Joint State and Parameter Estimation for a Target-Directed Nonlinear Dynamic System Model

Roberto Togneri, *Member, IEEE,* and Li Deng, *Senior Member, IEEE*

***Abstract*—In this paper, we present a new approach to joint state and parameter estimation for a target-directed, nonlinear dynamic system model with switching states. The model, which was recently proposed for representing speech dynamics, is also called the hidden dynamic model (HDM). The model parameters subject to statistical estimation consist of the target vector and the system matrix (also called the "time-constants"), as well as the parameters characterizing the nonlinear mapping from the hidden state to the observation. These latter parameters are implemented in the current work as the weights of a three-layer feedforward multilayer perceptron (MLP) network. The new estimation approach presented in this paper is based on the extended Kalman filter (EKF), and its performance is compared with the more traditional approach based on the expectation-maximization (EM) algorithm. Extensive simulation experiment results are presented using the proposed EKF-based and the EM algorithms and under the typical conditions for employing the HDM for speech modeling. The results demonstrate superior convergence performance of the EKF-based algorithm compared with the EM algorithm, but the former suffers from excessive computational loads when adopted for training the MLP weights. In all cases, the simulation results show that the simulated model output converges to the given observation sequence. However, only in the case where the MLP weights or the target vector are assumed known do the time-constant parameters converge to their true values. We also show that the MLP weights never converge to their true values, thus demonstrating the many-to-one mapping property of the feedforward MLP. We conclude from these simulation experiments that for the system to be identifiable, restrictions on the parameter space are needed.**

***Index Terms*—System modeling and representation.**

I. Introduction

THE work reported in this paper deals with the problem of state and parameter estimation for state-space dynamic models with switching states of the type of model similar to that in [35]. In this paper, we focus on a specific class of the model that is constrained to exhibit the target-directed property and implements a nonlinear observation equation. The problem of state and parameter estimation of dynamic models arises in many applications. In statistics, linear regression techniques can be generalized to a dynamic model that includes temporal evolution of the input variable. In control theory, a state-space dynamic has been widely used as a model for the noisy observations, assuming an underlying hidden state dynamic plant process. In adaptive signal processing, the Kalman filter technique adopts a state-space dynamic for formulating the minimum mean-square error (MMSE) linear filtering for complicated nonstationary problems. The specific class of dynamic model investigated in this paper has been used in recent research on continuous-state acoustic modeling for speech recognition, in [4] and [7], as the statistical coarticulatory model for speech production, and in [38] as the statistical hidden dynamic model evaluated against the discrete-state hidden Markov model [33] for a speech recognition task.

Our previous work on joint state and parameter estimation of the target-directed nonlinear model with switching states, which we will refer to as the hidden dynamic model (HDM), investigated the use of the extended Kalman filter (EKF) [1], [17], [28] and EM algorithms [2], [35], [36], for estimation of the parameters in the state equation only [37]. In this paper, we extend the results to include joint state and parameter estimation of both the state and observation equation parameters and perform more rigorous simulation experiments with equivalent complexity to the speech recognition task. We also solve the problem experienced in our earlier work related to the convergence of the EM algorithm. Comprehensive evaluation of the different algorithms in terms of the generative capabilities of the model, convergence to the true parameter values, computational complexity of the algorithm implementation, and implications for system identifiability will be presented in this paper.

The EKF algorithm that we have implemented for joint state and parameter estimation and parameter estimation only is known to perform suboptimally due to the first-order approximation of system nonlinearities. Recent work on the method of unscented transformations for more accurately calculating the distribution statistics of a nonlinear system has lead to the formulation of the unscented Kalman filter (UKF) [39], [41] and, together with efficient square-root numerical solutions, the SR-UKF algorithm [27], which is of the same complexity as the EKF but provides up to second-order accuracy in the nonlinearity. There are also different formulations of the problem, including sequential Monte Carlo methods (or particle filters) [9], [26], and variational learning methods applied directly to switching state-space models [11]. The traditional EKF algorithm was chosen due to its simple implementation and to establish a first-order analysis of an EKF-based method with the EM approach.

The organization of this paper is as follows. The HDM framework for this study is briefly discussed in Section II.

Manuscript received August 13, 2001; revised March 31, 2003. The associate editor coordinating the review of this paper and approving it for publication was Dr. Olivier Cappe.

