A MULTIMODAL VARIATIONAL APPROACH TO LEARNING AND INFERENCE IN SWITCHING STATE SPACE MODELS

Leo J. Lee[1,2]*, Hagai Attias*[2]*, Li Deng*[2] *and Paul Fieguth*[3]

University of Waterloo
[1]Electrical & Computer Engineering
[3]Systems Design Engineering
Waterloo, ON, N2L 3G1
Canada

[2]Microsoft Corporation
Microsoft Research
One Microsoft Way
Redmond, WA 98052-6339
USA

ABSTRACT

An important general model for discrete-time signal processing is the switching state space (SSS) model, which generalizes the hidden Markov model and the Gaussian state space model. Inference and parameter estimation in this model are known to be computationally intractable. This paper presents a powerful new approximation to the SSS model. The approximation is based on a variational technique that preserves the multimodal nature of the continuous state posterior distribution. Furthermore, by incorporating a windowing technique, the resulting EM algorithm has complexity that is just linear in the length of the time series. An alternative Viterbi decoding with frame-based likelihood is also presented which is crucial for the speech application that originally motivates this work. Our experiments focus on demonstrating the effectiveness of the algorithm by extensive simulations. A typical example in speech processing is also included to show the potential of this approach for practical applications.

1. INTRODUCTION

The switching state space (SSS) model is a probabilistic dynamic system which combines discrete and continuous dynamics. It generalizes the hidden Markov model (HMM) and the linear state space model. Whereas the state space model describes an observed time series in terms of a continuous hidden state vector whose dynamics is specified by the dependence of the current state on the previous one, in the SSS those dynamics depend on additional states which are discrete. Hence, the dynamics generally vary with time, producing a powerful model with applications in many domains, such as speech processing [1], control [2], machine vision [3] and financial analysis [4]. In the machine learning community, the SSS model belongs to a class of models termed conditional linear Gaussian (CLG) models[5], which has also been attracting interest recently.

Whereas inference and parameter estimation in HMM and the state space model can be done exactly using the EM algorithm (known as Baum-Welch for the former and Kalman-Rauch for the latter), it is well known that inference in SSS is computationally intractable [6]. More generally, it has been shown that inference in CLG models is NP-hard [5]. Here a new approximation scheme for the SSS model is presented, based on a multimodal variational technique. Variational methods are first applied to similar models by Ghahramani and Hinton [6], and some important differences between their work and ours will be pointed out as we develop our approach in the following sections. This paper builds on previously-published theoretical results [7][1]; here we will develop important extensions, backed up by extensive simulation results.

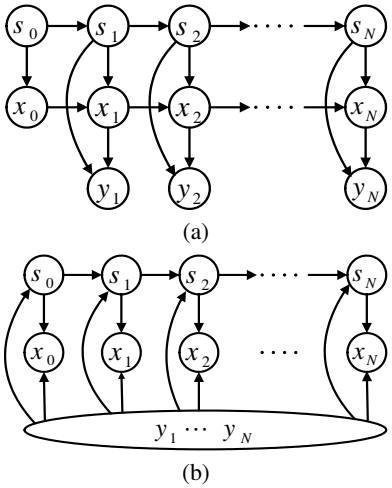

Fig. 1. The model (a) and the variational posterior (b) represented as Bayesian networks.

The remainder of the paper is organized as follows: The model used in this work is described in Section 2, followed by some details of the algorithm development in Section 3. Section 4 shows the effectiveness of the algorithm by simulation examples, and finally Section 5 concludes the paper by a typical speech processing example and a brief discussion.

2. THE SWITCHING STATE SPACE MODEL

We start with the definition of the switching state space (SSS) model. This is a probabilistic model that describes multivariable time series data in terms of two types of hidden variables (also termed states): discrete and continuous. The hidden variables form two first-order Markov chains, where the continuous chain is con-

[1]Two methods were developed in [7]: one is the progenitor to the present paper; the other, on which all the experimental results in [7] were based, is unrelated and not further pursued for its lack of efficiency as well as accuracy.

