An Expectation-Maximization Algorithm Based Kalman Smoother Approach for Event-Related Desynchronization (ERD) Estimation from EEG

Mohammad Emtiyaz Khan* and Deshpande Narayana Dutt

***Abstract*—We consider the problem of event-related desynchronization (ERD) estimation. In existing approaches, model parameters are usually found manually through experimentation, a tedious task that often leads to suboptimal estimates. We propose an expectation-maximization (EM) algorithm for model parameter estimation that is fully automatic and gives optimal estimates. Further, we apply a Kalman smoother to obtain ERD estimates. Results show that the EM algorithm significantly improves the performance of the Kalman smoother. Application of the proposed approach to the motor-imagery EEG data shows that useful ERD patterns can be obtained even without careful selection of frequency bands.**

***Index Terms*—Event-related desynchronization, expectation-maximization algorithm, Kalman smoother.**

I. Introduction

EVENT-RELATED desynchronization (ERD) and synchronization (ERS) are used to describe the decrease and increase in activity in an EEG signal, caused by *physical* events [1]. Experiments show that the preparation, planning and even imagination of specific movements result in ERD in mu and central-beta rhythms [2]–[4]. ERD also shows significant differences in EEG activity between left- and right-hand movements [5]. These differences can be used to build communication channels known as brain-computer interfaces (BCI) which have been very useful in providing assistance to paralyzed patients [6].

ERD has been studied extensively by researchers and many methods have been proposed for its estimation [7]–[11]. The intertrial variance (IV) method [7] is one of the first methods proposed for quantification of ERD. In this method, ERD estimates are obtained by computing the average IV of a band-pass filtered signal. Useful information about ERD time courses and the hemispherical asymmetry can be obtained with these estimates. However, the IV method cannot be used for on-line classification because it requires averaging over multiple trials [5]. Another problem is that it requires careful selection of frequency bands for ERD estimation. To overcome these problems, a method based on the adaptive-autoregressive (AAR) model has been proposed [8]. The AAR model is also known as the time-varying AR (TVAR) model, and has been applied extensively for EEG signal analysis [12], [13]. The TVAR coefficients are usually estimated with the recursive-least square (RLS) algorithm and classified with a linear-discriminator. It is shown in [5] that the TVAR coefficients capture the EEG patterns and improve classification accuracy. However, in this method, values of various parameters (e.g., model order, update coefficients) are required and are usually difficult to find.

The TVAR model can also be written as a state-space model. The advantage of this formulation is that the optimal estimates can be obtained using a Kalman filter [14]. The Kalman filter is an optimal estimator in the mean-square sense. Other adaptive algorithms like the RLS algorithm can be derived as a special case of the Kalman filter [15]. If *future* measurements are available, smoothing equations can be used to further improve estimation performance. The Kalman filter along with the smoothing equations is usually referred to as a Kalman smoother [16], [17]. The Kalman smoother has been used for ERD estimation in [18], it has been reported to improve the tracking of ERD patterns. However, in this formulation as in the AAR model formulation, setting model parameters is a problem. To make it easier to set the parameters, a very simple random-walk model is used.

In all the methods discussed above, finding the values of model parameters is a common issue. In this paper, we propose an expectation-maximization (EM) algorithm for model parameter estimation. We use the information present in large training datasets to estimate model parameters. The paper is organized as follows: In Section II, we describe the state-space formulation of the TVAR model. In Section III, we describe our algorithm for ERD estimation. In Section IV we discuss the results, followed by a conclusion in Section V

II. Time-Varying Autoregressive (TVAR) Model

We denote scalars/vectors/matrices by small/bold/capital letters. We denote the transpose of a matrix A by A'. We assume an EEG sequence follow the following TVAR model:

$$y_t = \sum_{k=1}^{p} a_k^t y_{t-k} + v_t. \tag{1}$$

Here, $\{a_k^t\}_{k=1}^p$ are the TVAR coefficients[1], p is the model order and v_t is the independent and identically distributed (i.i.d.) Gaussian noise with zero mean and variance σ_v^2. We denote a

Manuscript received December 13, 2005; revised October 29, 2006. *Asterisk indicates corresponding author.*

*M. E. Khan is with the Department of Computer Science, University of British Columbia, Vancouver, Canada, BC V6T1Z4, Canada (e-mail: emtiyaz@gmail.com).

D. N. Dutt is with the Department of Electrical Communication Engineering, the Indian Institute of Science, Bangalore 560012, India (e-mail: dndutt@ece.iisc.ernet.in).

Digital Object Identifier 10.1109/TBME.2007.894827

[1]These are also called TVAR "parameters." To avoid confusion with model parameters we will always use the term "coefficients" for these, and reserve the term "parameters" for model parameters.

sequence of measurements by $Y_{1:T} \equiv \{y_1, \ldots, y_T\}$. We also assume TVAR coefficients to follow a Gauss-Markov process:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t \tag{2}$$

where $\mathbf{x}_t \equiv [a_1^t a_2^t \ldots a_p^t]'$ is the array of TVAR coefficients, $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ is the i.i.d. noise, A is the state transition matrix and Q is a symmetric, positive definite matrix (both of size $p \times p$). These equations can be written as a single state-space model

$$\begin{aligned} \text{Measurement Equation}: y_t &= \mathbf{h}_t'\mathbf{x}_t + v_t \\ \text{State Equation}: \mathbf{x}_{t+1} &= A\mathbf{x}_t + \mathbf{w}_t \end{aligned} \tag{3}$$

where $\mathbf{h}_t \equiv [y_{t-1} y_{t-2} \ldots y_{t-p}]'$ is the vector of the past p measurements. $\mathbf{x}_t$ is called the state of the system. The initial state is assumed to be Gaussian: $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$. For simplicity, the initial state vector and noises are assumed to be independent of each other. All the model parameters together are denoted by $\Theta \equiv \{A, \sigma_v^2, Q, \boldsymbol{\mu}_0, \Sigma_0\}$.