R. Togneri is with the School of Electrical, Electronic, and Computer Engineering, The University of Western Australia, Crawley WA 6009, Australia (e-mail: roberto@ee.uwa.edu.au).

L. Deng is with Microsoft Research, Redmond WA 98052-6399 USA, and also with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: deng@microsoft.com).

Digital Object Identifier 10.1109/TSP.2003.819013

The proposed EM and EKF-based algorithms for joint state and parameter estimation are detailed in Section III. These algorithms are comparatively tested on the identical switching state-space, multiple-token training data generated from a hypothetical speech recognition task. The experimental setup used is explained in Section IV, and the results are presented in Section V. Finally, a summary of the results is made, and conclusions are drawn in Section VI.

II. Formulation of the Hidden Dynamic Model

The HDM studied in this paper represents one of a class of new dynamic acoustic modeling paradigms for speech recognition [4], [12], [31], [34], and the rationale for the form of the HDM adopted in this paper is provided in [7] with the model formulation detailed in [4], [7], and [38].

The first component of the HDM, which is also called the state equation, is a target-directed, continuously-valued (hidden) Markov process that is used to describe the vocal-tract resonance (VTR) dynamics according to

$$z(k+1) = \phi^j z(k) + (I_m - \phi^j)T^j + w(k) \tag{1}$$

where $z(k)$ is the $m\times 1$ VTR state vector, T^j is the $m\times 1$ phone target vector parameter, and ϕ^j is the $m \times m$ diagonal "time-constant" matrix parameter associated with the phone regime j. The phone regime is used to describe the segment of speech that is attributed to the phone identified by the model pair (ϕ^j, T^j). The process noise $w(k)$ is an i.i.d, zero-mean, Gaussian process with covariance Q. The target-directed nature of the process is evident by noting that $z(k) \to T^j$ as $k \to \infty$, independent of the initial value of the state.

The second component of the HDM is the observation equation used to describe the static mapping from the three-dimensional (3-D) hidden VTR state vector to the 12-dimensional observable acoustic feature vector. The general form of this mapping adopted in the current study assumes a static, multivariate nonlinear mapping function as follows:

$$O(k) = h^r(z(k)) + v(k) \tag{2}$$

where the $n\times 1$ acoustic observation $O(k)$ is the set of acoustic feature vectors for frame k, and $h^r(z(k))$ is the $n \times m$ static, nonlinear mapping function on the state vector $z(k)$ associated with the manner of articulation r. The manner of articulation describes how the phone is articulated to produce the acoustic observations arising from the speech production process and will usually be different for the different broad phonetic classes (e.g., vowels, voiced stops, etc.). The observation noise $v(k)$ is an i.i.d, zero-mean, Gaussian process with covariance R. The multivariate mapping function $h^r(z(k))$ is implemented by an m-J-n feedforward multilayer perceptron (MLP) with J hidden nodes, a linear activation function on the output layer, and the antisymmetric hyperbolic tangent function

$$g(x) = a\tanh(bx) \tag{3}$$

on the hidden layer, where $a = 1.72$ and $b = 2/3$ are chosen so that $g(x)$ exhibits useful unity slope and response at $x = 0$, and $x = \pm 1$ (see [13]). There is a unique MLP network for each distinct r.

The switching state behavior of this model is represented by an M-state discrete-time random sequence, where $j \equiv j(k) \in [1, 2 \ldots, M]$ is a random variable that takes on one of the M possible "phone" regimes (or states) at time k. An additional R-state discrete-time random sequence also exists where $r \equiv r(k) \in [1, 2, \ldots R]$ is a random variable that takes on one of the R possible manner of articulation states at time k. In practice, both sequences are unknown and need to be estimated, both when training the model (i.e., estimating the parameters) and testing (i.e., using the model to rescore or decode an unknown observation sequence).

An important property of this model is the continuity of the hidden state variable $z(k)$ across phone regimes: $z(0_{l+1}) = z(N_l)$, where N_l is the number of observation vectors in segment l, and 0_{l+1} is the initial observation vector for segment $l+1$. That is $z(k)$ at the start of segment $l+1$ is set to the value computed at the end of segment l. This provides a long-span continuity constraint across adjacent phone regimes that structurally models the inherent context dependencies and coarticulatory effects [7].