ditioned on the discrete one. Let $\mathbf{y}_n$ denote the observed data at time $n = 1 : N$, and s_n and $\mathbf{x}_n$ denote the discrete and continuous hidden states, respectively, at that time. The discrete states may assume one of S values, where $s = 1, ..., S$. We have

$$\begin{aligned} p(s_n = s \mid s_{n-1} = s') &= \pi_{ss'} \ , \\ p(\mathbf{x}_n \mid s_n = s, \mathbf{x}_{n-1}) &= \mathcal{N}(\mathbf{x}_n \mid \mathbf{A}_s \mathbf{x}_{n-1} + \mathbf{a}_s, \mathbf{B}_s^{-1}), \\ p(\mathbf{y}_n \mid s_n = s, \mathbf{x}_n) &= \mathcal{N}(\mathbf{y}_n \mid \mathbf{C}_s \mathbf{x}_n + \mathbf{c}_s, \mathbf{D}_s^{-1}), \end{aligned} \quad (1)$$

and the initial states at $n = 0$ are

$$p(s_0 = s) = \pi_s^0 \ , \qquad p(\mathbf{x}_0 \mid s_0 = s) = \mathcal{N}(\mathbf{x}_1 \mid \mathbf{a}_s^0, \mathbf{B}_s^0) \ . \quad (2)$$

The full joint distribution of the model is given by

$$\begin{aligned} p(y_{1:N}, x_{0:N}, s_{0:N} \mid \Theta) = \prod_{n=1}^{N} & p(\mathbf{y}_n \mid \mathbf{x}_n, s_n) p(\mathbf{x}_n \mid s_n, \mathbf{x}_{n-1}) \\ & \cdot p(s_n \mid s_{n-1}) p(\mathbf{x}_0 \mid s_0) p(s_0) \ , \end{aligned} \quad (3)$$

where the parameters are

$$\Theta = \{\pi_{ss'}, \pi_s^0, \mathbf{A}_s, \mathbf{a}_s, \mathbf{a}_s^0, \mathbf{B}_s, \mathbf{B}_s^0, \mathbf{C}_s, \mathbf{c}_s, \mathbf{D}_s\} \ . \quad (4)$$

Fig. 1 shows the graphical representation of the SSS model, where the conditional independence relations are illustrated clearly. Notice that our model implicitly forces a continuity constraint[2] on $\mathbf{x}_{1:N}$ so that it fits nicely into the Bayesian framework when $\mathbf{x}_{1:N}$ is treated as a smoothness prior for $\mathbf{y}_{1:N}$ [8]. Such a continuity constraint is not present in many other SSS models, such as the one in [6] where there are M underlying linear dynamic processes but only one of them is observed at a given time.

EM parameter estimation in probabilistic models generally involves iterating between an E-step, which updates the posterior distribution over the hidden states (and the moments of the posterior, a.k.a. sufficient statistics), and an M-step, which updates the model parameters. As is well-known, in the SSS model the E-step is computationally intractable, because the exact computation of the posterior distribution requires summing over all possible configurations of $s_{1:N}$, whose number is $\mathcal{O}(e^N)$. In the following section we derive a new approximation scheme for this posterior.

3. A MULTIMODAL VARIATIONAL APPROXIMATION

In the variational approach we approximate the exact posterior $p(s_{1:N}, \mathbf{x}_{1:N} \mid \mathbf{y}_{1:N})$ by a distribution with a tractable structure, denoted by q. Here we choose the following partially factorized structure shown graphically in Fig. 1:

$$\begin{aligned} & p(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) \approx q(s_{0:N}, \mathbf{x}_{0:N} \mid \mathbf{y}_{1:N}) \\ & = \prod_{n=1}^{N} q(\mathbf{x}_n \mid s_n) q(s_n \mid s_{n-1}) \cdot q(x_0 \mid s_0) q(s_0) \ . \end{aligned} \quad (5)$$

As is costumary to the notation in variational methods, the data dependence of the q's is omitted but always implied.

Whereas q is an approximation, it preserves important features of the exact posterior. In particular, (1) it incorporates temporal correlations via the Markov chain structure of $q(s_{1:N})$, (2) it incorporates correlations between the continuous and discrete states, and most significantly (3) it incorporates multimodality of the continuous states since $q(\mathbf{x}_n) = \sum_s q(\mathbf{x}_n \mid s_n = s) q(s_n = s)$ is a mixture distribution. On the other hand, it does not directly incorporate temporal correlations among $\mathbf{x}_n$'s; those are introduced indirectly via the variational equations below. Previous work on variational approach to such models [6] uses the factorized form $q = q(\mathbf{x}_{1:N}) q(s_{1:N})$. Whereas that form does incorporate temporal correlations among the $\mathbf{x}_n$, it results in a $q(\mathbf{x}_{1:N})$ which is a Gaussian, and thus unimodal. Nevertheless, such an approximation can also be applied to our model and a detail comparison study will be published separately.