We now compare our model with two previous approaches and show that they are special cases of our model. The first approach is based on an AAR model [8] wherein the RLS algorithm is used to estimate $\mathbf{x}_t$. It is shown in [15] that the model used by the RLS algorithm is a special case of the state-space model given in (3), and can be written as follows:

$$\begin{aligned} y_t &= \mathbf{h}_t'\mathbf{x}_t + v_t \\ \mathbf{x}_{t+1} &= \lambda^{-1/2}\mathbf{x}_t \end{aligned} \tag{4}$$

where λ is the forgetting factor for the RLS algorithm. Rewriting the AAR model as in (4) allows an easy comparison with our model. There are two important differences. First, there is no state noise in this model. Second, matrix A is constrained to a scaled identity matrix which depends on the choice of λ. Note that the only tuning parameter in the AAR model is λ.

The second approach, proposed in [18], uses a random-walk model given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{w}_t. \tag{5}$$

There are two differences again. First, the noise covariance is constrained to a scaled identity matrix: where $Q = \sigma_w^2 I$ (σ_w^2 is a nonnegative real number). Second, A is assumed to be an identity matrix. With these assumptions the only unknown parameter is σ_w^2. However, setting this parameter is even more difficult than λ as its range is not known ($\lambda \in (0, 1)$).

Both the AAR and the random-walk model impose constraints to reduce the number of tuning parameters. There are at least two major consequences because of this. First, the same model is assumed for all elements of the state vector. Second, all the elements are assumed to be independent of each other. These assumptions may deteriorate estimation performance (we will show this in Section IV-A). Another important point to note is that finding values of parameters is difficult even when the number of parameters is small. This is because it is usually done manually through trial-and-error. Most of the time, manual settings give suboptimal solutions and an equally good automatic tuning is always preferred. It is a well-known fact that if any *a priori* knowledge is available, then it should be used in formulation of the model [15]. We propose the use of an EM algorithm which allows model parameters to be estimated using training datasets. We describe the proposed approach in Section III.

III. EM Algorithm-Based Approach

We split ERD estimation into the following three subproblems:

1) estimation of the model parameter Θ;
2) estimation of the TVAR coefficients $\{\mathbf{x}_t\}$;
3) estimation of the ERD given TVAR coefficients.

We first present our solution to 2), followed by 1) and 3).

A. Estimation of TVAR Coefficients

Given the measurement sequence $Y_{1:T}$, we want to find estimates of the TVAR coefficients. For this purpose, we use the Kalman filter [14] which gives the optimal estimate in the mean-square sense (in this section, we assume that the model parameters are available). We use the following definitions for the conditional expectations of the states and the corresponding error covariances:

$$\begin{aligned} \hat{\mathbf{x}}_{t|s} &= E\left(\mathbf{x}_t | Y_{1:s}\right) \\ P_{t_1,t_2|s} &= E\left((\mathbf{x}_{t_1} - \hat{\mathbf{x}}_{t_1|s})(\mathbf{x}_{t_2} - \hat{\mathbf{x}}_{t_2|s})' | \mathbf{Y}_{1:s}\right). \end{aligned} \tag{6}$$

For convenience, when $t_1 = t_2 = t$, $P_{t_1,t_2|s}$ is written as $P_{t|s}$. The state estimate $(\hat{\mathbf{x}}_{t|t}, P_{t|t})$ can be obtained with the Kalman filter, which is given as follows:

$$\hat{\mathbf{x}}_{t|t-1} = A\hat{\mathbf{x}}_{t-1|t-1} \tag{7}$$

$$P_{t|t-1} = AP_{t-1|t-1}A' + Q \tag{8}$$

$$K_t = P_{t|t-1}\mathbf{h}_t(\mathbf{h}_t'P_{t|t-1}\mathbf{h}_t + \sigma_v^2)^{-1} \tag{9}$$

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + K_t(y_t - \mathbf{h}_t'\hat{\mathbf{x}}_{t|t-1}) \tag{10}$$

$$P_{t|t} = (I - K_t\mathbf{h}_t')P_{t|t-1} \tag{11}$$

with the initial condition $\mathbf{x}_{1|0} = \mu_1$ and $P_{1|0} = \Sigma_1$. Here, K_t is called the Kalman gain.

Note that the above Kalman filter is a time-varying filter as $\mathbf{h}_t$ depends on time. Hence, the gain and the error covariance will also vary with time and can not be computed *a priori*, unlike the classical Kalman filter [14]. Hence, it will require more computation than the classical Kalman filter. However, the increase in computation will not be very large as we are dealing with scalar measurements. Another important difference is in the convergence of the filter. As $P_{t|t}$ varies with the measurement sequence, it does not converge to a steady-state value. To monitor convergence we need to compute the expectation of $P_{t|t}$ with Monte Carlo simulations and check if it settles down to a value (see [19] for an example of a time-varying Kalman filter).

If the future measurements $Y_{t+1:T}$ are available, then these can be further used to improve the accuracy of the estimates. The *smoothed* estimates [20] can be obtained as follows:

$$J_t = P_{t|t}A'P_{t+1|t}^{-1} \tag{12}$$

$$\hat{\mathbf{x}}_{t|T} = \hat{\mathbf{x}}_{t|t} + J_t(\hat{\mathbf{x}}_{t+1|T} - \hat{\mathbf{x}}_{t+1|t}) \tag{13}$$

$$P_{t|T} = P_{t|t} + J_t(P_{t+1|T} - P_{t+1|t})J_t'. \tag{14}$$

Note that it is the designer's choice whether to use smoothing equations or not. For example, during an on-line analysis, the

Kalman smoother will give estimates only after the end of the experiment, which may not be acceptable. But for an off-line analysis, getting the estimates after the experiment may not matter.

B. Estimation of the Model Parameters With an EM Algorithm

In this section, we describe the estimation of model parameters with an EM algorithm. The objective is to compute an estimate of Θ given a measurement sequence. For Gaussian models, maximum-likelihood (ML) estimate is an obvious choice [20], which is given as follows: $\hat{\Theta}_{ML} = \arg\max_{\Theta} \log p(Y_{1:T}|\Theta)$, where $p(Y_{1:T}|\Theta)$ is the probability density function of the measurements (also called likelihood). Note that because of the dependence on the states, which are not available, direct maximization is not possible. The problem is to maximize the likelihood with respect to two unknowns: states and model parameters. The expectation-maximization (EM) algorithm takes an iterative approach by first maximizing the likelihood with respect to the states in the E-step, and then maximizing with respect to the parameters in the M-step. The EM algorithm was first introduced in [21], and has been used extensively for model parameter estimation [22]–[24]. The E-step maximum is given by the expected value of the complete log-likelihood function as follows:

$$\mathcal{Q} \equiv E_{X|Y}[\log p(Y_{1:T}X_{1:T}|\Theta)]. \tag{15}$$

The M-step involves the direct differentiation of $\mathcal{Q}$ to find the values of the parameters. These computations are done iteratively and the algorithm is guaranteed to converge [22].

