbol{w}) = E_{\boldsymbol{X}\in C_d}\left[K^{-1}\sum_{n=1}^{K}\hat{\boldsymbol{A}}_n^T\boldsymbol{w}\boldsymbol{w}^T\hat{\boldsymbol{A}}_n\right]$ for $d = 1, 2$. Then (10) can be written as

$$\max_{\boldsymbol{\theta}_i} \quad \hat{J}_2(\boldsymbol{\theta}_i|\boldsymbol{w}_i) = \frac{\boldsymbol{\theta}_i^T(\boldsymbol{Q}_1(\boldsymbol{w}_i) + \boldsymbol{Q}_2(\boldsymbol{w}_i))\boldsymbol{\theta}_i}{\boldsymbol{\theta}_i^T\boldsymbol{Q}_c(\boldsymbol{w}_i)\boldsymbol{\theta}_i}, \tag{13}$$
$$\text{subject to} \quad \boldsymbol{\theta}_i^T\boldsymbol{\theta}_j = 0, \quad j = 1, \ldots, i-1.$$

The solution to (13) is given by the following theorem.

Theorem 1: When matices $\boldsymbol{Q}_c(\boldsymbol{w}_i)$ and $\boldsymbol{Q}_1(\boldsymbol{w}_i)+\boldsymbol{Q}_2(\boldsymbol{w}_i)$ are nonsingular, the solution of (13) is given by the unit-length generalized eigenvector corresponding to the largest generalized eigenvalue of the generalized eigenvalue problem described as

$$\boldsymbol{G}(\boldsymbol{Q}_1(\boldsymbol{w}_i) + \boldsymbol{Q}_2(\boldsymbol{w}_i))\boldsymbol{\theta}_i = \zeta\boldsymbol{Q}_c(\boldsymbol{w}_i)\boldsymbol{\theta}_i, \tag{14}$$

where

$$\begin{aligned}\boldsymbol{G} =& \boldsymbol{I}_P - \boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{Q}_c(\boldsymbol{w}_i)\boldsymbol{H})^{-1} \\ &\times \boldsymbol{H}^T\left(\boldsymbol{Q}_1(\boldsymbol{w}_i) + \boldsymbol{Q}_2(\boldsymbol{w}_i)\right)^{-1},\end{aligned} \tag{15}$$

$\boldsymbol{H}$ is the matrix by the already optimized filters defined as

$$\boldsymbol{H} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}] \in \mathbb{R}^{P\times(i-1)}, \tag{16}$$

$\boldsymbol{I}_P$ is the $P \times P$ identify matrix, and ζ is an eigenvalue.

Proof: For convenience, we define $\boldsymbol{Q}_c = \boldsymbol{Q}_c(\boldsymbol{w}_i)$ and $\boldsymbol{Q} = \boldsymbol{Q}_1(\boldsymbol{w}_i) + \boldsymbol{Q}_2(\boldsymbol{w}_i)$. Any vector $\boldsymbol{\theta}_i$ can be normalized such that $\boldsymbol{\theta}_i^T\boldsymbol{Q}_c\boldsymbol{\theta}_i = 1$, its norm becomes any value, and $\hat{J}_2(\boldsymbol{\theta}_i|\boldsymbol{w}_i)$ keeps unchanged. Therefore, the maximization of $\hat{J}_2(\boldsymbol{\theta}_i|\boldsymbol{w}_i)$ is equivalent to the maximization of $\boldsymbol{\theta}_i^T\boldsymbol{Q}\boldsymbol{\theta}_i$ with the constraint that $\boldsymbol{\theta}_i^T\boldsymbol{Q}_c\boldsymbol{\theta}_i = 1$. Then the Lagrangian of (13) is

$$L = \boldsymbol{\theta}_i^T\boldsymbol{Q}\boldsymbol{\theta}_i - \zeta(\boldsymbol{\theta}_i\boldsymbol{Q}_c\boldsymbol{\theta}_i - 1) - \sum_{j=1}^{i-1}\nu_j\boldsymbol{\theta}_i^T\boldsymbol{\theta}_j, \tag{17}$$

where ζ and $\nu_1, \ldots, \nu_{i-1}$ are Lagrange multipliers. The partial derivative of L with respect to $\boldsymbol{\theta}_i$ is

$$\frac{\partial L}{\partial\boldsymbol{\theta}_i} = 2\boldsymbol{Q}\boldsymbol{\theta}_i - 2\zeta\boldsymbol{Q}_c\boldsymbol{\theta} - \sum_{j=1}^{i-1}\nu_j\boldsymbol{\theta}_j. \tag{18}$$