To derive q, we start with the functional

$$\begin{aligned} \mathcal{F}[q] = \sum_{s_{1:N}} \int d\mathbf{x}_{1:N} \ & q(s_{1:N}, \mathbf{x}_{1:N}) \cdot \\ & [\log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, s_{1:N}) - \log q(s_{1:N}, \mathbf{x}_{1:N})] \end{aligned} \quad (6)$$

and optimize it w.r.t. q, under the restriction that q has the above structure. This is done by setting the functional derivative $\delta \mathcal{F} / \delta q(\mathbf{x}_n \mid s_n)$ and the ordinary derivative $\partial \mathcal{F} / \partial q(s_n \mid s_{n-1})$ to zero, for each n, and solving the resulting recursive equations, equivalent to minimizing the KL distance of q to the exact posterior. Here we present the results only, rather than the full derivation, due to space limitations.

First, notice that the functional form of $q(\mathbf{x}_n \mid s_n)$ was not specified in advance. The optimal form, resulting from a free form optimization, turns out to be a Gaussian with state dependent mean $\boldsymbol{\rho}_{s,n}$ and precision $\boldsymbol{\Gamma}_{s,n}$,

$$q(\mathbf{x}_n \mid s_n = s) = \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\rho}_{s,n}, \boldsymbol{\Gamma}_{s,n}^{-1}) \ . \quad (7)$$

Next, we define the following variational posteriors,

$$\begin{aligned} \gamma_{s,n} &= q(s_n = s) \ , \\ \eta_{s's,n} &= q(s_n = s \mid s_{n-1} = s') \ , \\ \bar{\eta}_{s's,n} &= q(s_n = s \mid s_{n+1} = s') = \frac{\eta_{ss',n+1} \gamma_{s,n}}{\gamma_{s',n+1}} \ . \end{aligned} \quad (8)$$

We also introduce the notation $E_{s,n}$, denoting state-conditioned averaging, via

$$E_{s,n} g(\mathbf{x}_n) = \int d\mathbf{x}_n \ q(\mathbf{x}_n \mid s_n = s) g(\mathbf{x}_n) \ , \quad (9)$$

where g is an arbitrary function.

Computing $q(\mathbf{x}_n \mid s_n)$: The precision matrix of $q(x_n \mid s_n)$ is given by

$$\boldsymbol{\Gamma}_{s,n} = \mathbf{C}_s^T \mathbf{D}_s \mathbf{C}_s + \mathbf{B}_s + \sum_{s'} \eta_{ss',n+1} \mathbf{A}_{s'}^T \mathbf{B}_{s'} \mathbf{A}_{s'} \ . \quad (10)$$

The mean satisfies the linear equation

$$\begin{aligned} \boldsymbol{\Gamma}_{s,n} \boldsymbol{\rho}_{s,n} = \mathbf{C}_s^T \mathbf{D}_s (\mathbf{y}_n - \mathbf{c}_s) + \mathbf{B}_s (\mathbf{A}_s \sum_{s'} \bar{\eta}_{ss',n-1} \boldsymbol{\rho}_{s',n-1} + \mathbf{a}_s) \\ + \sum_{s'} \eta_{ss',n+1} \mathbf{A}_{s'}^T \mathbf{B}_{s'} (\boldsymbol{\rho}_{s',n+1} - \mathbf{a}_{s'}) \ . \end{aligned} \quad (11)$$

Notice that a brute force solution of the last equation by matrix inversion has complexity $\mathcal{O}((NS)^3)$, where S is the number of states. Below we discuss an efficient solution technique using overlapping windows, whose complexity is significantly lower.

[2] Further smoothness can be enforced by constraining the first and/or higher orders of derivatives to be continuous as well.