III. State and Parameter Estimation

The estimation problem that we investigate in this paper is as follows. Given multiple sets of observation sequences $O(k)$ for each distinct phone regime, we seek to determine the optimal estimates for the unknown values of the state-equation parameters ϕ and T and the observation-equation parameters W, which is the MLP weight vector of the nonlinear mapping function $h(z(k))$. For convenience and without causing loss of generality, we drop the j and r superscripts on the parameter variables. The hidden dynamic state vector $z(k)$ is usually also estimated simultaneously, giving rise to joint state and parameter estimation. However, state estimation is strictly not required, leading to the proposed parameter-only, EKF-based estimation algorithm detailed in Section III-C.

In this study, we assume that the phone sequence or segmentation of model regimes $j(k)$ is known in advance, which, in practice, requires training on phonetically transcribed speech corpora [4], [38]. In addition, for simplicity, we assume that there is only one manner of articulation (i.e., $r(k) = 1\ \forall k$). The former assumption is not unduly restrictive given the availability of phonetically transcribed data. However, estimation of the phone boundaries or the phone sequence is necessary when phonetic transcriptions are not available for training, and in testing when an unknown utterance is presented to the model for N-best or lattice rescoring. Solutions to this problem have been provided in [20] and [35] and will not be studied in this paper.

A. Joint State and Parameter Estimation by the EM Algorithm

The EM algorithm [2] is a widely used algorithm for the estimation of the parameters in the general state-space models [15], [36] and in the current research on the HDM [4], [7], [8]. The EM algorithm provides new estimates of the parameters after the set of all available N observation vectors have been presented. The EM algorithm can be considered a batch or offline

estimation method most suited to applications where all of the data is available. We now present the EM algorithm for the specific type of model given in Section II.

E-Step: For a sequence of N observation vectors, the E-step involves computation of the conditional expectation of the log joint likelihood between $Z = \{z(0),\ z(1),\ \ldots,\ z(N)\}$ and $O = \{O(0),\ O(1),\ \ldots,\ O(N)\}$, given the observation O and parameter set $\overline{\Theta}$ estimated at the previous step, that is

$$\begin{aligned} Q(\Theta|\overline{\Theta}) &= E\{\log L(Z,O|\Theta)|O,\overline{\Theta}\} \\ &= -\frac{1}{2}\sum_{k=0}^{N-1} E_N[e'_{k1}Q^{-1}e_{k1}|O,\overline{\Theta}] \\ &\quad -\frac{1}{2}\sum_{k=0}^{N-1} E_N[e'_{k2}R^{-1}e_{k2}|O,\overline{\Theta}] + \text{const} \end{aligned} \quad (4)$$

where $e_{k1} = [z(k+1) - \phi z(k) - (I-\phi)T]$ and $e_{k2} = [O(k) - h(z(k))]$, and E_N denotes the expectation based on N samples. The standard EKF smoother is used to provide estimates of the hidden dynamic variable $z(k) \equiv \widehat{z}(k|N) = E_N[z(k)|O,\overline{\Theta}]$ [22], [37]. The Jacobian matrix for the $n \times m$ nonlinear mapping function $h(z(k))$ used in the EKF recursion is given by

$$\begin{aligned} H_z^{ji}[\widehat{z}(k+1|k)] &= \left[\frac{\partial O_j(k+1)}{\partial \widehat{z}_i(k+1|k)}\right] \\ &= \left[\sum_{h=1}^{J} W_h^{2j} g'(W^{1h'} \cdot \widehat{z}(k+1|k)) W_i^{1h}\right] \end{aligned} \quad (5)$$

where $O_j(k)$ is the jth component of the observation vector at time k, $\widehat{z}_i(k+1|k)$ is the i^{th} component of the predicted state vector $\widehat{z}(k+1|k)$ at time k, W_i^{lh} is the i^{th} component of the MLP weight vector W^{lh} of node h in layer l (layer 1 is the hidden layer and layer 2 is the output layer), J is the number of nodes in the hidden layer, and $g'(x)$ is the derivative of the activation function in the hidden layer.

It should be noted that the continuity condition on $\widehat{z}(k)$ is also applied to the EKF error covariance $\widehat{P}(k)$.