Then, $\partial L/\partial\boldsymbol{\theta}_i$ is zero when $\zeta = (\boldsymbol{\theta}_i^T\boldsymbol{Q}\boldsymbol{\theta}_i)/(\boldsymbol{\theta}_i^T\boldsymbol{Q}_c\boldsymbol{\theta}_i)$, because multiplying $\partial L/\partial\boldsymbol{\theta}_i = 0$ by $\boldsymbol{\theta}_i^T$ is

$$2\boldsymbol{\theta}_i^T\boldsymbol{Q}\boldsymbol{\theta}_i - 2\zeta\boldsymbol{\theta}_i^T\boldsymbol{Q}_c\boldsymbol{\theta}_i - \sum_{j=1}^{i-1}\nu_j\boldsymbol{\theta}_i^T\boldsymbol{\theta}_j = 0, \tag{19}$$

where the third term in the left equation is zero because $\boldsymbol{\theta}_i^T\boldsymbol{\theta}_j = 0$ for $i \neq j$. Next, we use $\boldsymbol{H}$ defined as (16) and $\boldsymbol{\nu}$ defined as $\boldsymbol{\nu} = [\nu_1, \nu_2, \ldots, \nu_{i-1}]^T$. Then, multiplying $\partial L/\partial\boldsymbol{\theta}_i = 0$ by $\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}$ is $2\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{Q}\boldsymbol{\theta}_i - \boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{H}\boldsymbol{\nu} = 0$, because $2\zeta\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{Q}_c\boldsymbol{\theta}_i = 0$. Thus,

$$\boldsymbol{\nu} = 2(\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{Q}\boldsymbol{\theta}_i. \tag{20}$$

Algorithm 1 Design of a bank of spatio-temporal filters

Input: $\mathcal{C}_1$, $\mathcal{C}_2$: the sets of learning data of $\boldsymbol{X} \in \mathbb{R}^{M\times N}$.
Parameter: P: the filter order, F: the number of FIR filters.
Output: $\hat{\boldsymbol{w}}_i^{(m)}$, $\boldsymbol{\theta}_i$ (for $i = 1, \ldots, F$, $m = 1, \ldots, M$).

for $i = 1, \ldots, F$ **do**
  Initialize $\boldsymbol{\theta}_i$.
  Set the index of iteration as $k = 0$.
  repeat
    $k \leftarrow k + 1$
    Update $\boldsymbol{w}_i$ by solving (11).
    Update $\boldsymbol{\theta}_i$ by solving (13).
    Calculate cost, C_k from the cost function, $\hat{J}(\boldsymbol{w}_i, \boldsymbol{\theta}_i)$.
  until $C_k - C_{k-1}$ is sufficiently small.
  Obtain M spatial weights, $\hat{\boldsymbol{w}}_i^{(1)}, \ldots, \hat{\boldsymbol{w}}_i^{(M)}$, by (12).
end for

Substituting (20) into $\partial L/\partial\boldsymbol{\theta}_i = 0$ can be written as

$$\begin{aligned}2\boldsymbol{Q}\boldsymbol{\theta}_i - 2\zeta\boldsymbol{Q}\boldsymbol{\theta}_i - 2\boldsymbol{H}(\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{Q}_c^{-1}\boldsymbol{Q}\boldsymbol{\theta}_i = 0 \\ \boldsymbol{G}\boldsymbol{Q}\boldsymbol{\theta}_i = \zeta\boldsymbol{Q}_c\boldsymbol{\theta}_i,\end{aligned} \tag{21}$$

where $\boldsymbol{G}$ is defined in (15). Since ζ is the criterion to be maximized, the maximum solution of (13) is achieved by the unit-length generalized eigenvector corresponding to the largest generalized eigenvalue of (14). ■

As introduced in above, we alternately optimize $\boldsymbol{w}_i$ and $\boldsymbol{\theta}_i$ by solving optimization problem (11) and (13) for each index, i. We optimize the filters and the spatial weights in the order of $\{\boldsymbol{w}_1, \boldsymbol{\theta}_1\}$ to $\{\boldsymbol{w}_F, \boldsymbol{\theta}_F\}$. In the optimization, the initialization of $\boldsymbol{\theta}_i$ and $\boldsymbol{w}_i$ is an important topic. However, we do not discuss the initialization problem in this paper. We adopt a simple initialization way as follows: initialize $\boldsymbol{\theta}_i$ as a random vector which is orthnormalized from $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}$ by the Gram-Schmidt orthnormalization.