Computing $q(s_n \mid s_{n-1})$: We introduce the quantity $z_{s,n}$ for each time n and state s. It turns out to be the normalization factor of the posterior transition probability $\eta_{s's,n} = q(s_n \mid s_{n-1})$. These probabilities are computed recursively by a backward pass as follows. First, we initialize $z_{s,N+1} = 1$ for all s. Next, for $n = N, ..., 1$ we compute

$$\eta_{s's,n} = \frac{1}{z_{s,n}} e^{f_{s's,n}} z_{s',n+1} \ , \quad z_{s,n} = \sum_{s'} e^{f_{s's,n}} z_{s',n+1} \ , \quad (12)$$

and for $n = 0$ we compute

$$\gamma_{s,0} = \frac{1}{z_0} e^{f_{s,0}} z_{s,1} \ , \quad z_0 = \sum_{s} e^{f_{s,0}} z_{s,1} \ . \quad (13)$$

The quantities $f_{s's,n}$ are given by

$$\begin{aligned} f_{s's,n} = & E_{s',n} \left[\log p(\mathbf{y}_n \mid \mathbf{x}_n, s_n = s') - \log q(\mathbf{x}_n \mid s_n = s') \right] \\ & + E_{s',n} E_{s,n-1} \log p(\mathbf{x}_n \mid \mathbf{x}_{n-1}, s_n = s') \\ & + \log p(s_n = s' \mid s_{n-1} = s) \ , \end{aligned} \quad (14)$$

where the averages are straightforward to compute analytically but too lengthy to show due to the space constraint here. A similar result is obtained for $f_{s,0}$.

Computing $q(s_n)$: In addition to the posterior transition probabilities, estimation of the model parameters Θ (M-step) also requires the posterior marginals, which are computed recursively by a forward pass for $n = 1, ..., N$,

$$\gamma_{s,n} = \sum_{s'} \eta_{s's,n} \gamma_{s',n-1} \ . \quad (15)$$

Overlapping windows. As mentioned above, a direct solution of Eq. (11) for the whole N-point-long time series has complexity which is cubic in N and is thus very expensive. Although a sparse matrix technique has been applied before [7], the complexity still doesn't scale well enough to handle large N. Here we use a procedure motivated by the following observation. Assume the first N_1 data points $\mathbf{y}_{1:N_1}$ have been observed and the posterior over $s_{1:N_1}, \mathbf{x}_{1:N_1}$ has been computed. After observing the next data point, the posterior is recomputed. Now, if N_1 is sufficiently large, the effect of this new data point on the posterior over the early states $s_1, \mathbf{x}_1$ may be vanishingly small.

Hence, we proceed to solve Eq. (11) as follows. We apply an M-point-long window to the time series and repeatedly increment its start point by J points. At each increment, we solve Eq. (11) for the data inside the window to obtain the sufficient statistics. This procedure has complexity $\mathcal{O}((MS)^2 N/J)$, which is linear in N. Generally, we choose the smallest M that produces a desired accuracy; note that this value depends on the temporal structure of the time series, but is independent of its length N.

E-step: sufficient statistics. As usual, the variational equations above are coupled, with the equations for $\boldsymbol{\rho}_{s,n}, \boldsymbol{\Gamma}_{s,n}$ depend on $\eta_{s's,n}$, $\gamma_{s,n}$ and vice versa. These equations are solved iteratively starting from a random or more suitable initialization if available. The solution is the set of sufficient statistics

$$\varphi = \{\rho_{s,n}, \Gamma_{s,n}, \eta_{ss',n}, \gamma_{s,n}\} \quad (16)$$

which are moments of the variational posterior.

M-step: parameter estimation. Given the sufficient statistics φ, the derivation of the M-step is achieved by taking derivatives of $\mathcal{F}$ w.r.t. the model parameters (details omitted).

Recovering hidden states. It is often needed to estimate the state sequences $\hat{\mathbf{x}}_{1:N}$ and $\hat{s}_{1:N}$ from the data. For the continuous states we use the MMSE estimator, defined w.r.t. the variational posterior, to obtain

$$\hat{\mathbf{x}}_n = \int dx_{1:N} \ q(\mathbf{x}_{1:N}) \mathbf{x}_n = \sum_s \gamma_{s,n} \boldsymbol{\rho}_{s,n} \ . \quad (17)$$

For the discrete states the Viterbi algorithm can be applied based on the variational posterior η, e.g., the initialization and induction equations for the scoring turn out to be:

$$V_1(s) = \max_{s'} [\pi_{s'0} \eta_{s's,1}], \ V_n(s) = \max_{s'} [V_{n-1}(s') \eta_{s's,n}]. \quad (18)$$

Interestingly, it can be shown that identical inference can also be obtained by applying the Viterbi algorithm on f, where the initialization and induction equations are:

$$V_1(s) = \max_{s'} [\log \pi_{s'0} + f_{s's,1}], \quad (19)$$

$$V_n(s) = \max_{s'} [V_{n-1}(s') + f_{s's,n}]. \quad (20)$$

Here f plays the same role as the frame-based likelihood in a standard HMM. The alternative Viterbi algorithm based on f is crucial for applications where external sources of information needs to be included when recovering the discrete hidden states (e.g., language model score in speech recognition). The external information can be simply added as an extra term in (19) and (20).