M-Step: In the M-step, the Q function in (4) is maximized with respect to the parameter set $\Theta = (T, \phi, W)$. We consider the first summation involving the parameters T and ϕ:

$$Q_1(Z,O,\Theta) = \sum_{k=0}^{N-1} E_N[e'_{k1}Q^{-1}e_{k1}|O,\overline{\Theta}].$$

Minimization of Q_1, which implies maximization of Q, proceeds by setting the partial derivatives with respect to T and ϕ to zero, that is

$$\begin{aligned} \frac{\partial Q_1}{\partial \phi} &\propto \sum_{k=0}^{N-1} E_N\{[z(k+1) - \phi z(k) - (I-\phi)T][T - z(k)]'|O,\overline{\Theta}\} \\ &= 0 \\ \frac{\partial Q_1}{\partial T} &\propto \sum_{k=0}^{N-1} E_N\{[z(k+1) - \phi z(k) - (I-\phi)T]|O,\overline{\Theta}\} = 0. \end{aligned}$$

The resulting equations to be solved are nonlinear high-order equations in terms of ϕ and T:

$$N\phi TT' - \phi TB' - \phi BT' - NTT' + \phi D + TB' + AT' - C = 0 \quad (6)$$

$$A - \phi B - NT + N\phi T = 0 \quad (7)$$

where

$$A = \sum_{k=0}^{N-1} E_N[z(k+1)|O,\overline{\Theta}],\quad C = \sum_{k=0}^{N-1} E_N[z(k+1)z(k)'|O,\overline{\Theta}]$$

$$B = \sum_{k=0}^{N-1} E_N[z(k)|O,\overline{\Theta}],\quad D = \sum_{k=0}^{N-1} E_N[z(k)z(k)'|O,\overline{\Theta}]$$

are the relevant sufficient statistics that are computed by the EKF smoother during the E-step.

Direct solutions of (6) and (7) for either parameter can be easily derived, assuming the other parameter is fixed (i.e., known). However, for joint estimation of ϕ and T, a direct solution is not evident. One alternative is to apply the ECM algorithm described in [25], which has been shown to hold the same convergence properties as the EM algorithm. In our previous work [37], however, we found that the ECM procedure did not converge. This is due to the constrained nature of the parameter space for ϕ, which must lie in the range [0,1). The convergence properties of the ECM (and any generalized EM algorithm) assume an unconstrained parameter space. Specifically, the problem in our case is that in the ECM formulation for T, it is assumed that $\phi \neq I$. In cases where ϕ was re-estimated close to I, the re-estimated value for T would be ill-defined, and this was the cause for the ECM algorithm failing to converge. One solution would be to reset ϕ to a reasonable value when it is re-estimated close to I; however, this makes it difficult to analyze the convergence properties of the ECM. An alternative is to use a locally optimal gradient method, as discussed in [16], with a suitable initialization of the parameters and step-size. In this paper, we propose to maximize (6) and (7) jointly by a simple gradient descent method with the gain step and number of iterations empirically chosen to ensure convergence.

We now consider the second summation of the Q function in (4) involving the parameter W:

$$Q_2(Z,O,\Theta) = \sum_{k=0}^{N-1} E_N[e'_{k2}R^{-1}e_{k2}|O,\overline{\Theta}].$$

Minimization of Q_2, which leads to maximization of Q, proceeds by setting the partial derivatives with respect to W to zero, that is

$$\frac{\partial Q_2}{\partial W} \propto \sum_{k=0}^{N-1} E_N\left[\frac{\partial}{\partial W}\{[O(k) - h(z(k)]^2\}|O,\overline{\Theta}\right] = 0.$$

That is, Q_2 is minimized when the error signal $e(k) = O(k) - h(z(k))$ is minimized. Since the multivariate mapping function is a feedforward MLP network, then the standard back-propagation [13] is used with $\widehat{z}(k|N)$ as the input and $O(k)$ as the desired output to provide estimates of the MLP weights W.