We now describe an EM algorithm for our model. For derivation, we follow the procedure given in [23], and details are given in Appendix A. A summary is given below.

1) E-Step: This step involves the computation of $\mathcal{Q}$ given the measurements $Y_{1:T}$ and an estimate of the model parameter from the previous iteration, $\hat{\Theta}_k$. As shown in Appendix A, $\mathcal{Q}$ depends on the following three quantities:

$$\hat{\mathbf{x}}_{t|T} \equiv E(\mathbf{x}_t|Y_{1:T}) \tag{16}$$

$$S_{t|T} \equiv E(\mathbf{x}_t\mathbf{x}_t'|Y_{1:T}) = P_{t|T} + \hat{\mathbf{x}}_{t|T}\hat{\mathbf{x}}_{t|T}' \tag{17}$$

$$S_{t,t-1|T} \equiv E(\mathbf{x}_t\mathbf{x}_{t-1}'|Y_{1:T}) = P_{t,t-1|T} + \hat{\mathbf{x}}_{t|T}\hat{\mathbf{x}}_{t-1|T}'. \tag{18}$$

The first two quantities can be obtained using the Kalman smoother as described in Section III-A. The last quantity can be obtained as described in [20] with the following equation:

$$P_{t,t-1|T} = J_{t-1}P_{t|T}. \tag{19}$$

$\mathcal{Q}$ is then obtained using (34) given in Appendix A.

2) M-Step: By direct differentiation of $\mathcal{Q}$, we get the following expressions of the model parameter estimates:

$$\hat{A}^{k+1} = \Big(\sum_{t=2}^{T} S_{t,t-1|T}\Big)\Big(\sum_{t=2}^{T} S_{t-1|T}\Big)^{-1} \tag{20}$$

$$\hat{Q}^{k+1} = \frac{1}{T-1}\Big(\sum_{t=2}^{T} S_{t|T} - \hat{A}^{k+1}\sum_{t=2}^{T} S_{t-1,t|T}\Big) \tag{21}$$

$$\hat{\sigma_v^2}^{k+1} = \frac{1}{T}\sum_{t=1}^{T}(y_t^2 - 2\,\mathbf{h}_t'\hat{\mathbf{x}}_{t|T}y_t + \mathbf{h}_t'S_{t|T}\mathbf{h}_t) \tag{22}$$

$$\hat{\boldsymbol{\mu}}_1^{k+1} = \hat{\mathbf{x}}_{1|T} \tag{23}$$

$$\hat{\Sigma}_1^{k+1} = S_1 - \hat{\mathbf{x}}_{1|T}\hat{\mathbf{x}}_{1|T}' \tag{24}$$

where k denotes the current iteration. We denote all these estimates together as $\hat{\Theta}^{k+1}$.

Both E- and M-steps are iterated, and convergence is monitored with the conditional likelihood function, obtained as follows:

$$\log p(Y_{1:T}|\hat{\Theta}^k) = \sum_{t=1}^{T} \log\Big(\mathcal{N}(\mathbf{h}_t'\hat{\mathbf{x}}_{t|t-1}, \mathbf{h}_t'P_{t|t-1}\mathbf{h}_t + \sigma_v^2)\Big). \tag{25}$$

The algorithm is said to have converged if the relative increase in the likelihood at the current time step compared to the previous time is below a certain threshold.

The above algorithm can be easily extended to multiple measurements. Assuming trials to be i.i.d., the Kalman smoother estimates must be averaged over all measurement sequences. Substitution in M-step equations will then give the estimate of the parameters corresponding to the multiple measurements.

There are a few practical issues which need to be addressed when implementing the above algorithm. The first issue is of numerical error. Because of its iterative nature, the algorithm is susceptible to numerical round-off errors and can diverge. To solve the numerical problem, we used a square-root filter [25] implementation. The other issue concerns initialization. Some methods are available for initialization (e.g., subspace identification method in [24], [26]). In this paper we use a simpler method by assuming local stationarity. We divide the dataset into overlapping windows, and for each window, we find $\mathbf{x}_t$ and σ_v^2 using MATLAB's ARYULE function. From these local estimates, we find ML estimates of Q. We set A to identity and the initial state mean and covariance to zero and identity matrix, respectively.

C. Estimation of ERD

In this section, we describe the estimation of ERD using the TVAR coefficient estimates obtained with the Kalman smoother. The approach is motivated by an earlier analysis using an AR spectrum, discussed in [27]. We use a similar method, but with a TVAR spectrum. Given TVAR coefficients, time-varying spectrum estimates can be obtained as follows:

$$H(t,f) = \frac{\hat{\sigma_v}}{|1 - \sum_{i=1}^{p} \hat{a}_t^i e^{-2\pi i f/f_s}|}. \tag{26}$$

Here, f_s is the sampling frequency, $\hat{a}_t^i$ is the i^{th} element of the estimated state-vector, and f is the frequency in the range $[0, f_s/2]$. As ERD is seen only in specific frequency bands, we average the spectrum to get band-power P_B

$$P_B(t) = \sum_{f=f_1}^{f_2} H(t,f)^2 \tag{27}$$

where (f_1, f_2) is the band of interest. The band can be set through visual inspection or by using a threshold. We will show later that a very precise selection of the frequency band is not required, and that a rough setting serves well.