C. Feature Vector Definition

We use the eigenvectors given by (12) as the spatial weights for feature extraction. Let $\boldsymbol{\theta}_i$, $i = 1, \ldots, F$ be the filters given by (8). By solving (12) with $\boldsymbol{\theta}_i$, we obtain $M \cdot F$ eigenvectors as $\hat{\boldsymbol{w}}_i^{(m)}$ for $i = 1, \ldots, F$ and $m = 1, \ldots, M$, where $\hat{\boldsymbol{w}}_i^{(m)}$ is the unit-length eigenvector corresponding to m-th largest eigenvalue of (12). We assume that the $2r$ eigenvectors for each $\boldsymbol{\theta}_i$ are used for feature extraction. Then the feature vector, $\boldsymbol{y} \in \mathbb{R}^{2r\cdot F}$, is defined as

$$\begin{aligned}\boldsymbol{y} =& [\alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_1^{(1)}, \boldsymbol{\theta}_1), \ldots, \alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_1^{(r)}, \boldsymbol{\theta}_1), \\ &\alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_1^{(M-r+1)}, \boldsymbol{\theta}_1), \ldots, \alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_1^{(M)}, \boldsymbol{\theta}_1), \ldots, \\ &\alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_F^{(1)}, \boldsymbol{\theta}_F), \ldots, \alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_F^{(r)}, \boldsymbol{\theta}_F), \\ &\alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_F^{(M-r+1)}, \boldsymbol{\theta}_F), \ldots, \alpha(\boldsymbol{X}, \hat{\boldsymbol{w}}_F^{(M)}, \boldsymbol{\theta}_F)].\end{aligned} \tag{22}$$

As well as the case of CSP, we input $\boldsymbol{y}$ to a classifier. The procedure to design frequency/spatial filters is summarized in Algorithm 1 as a pseudo-code.

TABLE I: Classification accuracy [%] given by 5×5 CV. The figures in the round brackets beside accuracies represent the number of dimensions of the feature vector. The column labeled "(Param)" shows the value of the parameters used in the proposed method. In the brackets in the column, the first and second elements represent the number of filters, F, and the number of spatial weight vector, $2r$, for each filter.

Method	Subject *aa*	*al*	*av*	*aw*	*ay*
CSP-Ref	88.4 (4)	98.6 (2)	74.4 (10)	99.9 (6)	92.4 (4)
CSP	81.7 (2)	94.6 (6)	68.3 (8)	95.9 (6)	89.6 (2)
CSSP	84.1 (2)	95.4 (10)	69.9 (8)	96.9 (6)	90.5 (2)
SPEC-CSP	84.7 (10)	95.3 (6)	59.0 (20)	96.9 (6)	83.4 (4)
Proposed	**89.4**	**98.1**	**70.1**	**99.1**	**95.0**
(Param.)	(10, 4)	(1, 16)	(2, 4)	(2, 18)	(5, 14)

IV. EXPERIMENT

We compare performance in classifying EEG signals during motor imagery using the proposed method to that using existing methods (CSP, CSSP, and SPEC-CSP).

A. Data Description

We used dataset IVa from BCI competition III [7] (for details of the dataset, see `http://www.bbci.de/competition/iii/`). This dataset consists of EEG signals during right hand and right foot motor-imageries. The EEG signals were recorded from five subjects labeled *aa*, *al*, *av*, *aw*, and *ay*. The measured signal was bandpass filtered with the passband of 0.05–200 Hz, and then digitized at 1000 Hz. Moreover, we applied to this data the lowpass filter whose the cutoff frequency is 50 Hz, and downsampled to 100 Hz. The dataset for each subject consisted of signals of 140 trials. A signal of one trial was measured for 3.5 seconds.