B. Joint State and Parameter Estimation by the EKF Algorithm

The use of the traditional EKF for joint state and parameter estimation is not new [17], [28], and its application to the HDM has been detailed in our earlier work [37], [38] for the parameter set $\Theta = (T, \phi)$. Parameter estimation based on the EKF algorithm differs fundamentally from estimation using the EM algorithm in that new estimates of the parameters are provided immediately after the presentation of the current observation

vector. Thus, EKF-based parameter estimation is a recursive, online method suitable for applications requiring continuous parameter updates at each observation time-step and where not all of the data needs to be available. The use of the EKF algorithm for joint state and parameter estimation involves extending the standard EKF algorithm to the complete parameter set $\Theta = (T, \phi, W)$. This is achieved by defining the augmented state vector

$$\theta(k) = \begin{pmatrix} z(k) \\ \widetilde{\phi}(k) \\ T(k) \\ \widetilde{W}(k) \end{pmatrix} \tag{8}$$

where $T(k)$ is the target vector at time k. The "super"-vector

$$\widetilde{\phi}(k) = \begin{pmatrix} \phi_1'(k) \\ \phi_2'(k) \\ \vdots \\ \phi_m'(k) \end{pmatrix} \tag{9}$$

is the $m^2 \times 1$ time-constant "vector" at time k, where $\phi_i(k)$ is row i of $\phi(k)$, and the "super"-vector

$$\widetilde{W}(k) = \begin{pmatrix} W^{11}(k) \\ W^{12}(k) \\ \vdots \\ W^{1J}(k) \\ W^{21}(k) \\ W^{22}(k) \\ \vdots \\ W^{2n}(k) \end{pmatrix} \tag{10}$$

consists of all MLP weights, where $W^{lh}(k)$ is the MLP weight vector of node h in layer l at time k.

After the definition of the augmented state vector, the new state equation becomes

$$\theta(k+1) = f(\theta(k)) + w(k)$$

which is now nonlinear in the state variable $\theta(k)$ and can be decomposed as

$$\begin{pmatrix} z(k+1) \\ \widetilde{\phi}(k+1) \\ T(k+1) \\ \widetilde{W}(k+1) \end{pmatrix} = \begin{pmatrix} \phi(k)z(k) + (I - \phi(k))T(k) \\ \widetilde{\phi}(k) \\ T(k) \\ \widetilde{W}(k) \end{pmatrix} + \begin{pmatrix} w_z(k) \\ w_\phi(k) \\ w_T(k) \\ w_W(k) \end{pmatrix}.$$

The measurement equation now becomes

$$O(k) = h(\theta(k)) + v(k).$$

The state equation error covariance matrix Q is also augmented to include the covariance of the parameter noise processes $w_\phi(k)$, $w_T(k)$, and $w_W(k)$ in addition to the noise process of the dynamic state $w_z(k)$. The addition of the noise parameter changes the modeling paradigm in an important way from EM-based parameter estimation, where there is no "noise" process associated with the parameters.

The standard EKF recursion can now be used to yield joint state and parameter estimates at each time-step. The expression for the $(2m + m^2 + p) \times (2m + m^2 + p)$ state equation Jacobian matrix for the nonlinear function $f(\theta(k))$ used in the EKF recursion has been derived to have the following form:

$$F_\theta[\widehat{\theta}(k|k)] = \begin{bmatrix} \widehat{\phi}(k|k) & \frac{\partial f}{\partial \phi}(k|k) & I_m - \widehat{\phi}(k|k) & 0 \\ 0 & I_{m^2} & 0 & 0 \\ 0 & 0 & I_m & 0 \\ 0 & 0 & 0 & I_p \end{bmatrix}$$

where $\widehat{\phi}(k|k)$ is the current estimate of ϕ, $p = (n+1) \times J + (J+1) \times m$ is the number of weights in a m-J-n feedforward MLP network (including the bias terms), and

$$\frac{\partial f}{\partial \phi}(k|k) = \begin{bmatrix} [\widehat{z}(k|k) - \widehat{T}(k|k)]' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [\widehat{z}(k|k) - \widehat{T}(k|k)]' \end{bmatrix} \tag{11}$$

is the $m \times m^2$ partial derivative submatrix expression for $\partial f / \partial \phi$.

The expression for the $(n) \times (2m + m^2 + p)$ measurement equation Jacobian matrix for the nonlinear function $h(\theta(k))$ used in the EKF recursion is strictly dependent only on $\widehat{z}(k+1|k)$ and $\widehat{W}(k+1|k)$. It can be expressed as

$$H_\theta[\widehat{\theta}(k+1|k)] = \begin{bmatrix} H_z^{ji}[\widehat{z}(k+1|k)] & 0 & 0 & H_W^{jw}[\widehat{W}(k+1|k)] \end{bmatrix}$$

where the expression for the $n \times m$ Jacobian matrix $H_z^{ji}[\widehat{z}(k+1|k)]$ is given by (5), and (12), shown at the bottom of the page, is an $n \times p$ Jacobian matrix.