An ERD estimate is then found by computing the relative band-power with respect to a reference window. First, a reference power is obtained by averaging band-power over a time interval, where the ERD pattern is expected to be absent (most probably at the start of the experiment). ERD estimates are then obtained with the following equation:

$$\mathrm{ERD}(t) = \frac{P_B(t) - P_{\mathrm{ref}}}{P_{\mathrm{ref}}} \tag{28}$$

where $P_{\mathrm{ref}} = \sum_{t=T_1}^{T_2} P_B(t)$ is the reference band-power for time T_1 to T_2. The ERD estimates obtained are further smoothed by averaging over a time window. The above procedure is similar to the IV method [7] where ERD estimates are obtained in the time domain by computing the variance of a bandpass filtered EEG. The difference is that the IV method does computation in the time domain, while our method operates in the frequency domain. A careful selection of the frequency band is required for the IV method. We will show in Section IV that our approach does not require such precision for the frequency band.

IV. Results

In this section, we study the effect of model parameter estimation with the EM algorithm. We compare the proposed approach with two previous approaches based on the RLS algorithm and the Kalman smoother (KS) as discussed in [8] and [18], respectively (see Section II for details of these approaches). In the rest of the paper, we will refer to these approaches as RLS and KS while we call our approach EMKS.

A. Simulation Results

We compare the approaches for two criteria relevant to the estimation of ERD: 1) tracking of the TVAR coefficients; 2) spectrum estimation of a nonstationary signal. Note that this evaluation requires time-varying simulation data. To generate a smoothly time-varying signal, we consider nonlinear models. This helps us to study the effect of approximating a nonlinear signal, such as an EEG signal, with a TVAR model. However, a direct comparison of the model parameter estimates is not possible for these cases as the actual model will be nonlinear. Hence, we base our comparison on the performance of a filter using the estimated model.

For the first criteria, we generate a smoothly varying AR(2) process (see [18] for simulation details). The trace of the simulated model root and a typical realization are shown in Fig. 1(a). A signal is generated for 2 s, sampled at 128 Hz and the noise variance is set to 0.2. The model order is set to $p = 2$, equal to the actual model order. The model parameters are estimated with the EM algorithm using a dataset of 100 sequences. The same dataset is used to set parameters for RLS and KS. λ and σ_w^2 are optimized for minimum mean-square error. The optimization results are shown in Fig. 1(a), and the values obtained are $\lambda = 0.898$ and $\sigma_w^2 = 0.037$. TVAR coefficients are then estimated with these parameters.

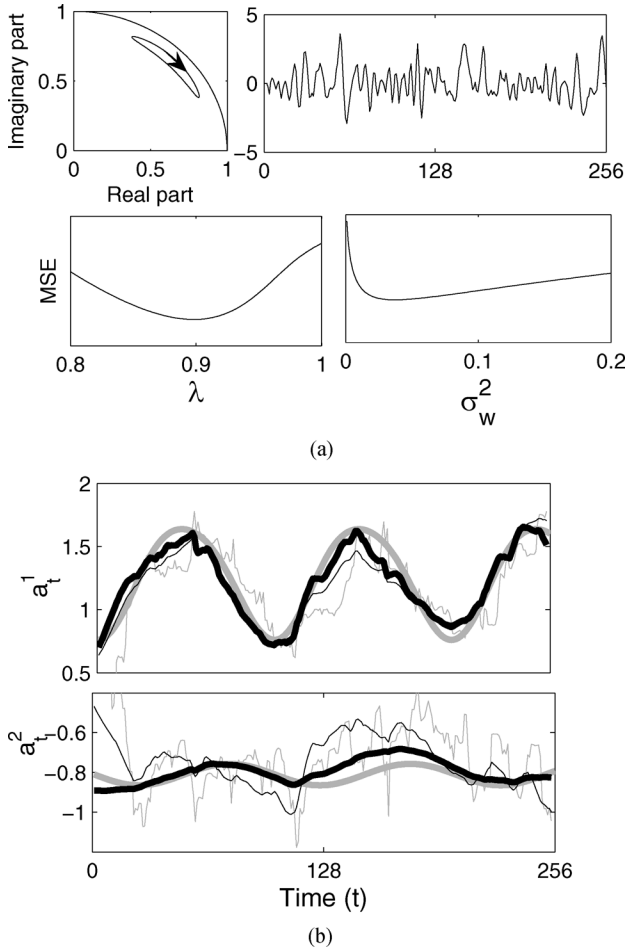

Fig. 1. (a) The root evolution and a typical realization of the AR(2) process along with the optimization of λ and σ_w^2. (b) The TVAR coefficient estimates with EMKS (thick black line), KS (thin black line) and RLS (thin gray line). The actual TVAR coefficients are shown with a thick gray line.

Estimates for one realization are shown in Fig. 1(b). From these figures, it is clear that EMKS performance best. Although RLS and KS track the first coefficient to some extent, they do not track the second coefficient very well. This is because the same model is assumed for both coefficients (see Section II). The optimization function is biased towards the first coefficient as its magnitude is higher, and the estimate for the second coefficient suffers. The model parameters estimated with EM algorithm do not impose any such constraint on the model, and both coefficients have different models. The means and variances of the estimates for 100 realizations are shown in Fig. 2 and they show the same trends for the performance of the algorithms. Hence, we conclude that the better performance of EMKS is due to the better model parameter estimates.

Next we compare the performance for spectrum estimation. For this purpose, we consider a frequency modulated signal given by the following equation:

$$y_t = 5\sin(2\pi f_t t) + u_t \tag{29}$$

where f_t is called the instantaneous frequency (IF) and u_t is zero-mean Gaussian noise with variance σ_u^2. We choose linear frequency modulation: $f_t = 10t$. The signal is generated for

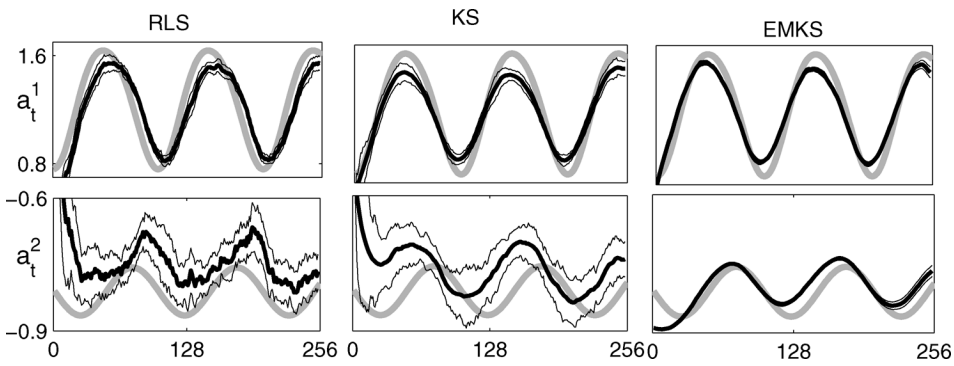

Fig. 2. The average behavior: mean (thick black line) and "mean $\pm 3\sigma$" (thin black line) with RLS, KS, and EMKS. The actual TVAR coefficients are shown with a thick gray line.