B. Result

Table I shows the classification accuracies given by each method. In CSP, the feature vector defined in (4) with the bandpass filtering with the passband of 7–30 Hz was used. In CSSP, the delay sample was determined by 5×5 CV in learning dataset. In SPEC-CSP, we assumed that the filter coefficients not corresponding to 7–30 Hz are zero. For reference, CSP-Ref shows the classification accuracy rates when the passbands of the bandpass filters used in preprocessing were 11–16 Hz (*aa*), 12–16 Hz (*al*), 21–26 Hz (*av*), 11–18 Hz (*aw*), and 9–12 Hz (*ay*) in CSP procedure. These passbands were chosen by 5×5 CV from f_l–f_u Hz for $f_l = 1, \ldots, 25$ and $f_u = f_l + 1, \ldots, 30$ and performed the best accuracy among them. In the proposed method, we set the filter order, P, to 20. The dimension of the feature vector that performs the best classification accuracy was adopted. The extracted feature vector was classified by linear SVM with the soft margin parameter of 50 [8]. The linear SVM was implemented with SVM-Light [9]. Table I shows that the proposed method has the equal classification performance to CSP-Ref. Moreover, the proposed method outperforms CSP-Ref in classification accuracy of subjects *aa* and *ay* and many filters are used in classification for these subjects.

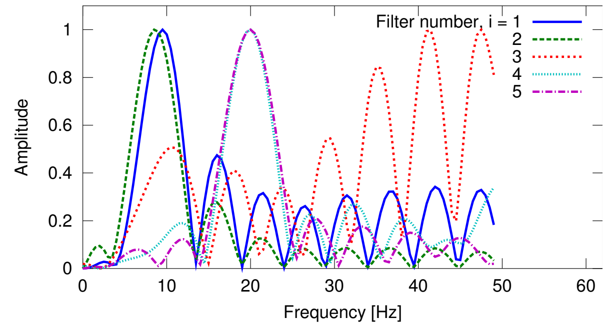

Fig. 1: Amplitude characteristics of the filters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_5$ given by the proposed method in subject *ay*.

Figure 1 shows amplitude characteristics of filters given by the proposed method. The filters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have a passband of about 5–14 Hz including the band called mu rhythm. Moreover, the band of about 15–25 Hz including the band called beta rhythm are specified by the filters $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$.

V. CONCLUSION

We have proposed a novel supervised feature extraction method by extending existing CSP. The proposed method extracts the feature by plural FIR filters and the associated spatial weights. The objective of the proposed method is to extract features associated with plural brain activities, which are observed in different frequency bands and have different spatial patterns. By the experiments, we have demonstrated that the proposed method performs high classification accuracy for the MI-BCI and the method is competitive to existing CSP algorithm. Interestingly, a passband of the resulting filters corresponds to a band of mu or beta rhythm. This suggests that the proposed method may be able to separately extract different brain activities.

REFERENCES

[1] S. Sanei and J. Chambers. *EEG Signal Processing.* Hoboken, NJ: John Wiley & Sons, 2007.

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

[3] J. Müller-Gerling, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filters for single-trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, pp. 787–789, 1999.

[4] S. Lemm, B. Blankertz, G. Curio, and K. R. Müller, "Spatio-spectral filters for improved classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 441–446, 2005.

[5] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K. R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274–2281, 2006.

[6] R. Tomioka, G. Dornhege, G. Nolte, K. Aihara, and K. R. Müller, "Optimizing spectral filters for single trial EEG classification," *Lecture Notes in Computer Science*, vol. 4174, pp. 414–423, 2006.

[7] B. Blankertz, K. R. Müller, D. J. Krusienski, G. Schalk, J. R. Wolpaw, A. Schlogl, G. Pfurtscheller, J. R. Millan, M. Schroder, and N. Birdaumer, "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Trnas. Neural Sys. Rehab. Eng.,*, vol. 14, no. 2, pp. 153–159, 2006.

[8] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[9] T. Joachims. "Making large-scale SVM learning practical" in *Advances in Kernel Method - Support Vector Learning.* B. Schölkopf, C. Burges, and A. Smola, Ed. MIT-Press, 1999.