4. SIMULATION EXPERIMENTS

Extensive simulations have been carried out to verify the correctness and effectiveness of the algorithm. The example used here has four discrete states with the following model parameters:

$A_1 = 0.7, \ a_1 = 0.6, \ B_1 = 1000, \ C_1 = [0.8\,0.3\,0.2]^T, \ c_1 = -[3\,2\,1]^T,$
$A_2 = 0.8, \ a_2 = 0.5, \ B_2 = 4000, \ C_2 = [1.0\,0.2\,0.1]^T, \ c_2 = [-1\,0\,1]^T,$
$A_3 = 0.9, \ a_3 = 0.18, \ B_3 = 694.4, \ C_3 = [0.5\,0.4\,0.2]^T, \ c_3 = [0\,1\,2]^T,$
$A_4 = 0.6, \ a_4 = 0.88, \ B_4 = 1563, \ C_4 = [0.1\,0.7\,0.8]^T, \ c_4 = [1\,2\,3]^T,$

where $\mathbf{D}$ is 100 times the identity matrix for all four states and π is uniform. Since the E step is an iterative process itself, we have to initialize $\boldsymbol{\rho}$ and $\boldsymbol{\Gamma}$ or η and γ, and a suitable initialization scheme can be very application dependent. In our simulation the $\boldsymbol{\rho}$'s are initialized to be a weighted sum of the pseudo-inverse estimation from $\mathbf{y}$ and the target value of $\mathbf{x}$ for each s[3], and the weights are determined by the ratio of $\|\mathbf{B}_s\|$ and $\|\mathbf{D}_s\|$. The $\boldsymbol{\Gamma}$'s are initialized by

$$\boldsymbol{\Gamma}_{s,n} = \mathbf{C}_s^T \mathbf{D}_s \mathbf{C}_s + \mathbf{B}_s + \mathbf{A}_s^T \mathbf{B}_s \mathbf{A}_s. \quad (21)$$

Fig. 2 tests the sensitivity of algorithm inference (E step) to levels of process and observation noise, which are measured by the precision matrices $\mathbf{B}$ and $\mathbf{D}$. In all cases, the discrete state sequence (indicated by vertical lines) is recovered correctly by Viterbi decoding on f. The continous states, generally more difficult to estimate than the hidden discrete states, are recovered well under moderate noise (a), degraded as expected as the noise levels increase (b,c), and remain reasonable even in severe noise (d).

The effectiveness of the windowing technique is shown in Fig. 3, where a window size of ten strikes a good balance between

[3] $\mathbf{x}$ will reach a target as long as $\mathbf{A}$ is stable, which is probably the only case of interest for practical applications.

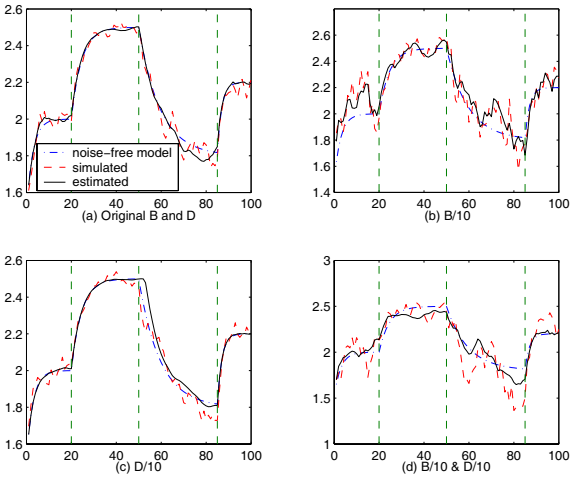

Fig. 2. Hidden state recovery under different noise levels: process noise $\mathbf{B}^{-1}$; observation noise $\mathbf{D}^{-1}$.