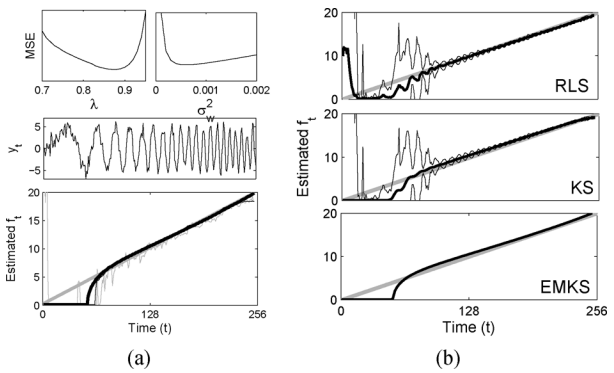

Fig. 3. (a) Optimization of parameters along with a realization of linear FM signal and the IF estimates with EMKS (thick black line), KS (thin black line), RLS (thin gray line), and the actual f_t (thick gray line). (b) Mean (thick black line) and "mean $\pm 3\sigma$ limit" (thin black line) for each method.

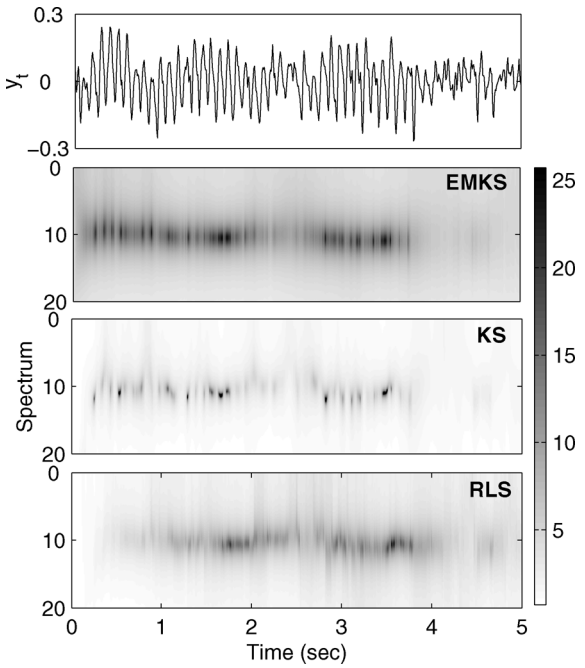

Fig. 4. The time-varying spectral estimates of an EEG signal for a right-hand motor-imagery experiment at position C_3.

2 s, sampled at 128 Hz and the noise variance is set to 1. As the simulated signal contains a single frequency component, we need 2 poles to model it. However, empirical evidence suggests that $p = 4$ is more appropriate for noisy data. Model parameters are obtained with the same method used in the first simulation. Optimized values of λ and σ_w^2 are found to be 0.87 and 0.0006. IF estimates are obtained by picking the peaks of the spectrum obtained using estimated TVAR coefficients. Fig. 3(a) shows the estimates for a realization. It can be seen that EMKS shows smooth convergence, and lowest steady-state error. While performance of RLS is quite poor, KS seems to track as well as EMKS. However, the average performance in Fig. 3(b) shows that variance of the estimates with KS is larger than that of EMKS. In addition, both RLS and KS show oscillation in convergence, while EMKS shows a slightly over-damped response. Results for a fast varying FM signal show similar trends [28].

B. Motor-Imagery EEG Data

In this section, we apply our method to the motor-imagery dataset provided by the Graz University of Technology. A detailed description of the dataset can be found in [29]. In the experiment, the subject's task was to control a bar in one dimension by imagining left- or right-hand movements. The experiment included 7 runs with 40 trials, each of 9 s (hence, 280 trials). Three bipolar EEG signals were measured over positions C_3, C_z, and C_4. The first 2 s were quiet and at $t = 2$ s, an acoustic stimulus indicated the beginning of the trial. A cross ("+") was displayed for 1 s. Then, at $t = 3$ s, an arrow pointing either to the left or right was displayed as a cue stimulus. The subject was asked to imagine moving the bar in the direction of the cue. The number of left-hand cues were equal to the number of right-hand cues. For our analysis, we use a model order of $p = 5$, and set $\lambda = 0.97$ and $\sigma_w^2 = 0.001$ for RLS and KS, respectively. These parameters are chosen to the best of our ability based on visual inspection. For the EM algorithm, model parameters are estimated with 50 trials. For single-trial results, the chosen dataset does not belong to the training dataset. However, for average behavior the training dataset is included, because there would be too little data otherwise.

Fig. 4 shows the time-varying spectrum estimates for the first 5 s of a trial. This trial shows a decrease in activity between 2 and 3 s and then after 4 s. We can clearly see that the EMKS estimates capture these patterns accurately. Although KS detects the decrease in activity, the estimates have noisy peaks and