The same continuity condition on $\widehat{z}(k)$ is also applied to $\widehat{W}(k)$ and the error covariance $\widehat{P}(k)$. For the phone dependent parameters $(\widehat{T}(k), \widehat{\phi}(k))$, the continuity condition is slightly more complex: $\widehat{\phi}^j(0_{l^j+1}) = \widehat{\phi}^j(N_{l^j})$ and $\widehat{T}^j(0_{l^j+1}) = \widehat{T}^j(N_{l^j})$. Here, l^j and $l^j + 1$ are the successive (but not necessarily adjacent) segments of the same state-space phone model j, N_{l^j} is the number of observation vectors in segment l^j (the lth segment for phone j), and 0_{l^j+1} is the

$$H_W^{jw}[\widehat{W}(k+1|k)] = \left[\frac{\partial O_j(k+1)}{\partial \widehat{W}_i^{lh}(k+1|k)} \right] = \left[\begin{cases} g(\widehat{W}^{1i'}(k+1|k) \cdot \widehat{z}(k+1|k)), & \text{if } l = 2,\ h = j \\ \widehat{W}_h^{2j}(k+1|k) g'(\widehat{W}^{1h'}(k+1|k).\widehat{z}(k+1|k)) \widehat{z}_i(k+1|k), & \text{if } l = 1 \\ 0, & \text{otherwise} \end{cases} \right] \tag{12}$$

initial observation for segment $l^j + 1$ (the $(l+1)$th segment for phone j).

C. Parameter Estimation by the EKF Algorithm

The use of the EKF for joint state and parameter estimation strictly does not require estimation of the state since we are only concerned with parameter estimation. If the hidden dynamic state is assumed to be a deterministic process then the EKF can be used for parameter-only estimation (i.e., without the need to estimate the state sequence). The method described here is an extension of the EKF training algorithm of a recurrent neural network [10], [14], [32] to the HDM using static linearizations.

We define the augmented state vector as

$$\theta(k) = \begin{pmatrix} \widetilde{\phi}(k) \\ T(k) \\ \widetilde{W}(k) \end{pmatrix} \tag{13}$$

where $\widetilde{\phi}(k)$ is given by (9), $\widetilde{W}(k)$ is given by (10), and $T(k)$ is the target vector at time k.

The state equation becomes

$$\theta(k+1) = \theta(k) + w(k)$$

which is a simple linear function in the state variable $\theta(k)$. The noise process $w(k)$ includes the parameter noise processes $w_\phi(k)$, $w_T(k)$, and $w_W(k)$. It should be noted that the system state is still $z(k)$, but the state for the purposes of the EKF recursion is $\theta(k)$.

The measurement equation becomes

$$O(k) = h(\theta(k), z(k)) + v(k).$$

The standard EKF recursion is used to yield the parameter estimates at each time-step, where the state equation system matrix $F_\theta = I_{m+m^2+p}$ and the "true" state is recursively computed from the relation

$$z(k+1) = \widehat{\phi}(k|k)z(k) + (I_m - \widehat{\phi}(k|k))\widehat{T}(k|k). \tag{14}$$

Since the "true" state is itself a function of the parameters being estimated and the previous estimate of the state, a recurrent or dynamic derivation of the Jacobian matrix derivatives for the nonlinear function $h(\theta(k), z(k))$ is needed [32], [40]. However, to avoid the complexity of formulating these recurrent derivatives, an approximation based on static derivatives is used assuming the "true" state is known and given by (14).

The expression for the $(n)\times(m+m^2+p)$ measurement equation Jacobian matrix for the nonlinear function $h(\theta(k))$ used in the EKF recursion can be expressed as the equation at the bottom of the page, where $H_W^{jw}[\widehat{W}(k+1|k)]$ is given by (12).