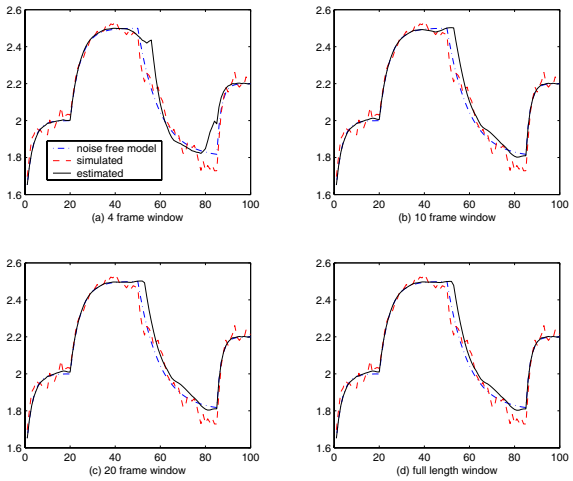

Fig. 3. The effect of different window sizes.

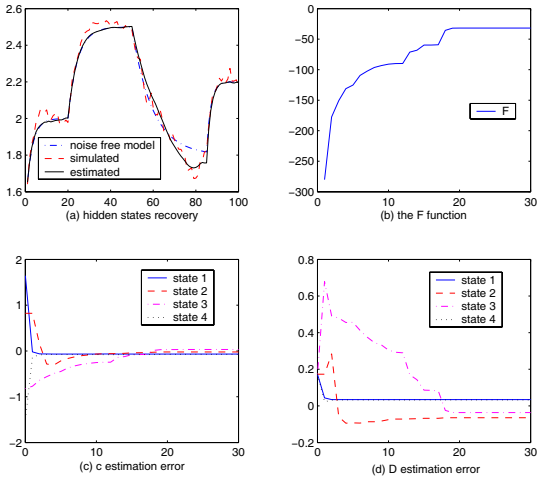

Fig. 4. Model parameter estimation.

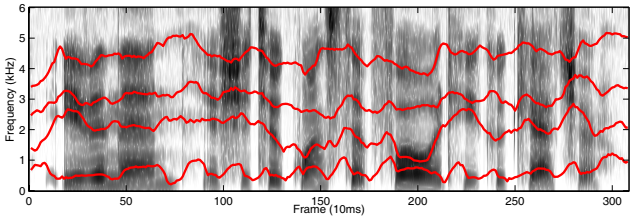

Fig. 5. Tracking VTRs for a speech sentence.

estimation accuracy and computational intensity for this particular example. Finally Fig. 4 tests the estimation of model parameters $\mathbf{c}$ and $\mathbf{D}$. Given initial values $\mathbf{c}' = \mathbf{c} - 1$ and $\mathbf{D}' = \mathbf{D}/2$, it can be seen that the variational EM procedure works well: the hidden dynamics are recovered well (a) and $\mathcal{F}$ is monotonically increasing and quickly converging (b). The estimates of c and D are accurate, evidenced by the small error norms (c,d).

5. APPLICATION AND DISCUSSION

The novel variational EM algorithm for SSS models developed in this paper admits a broad range of applications in signal processing and beyond. Fig. 5 shows one typical example in speech processing: vocal-tract-resonance (VTR) tracking for a TIMIT sentence. The tracking works well not only in the clear, vocalized regions, but also in more difficult consonant regions due to the built-in smoothness constraint of the model. How such models can be used in speech applications, especially for speech recognition, has been explored in [7]. The key point is that SSS models can capture the internal dynamics of speech which are completely missing in the state of the art technology; further developments and results are the subject of ongoing research.

6. REFERENCES

[1] L. Deng and J. Z. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal tract dynamics," *J. Acoust. Soc. Am.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[2] Y. Bar-Shalom and X.-R. Li, *Estimation and Tracking*, Artech House, Boston, 1993.

[3] V. Pavlovic, J. Rehg, T.-J. Cham, and K. Murphy, "A dynamic bayesian network approach to figure tracking using learned dynamic models," in *Proc. ICCV*, 1999, pp. 94–101.

[4] J. D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, pp. 357–384, 1994.

[5] U. Lerner and R. Parr, "Inference in hybrid networks: theoretical limits and practical algorithms," in *Proc. UAI*, 2001, pp. 310–318.

[6] Z. Ghahramani and G. E. Hintonn, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, pp. 831–864, 2000.

[7] L. J. Lee, H. Attias, and L. Deng, "Variational inference and learning for segmental state space models of hidden speech dynamics," in *Proc. ICASSP*, 2003, pp. 920–923.

[8] G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Time Series*, Springer-Verlag, New York, 1996.