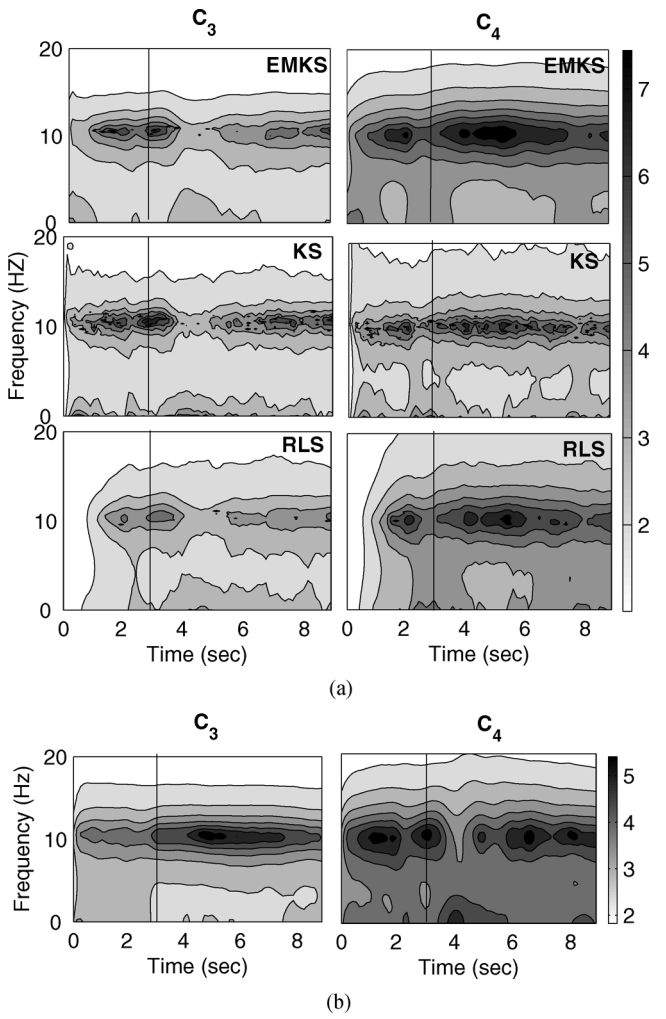

Fig. 5. The average spectrum for (a) the right-hand motor-imagery EEG data and (b) the left-hand motor-imagery data. The cue is indicated with a vertical line at $t = 3$ s.

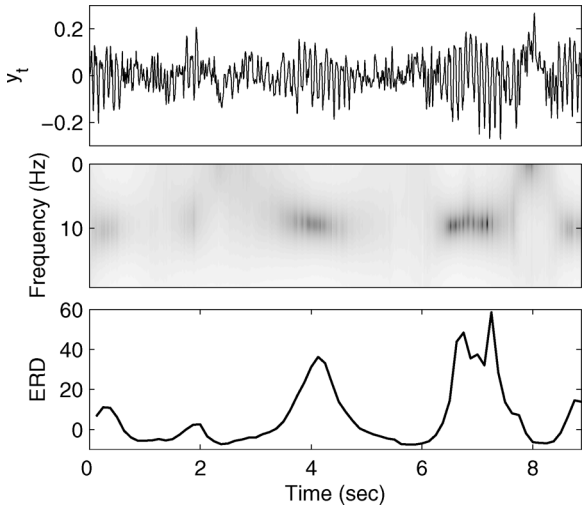

Fig. 6. A motor-imagery trial (top) chosen from the right-hand movement experiment at position C_3, along with the estimated spectrum and ERD.

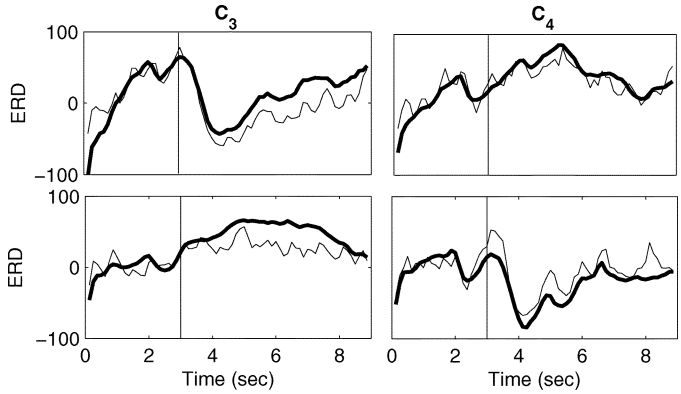

Fig. 7. ERD estimates with EMKS (thick line) and IV (thin line) for imagination of the right-hand (top) and the left-hand (bottom) movements at positions C_3 and C_4. The event time is shown with a vertical line at $t = 3$ s.

are not smooth. RLS also does not estimate the pattern properly. Note that all of these estimates show activity in the alpha band (8–12 Hz) which is expected for a motor-imagery experiment. Fig. 5(a) shows the mean of the spectrum for all 70 trials of right-hand datasets at positions C_3 and C_4. We can see that for all the methods there is a significant decrease in activity in the alpha band-power at position C_3 after the cue is presented, while there is no such pattern at position C_4. Hence, on average, the estimates show ERD. Comparisons between methods show the same trend as performance for a single trial: EMKS estimates are smooth, while KS and RLS are noisy. In addition, EMKS and KS both show better convergence than RLS. The poor convergence may affect the ERD estimates. This is because the reference level is obtained using initial estimates. For completeness, Fig. 5(b) shows the EMKS spectrum estimates for left-hand data. The ERD patterns are reversed here, estimates for position C_4 show ERD, while those for position C_3 do not. This clearly demonstrates the expected hemispherical asymmetry due to the motor-imagery experiment.

We now discuss the results for ERD estimation. Fig. 6 shows a trial of right-hand data at position C_3, its spectrum, and ERD estimates. ERD estimates are obtained with the following settings: the frequency band for band-power is chosen to be 8–15 Hz, reference power is obtained by averaging the band-power from 0 to 2 s, and ERD estimates are smoothed over a window length of 16 samples. We observe that the derived ERD pattern is in accordance with activity changes in the spectrum. However, because of high variability between trials, it is difficult to draw any conclusion about the general behavior of ERD estimates from single-trial estimates. To prove the consistency of the ERD estimate on average, we compare it with the standard IV method [7]. Note that the IV method gives a good estimate of ERD, but is sensitive to the selection of the frequency band. With visual inspection, a frequency band for the IV method is chosen to be 9–12Hz. A dataset of 70 trials is used for estimation. Referencing and smoothing are done with the same parameters used for EMKS. ERD estimates are shown in Fig. 7. We can see that both estimates show similar patterns. Also, both right- and left-hand data show desynchronization. Note that the frequency range chosen for EMKS is quite large (8–15 Hz), and does not have to be chosen very precisely. This is due to a better time-frequency resolution of spectrum estimates with EMKS as compared to other methods.