We form the $n \times m^2$ expression

$$H_\phi^{jp}[\widehat{\phi}(k+1|k)] = \sum_{k=1}^{m} \left[\frac{\partial O_j(k+1)}{\partial z_k(k+1)}\right] \left[\frac{\partial z_k(k+1)}{\partial \widehat{\phi}_p(k+1|k)}\right] = \sum_{k=1}^{m} H_z^{jk}[z(k+1)] \left[\frac{\partial f}{\partial \phi}(k+1|k)\right]^{kp}$$

where $H_z^{jk}[z(k+1)]$ is given by (5), $[\partial f/\partial\phi(k+1|k)]$ is given by (11), and the $n \times m$ matrix

$$H_T^{ji}[\widehat{T}(k+1|k)] = \sum_{k=1}^{m} \left[\frac{\partial O_j(k+1)}{\partial z_k(k+1)}\right] \left[\frac{\partial z_k(k+1)}{\partial \widehat{T}_i(k+1|k)}\right] = \sum_{k=1}^{m} H_z^{jk}[z(k+1)] \left[I_m - \widehat{\phi}(k+1|k)\right]^{ki}.$$

The continuity condition that applies to $z(k)$, $\widehat{W}(k)$, and $(\widehat{T}(k), \widehat{\phi}(k))$ is as described previously.

IV. Simulation Experiments—Conditions

Simulated data was used to evaluate the performance of several estimation algorithms described in Section III for $\Theta = (T, \phi, W)$. All experiments were based on a hidden dynamic $z(k)$ of dimension $m = 3$, an acoustic feature vector $O(k)$ of dimension $n = 12$, and a 3–8-12 feedforward MLP network. The choice of eight nodes in the hidden layer was made as a compromise between having too many parameters to train and not enough hidden units to allow the network sufficient nonlinearity in the mapping.

The "time-constant" ϕ is a diagonal 3×3 matrix comprising the three diagonal terms. Estimation of a diagonal ϕ was achieved by diagonalizing the ensuing full matrix that is computed at each M-step of the EM algorithm and at each time-step of the EKF algorithms. There are a total of six scalar parameters for the state parameter vectors ϕ and T and 140 scalar parameters for the nonlinear mapping function MLP weights of a 3–8–12 network. Thus, the estimation of the MLP weights was by far the most time-consuming task, but the ϕ and T are by far the most important parameters since they characterize each phone regime.

Both the experimental setup and simulated data were based on typical conditions found in using the HDM for acoustic modeling [4], [7], [38]. The simulated data was generated using (1) with $w(k) = 0$ and $z(0) = [100, 700, 1700]'$ and (2) with $v(k) = N(0, 0.0625)$.

Two 3–8–12 MLP networks were randomly generated: One was used to generate the simulated data and represented the "true" MLP network, and the other was used as the initial network in experiments where the MLP weight vector W had to be estimated. For optimum performance, the inputs to an MLP should be normalized [13], and this was achieved by presenting $z_{\text{in}}(k) = z(k) - E(z)$, where $E(z) = [300, 1200, 2000]'$ to the

$$H_\theta[\widehat{\theta}(k+1|k)] = \left[\, H_\phi^{jp}[\widehat{\theta}(k+1|k)] \quad H_T^{ji}[\widehat{T}(k+1|k)] \quad H_W^{jw}[\widehat{W}(k+1|k)] \,\right]$$

MLP network. $E(z)$ is an empirical estimated of the expected value of the hidden dynamic state $z(k)$.

The number of phone segment models were chosen to correspond to 31 of the phonemes of the English language. The "true" target values T were based on the Klatt synthesizer setup, as described in [7], and the "true" time-constant ϕ were randomly generated values in the range $[0.5 \ldots 1.0]$. A total of 64 utterances were generated based on random phonetic transcriptions with $[13 \ldots 23]$ phone models per utterance and $[7 \ldots 17]$ frames per model segment. This produced a training data set with a total of 12 695 observations.

In all experiments the (ϕ, T) parameters were initialized by

$$\widehat{\phi}(0|0)=\begin{bmatrix} 0.75 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}=\text{diag}^3(0.75),\ \widehat{T}(0|0)=\begin{bmatrix} 300 \\ 1200 \\ 2000 \end{bmatrix}$$

and the hidden dynamic state was initialized by $\widehat{z}(0|0) = [100, 700, 1700]'$.