Finally, we compare the classification accuracy obtained using ERD estimates. We use a linear discrimination method

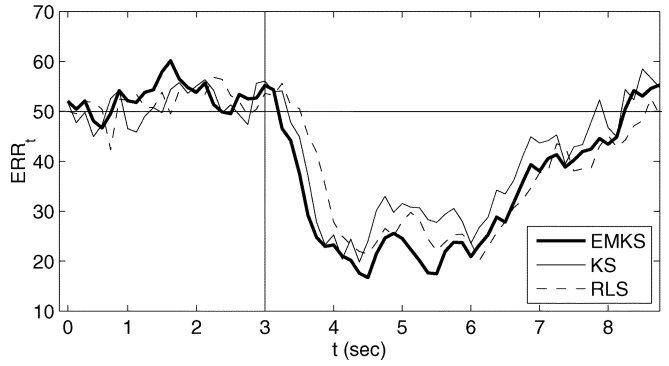

Fig. 8. Time course of smoothed error rate ERR_t with EMKS, KS, and RLS algorithms.

similar to one described in [5]. Training data consists of 140 trials (70 each for right- and left-hand imagery) at positions C_3 and C_4. Four sets of model parameters are estimated with the EM algorithm corresponding to left- and right-hand at positions C_3 and C_4. TVAR coefficients are obtained with these models, and a feature vector is formed as follows:

$$\mathbf{d}_t = \begin{bmatrix} \mathbf{x}_{R3}^t - \mathbf{x}_{L3}^t \\ \mathbf{x}_{R4}^t - \mathbf{x}_{L4}^t \end{bmatrix}. \tag{30}$$

Here, $\mathbf{x}_{R3}^t$(or $\mathbf{x}_{L3}^t$) denotes the TVAR coefficients of the signal at position C_3 using the right-hand (or left-hand) data model. Similar notations are used for other variables. A distance D can be computed for a signal, using a linear discrimination function as follows:

$$D_t = \mathbf{w}_t^T \mathbf{d}_t - w_0 \tag{31}$$

where $\mathbf{w}_t$ is the weight vector and w_0 is the offset. $D_t > 0\,(< 0)$ means that the signal is classified as a left-hand (right-hand) trial. $\mathbf{w}_t$ and w_0 are found with a support vector machine (SVM) [30]. A test dataset of 140 trials is classified using the above discrimination function, and a ten-times tenfold cross-validation is applied every 125 ms [5]. A time-course of error ERR_t is then obtained. Fig. 8 shows the ERR_t smoothed over a window of 16 samples. As expected before the event, the error rate is close to 50%, and it drops after the cue is presented. The lowest classification error obtained is 15.4% at time point 4.6 s with EMKS, 19.6% at 4.6 s with KS and 20.8 at 6.1 s for RLS. We see that EMKS gives the lower error rate. Also note that the lowest error is obtained at a later time by using RLS as compared to EMKS and KS. This is because of the tracking lag introduced by the RLS algorithm.

V. Conclusion

In this paper, we propose an EM algorithm based Kalman smoother approach for ERD estimation. Previous approaches impose several constraints on the AR model to make model parameter setting easier. We show that such constraints may deteriorate estimation performance. The proposed method does not require any constraints or manual setting. In addition, optimal estimates in the ML sense are obtained. Another advantage of the proposed approach is that the Kalman smoother can be used for coefficient estimation with these estimated model parameters. This further improves estimation performance compared to RLS-based approaches. We show that the proposed approach significantly improves tracking and spectrum estimation performance. Application to real world EEG data shows that the spectrum estimates are smooth and show good convergence. Useful ERD patterns are obtained with the proposed method for ERD estimation. The advantage is that the method does not require a careful selection of the frequency band, in contrast to previous approaches. This paper also confirms the hemispherical asymmetry obtained with ERD, and supports its use for BCIs.

Although the use of the EM algorithm is promising, there are a few issues. The first one is related to convergence. We found that convergence becomes very slow after a few cycles, and training takes a lot of time. Also, to obtain a value close to the true model parameter, a large dataset is necessary. Further work on increasing the rate of convergence could be useful. The second issue is about the validation of the above results. The proposed approach shows very clear results for the dataset considered. Although we do not expect a poor performance on other datasets, validation with more datasets and multiple subjects will confirm our method's applicability in a practical BCI system.

Appendix
EM Algorithm: Log-Likelihood Derivation and M-Step

The joint probability distribution of $X_{1:T}, Y_{1:T}$ can be written as

$$p(X_{1:T}, Y_{1:T}|\Theta) = p(\mathbf{x}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})\prod_{t=1}^{T} p(y_t|\mathbf{x}_t, \mathbf{h}_t). \tag{32}$$

Taking log and expectation, we get the expectation of joint log-likelihood with respect to the conditional expectation

$$\begin{aligned}\mathcal{Q} &= E_{X|Y}[\log p(X_{1:T}, Y_{1:T}|\Theta)] \qquad (33)\\ &= -\frac{T}{2}\ln {\sigma_v}^2 - \frac{1}{2{\sigma_v}^2}\sum_{t=1}^{T}[{y_t}^2 - 2\mathbf{h}_t'\hat{\mathbf{x}}_t y_t + \mathbf{h}_t' S_{t|T}\mathbf{h}_t] \\ &\quad - \frac{1}{2}\sum_{t=2}^{T} trace[Q^{-1}(S_{t|T} - AS_{t-1,t|T} - S_{t,t-1|T}A' \\ &\quad + AS_{t-1|T}A')] \\ &\quad - \frac{1}{2}trace[V_1^{-1}(P_{1|T} - 2\pi_1\hat{\mathbf{x}}_1{}' + \pi_1\pi_1{}')] - \frac{1}{2}\ln|V_1| \\ &\quad - \frac{T-1}{2}\ln|Q| - \frac{(p+1)T}{2}\ln 2\pi. \qquad (34)\end{aligned}$$

For the M-step, we take the derivative of $\mathcal{Q}$ with respect to each model parameter, and set it to zero to get the estimate, e.g., an update for A can be found as

$$\frac{\partial \mathcal{Q}}{\partial A} = -\frac{1}{2}\sum_{t=2}^{T}\Big[-2S_{t,t-1|T} + 2AS_{t-1|T}\Big] = 0 \tag{35}$$

which gives

$$A^{k+1} = \Big(\sum_{t=2}^{T} S_{t,t-1|T}\Big)\Big(\sum_{t=2}^{T} S_{t-1|T}\Big)^{-1}. \tag{36}$$

Updates for other parameters can be obtained similarly.

ACKNOWLEDGMENT

The authors would like to thank Dr. C. Sekhar from the Indian Institute of Science, Bangalore, for his valuable suggestions for simulations, Prof. Z. Gharhamani from the University of Cambridge for the EM algorithm code, Graz BCI researchers for providing the dataset used in this paper, and the reviewers for their valuable suggestions. They would also like to thank L. Rizoli, L. Burbidge, and U. Schmidt from the University of British Columbia, Vancouver, BC, Canada, for proofreading this paper.

REFERENCES

[1] G. Pfurtscheller and F. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Electroencephalogr. Clin. Neurophysiol.*, vol. 110, pp. 1842–1857, 1999.

[2] G. Pfurtscheller and C. Neuper, "Event-related synchronization of mu rhythm in the EEG over the cortical hand area in man," *Neurosci. Lett.*, vol. 174, pp. 93–96, 1994.

[3] G. Pfurtscheller and C. Neuper, "Motor imagery activates primary sensorimotor area in man," *Neurosci Lett.*, vol. 239, pp. 65–68, 1997.

[4] G. Pfurtscheller, C. Neuper, H. Ramoser, and J. Muller-Gerking, "Visually guided motor imagery activates sensotimotor areas in humans," *Neurosci. Lett.*, vol. 269, pp. 153–156, 1999.

[5] G. Pfurtscheller, C. Neuper, A. Schlogl, and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *IEEE Trans. Rehabil. Eng.*, vol. 6, pp. 316–325, 3, Sep. 1998.

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

[7] J. Kalcher and G. Pfurtscheller, "Discrimination between phase-locked and non-phase-locked event-related EEG activity," *Electroencephalogr. Clin. Neurophysiol.*, vol. 94, pp. 381–483, 1995.

[8] A. Schlogl, D. Flotzinger, and G. Pfurtscheller, "Adaptive autoregressive modeling used for single-trial EEG classification," *Biomed. Technik*, vol. 42, pp. 162–167, 1997.

[9] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[10] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden Markov models for online classification of single trial EEG data," *Pattern Recognit. Lett.*, vol. 22, pp. 1299–1309, 2001.

[11] G. Pfurtscheller, J. Kalcher, C. Neuper, D. Flotzinger, and M. Pregenzer, "On-line EEG classification during externally paced hand movements using a neural network-based classifier," *Electroencephalogr. Clin. Neurophysiol.*, vol. 99, pp. 416–425, 1996.

[12] M. Arnold, W. H. R. Miltner, H. Witte, R. Bauer, and C. Braun, "Adaptive AR modeling of nonstationary time-series by means of Kalman filtering," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 5, pp. 553–562, May 1998.

[13] M. J. Cassidy and W. D. Penny, "Bayesian nonstationary autoregressive models for biomedical signal analysis," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 10, pp. 1142–1152, Oct. 2002.

[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. Am. Soc. Mech. Eng.,* , ser. D–J. Basic Eng., vol. 82D, pp. 35–45, Mar. 1960.

[15] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ: Pearson Education, 2002.

[16] H. E. Rauch, "Soltutions to the linear smoothing problem," *IEEE Trans. Autom. Control*, vol. 8, pp. 371–372, 1963.

[17] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *J. Am. Inst. Aeronautics Astronautics*, vol. 3, no. 8, pp. 1445–1450, 1965.

[18] M. Tarvainen, J. Hiltunen, P. Ranta-aho, and P. Karjalainen, "Estimation of nonstationary EEG with Kalman smoother approach: An application to event-related synchronization (ERS)," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 3, pp. 516–524, Mar. 2004.

[19] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. Jordan, and S. Sastry, "Kalman filtering with intermittent observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1453–1464, Sep. 2004.

[20] R. Shumway and D. Stoffer, *Time Series Analysis and Its Applications*. New York: Springer, 2000.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incompleter data via EM algorithm," *J. Roy. Statist. Soc.* ser. B–Methodological, vol. 39, pp. 1–38, 1998.

[22] R. Shumway and D. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[23] Z. Ghahramani and G. E. Hinton, "Parameter Estimation for Linear Dynamical Systems," Tech. Rep. Univ. Toronto, Dept. Comput. Sci., Toronto, ON, Canada, 1996.

[24] H. Raghavan, A. Tangirala, R. Gopaluni, and S. Shah, "Identification of chemical processes with irregular output sampling," *Control Eng. Practice*, Jan. 2005.

[25] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

[26] S. Kung, "A new identification and model reduction algorithm via singular value decomposition," presented at the 12th Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, 1978.

[27] G. Florian and G. Pfurtscheller, "Dynamic spectral analysis of event-related EEG data," *Electroencephalogr. Clin. Neurophysiol.*, vol. 95, pp. 393–396, 1995.

[28] M. E. Khan and D. N. Dutt, "Expectation-maximization (EM) algorithm for instantaneous frequency estimation with Kalman smoother," presented at the 12th Eur. Signal Processing Conf. (EUSIPCO), Vienna, Austria, 2004.

[29] B. Blankertz, K. R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1044–1051, Jun. 2004.

[30] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience, 1973.

Mohammad Emtiyaz Khan received the M.Sc. degree in 2004 from the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India. Currently, he is a graduate student in the Department of Computer Science at the University of British Columbia.

He worked for two years in the research and technology group at Honeywell Technology Solutions Lab, Bangalore. His research interests are in the area of statistical methods applied to machine-learning, control, imaging, and biomedicine.

Deshpande Narayana Dutt received the Ph.D. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India.

Currently, he is an Associate Professor at the same institute. He has worked in the areas of acoustics and speech signal processing. His research interests are in the area of digital signal processing applied to the analysis of biomedical signals, in particular brain signals. He has published a large number of papers in this area in leading international journals.