The error covariance for the hidden state variable was initialized to $\widehat{P}(0|0) = 0$, corresponding to the condition of no errors in the initial estimate for $\widehat{z}(0|0)$. For the two EKF parameter estimation algorithms, the error covariance corresponding to the parameters in the augmented state vector was initialized to $\widehat{P}(0|0) \equiv [\text{diag}^3(\phi_\epsilon)\ \text{diag}^3(T_\epsilon)\ \text{diag}^p(W_\epsilon)]'$, where $(\phi_\epsilon, T_\epsilon, W_\epsilon) \cong (10^{-5}, 0.1, 10^{-10})$ are empirically chosen values proportional to the expected error in the corresponding parameter to ensure smooth convergence of the EKF algorithms.

The process and observation noise covariance Q and R matrices are set to fixed values and do not form part of the estimation process in the EKF recursions. Although the use of fixed values may produce questionable convergence results, this was found not be the case for the results reported in this paper if reasonable choices for the parameters were made. For the EM parameter estimation algorithm $Q = \text{diag}^3(10)$ was used to describe the error in the predicted value of the state $\widehat{z}(k+1|k)$ arising from the incorrect state parameter values (ideally, Q should be annealed to 0 as the state parameter values converge to their correct values). For the EKF parameter estimation algorithm, $Q = \text{diag}^{3+3+p}(0)$, which indicates the absence of any "noise" in the augmented state equation [that is, we assume $w_\phi(k) = w_T(k) = w_W(k) = 0$ and that errors in the parameter values are described by the error covariance matrix $P(k)$]. For the EKF joint state and parameter estimation algorithm, we additionally set $Q = \text{diag}^3(10^{-5})$ for the covariance of the noise process $w_z(k)$, where 10^{-5} is an arbitrary fictitious noise for the uncertainty in the hidden state variable equation. For all algorithms, the observation noise covariance $R = \text{diag}^{12}(0.0625)$, which describes the effect of the added Gaussian observation noise.

V. Simulation Experiment—Results

The EM and EKF algorithms described in Section III were evaluated by different parameter estimation trials based on the simulated data experimental setup described in Section IV.

The three estimation algorithms evaluated were the following:

EM—joint state and parameter estimation by the EM algorithm (Section III-A);
EKFZ—joint state and parameter estimation by the EKF algorithm (Section III-B);
EKFP—parameter estimation by the EKF algorithm (Section III-C).

There were four different experimental evaluations carried out involving different combinations of parameters to be estimated. These were the following:

1) parameter set to be estimated consisted of the state parameters $\Theta = (\phi, T)$, with observation parameter W assumed known (i.e., fixed to the "true" value);
2) parameter set to be estimated consisted of the complete parameter set $\Theta = (\phi, T, W)$.
3) parameter set to be estimated consisted of $\Theta = (\phi, W)$ with the target vector T known;
4) parameter set to be estimated consisted of $\Theta = (T, W)$ with the time-constant ϕ known.

For each simulation experiment, results are presented after 100 iterations of the algorithm, where an iteration is defined as one run or pass through all of the training data (corresponding to 12 695 observations per iteration).

The average percentage deviation of the estimated parameters from the known "true" values was calculated to indicate the convergence of the algorithm and identifiability of the system. The algorithm performance in minimizing the innovation sequence $e(k) = O(k) - h(\widehat{z}(k|k))$ was examined by calculating the difference between the observation sequence $O(k)$ and the sequence generated by HDM during training $\widehat{O}(k) = h(\widehat{z}(k|k))$. This difference is presented as both an average mean-square-error (MSE) and an average percentage deviation. The generative capabilities of the HDM was examined by plotting the sixth component of the observation vector $O(k)$ together with the HDM generated output $h(\widehat{z}(k|k))$ between sample times 5800 and 6000.

To gauge the computational load of the proposed algorithms, the CPU time (user and system time) was measured for 100 iterations of the algorithm.

The significance of each algorithm's performance was verified by including the average mean-square-error and average percentage deviation results based on the initial values of the parameters prior to estimation. These results are indicated by the column labeled "Untrained" and represent the worst-case performance.

A. Parameter Set $\Theta = (\phi, T)$ and Known W

From the results in Table I, it is evident that the state parameter set $\Theta = (\phi, T)$ converged to the true values, and hence, this system is identifiable for both the EM and EKF algorithms, with EKFZ exhibiting the the smallest deviation for both the parameters and observation sequence. Among the three algorithms, the EM algorithm was the most expensive computationally and had the largest parameter deviation. The EKFP was marginally faster than the EKFZ but was also slightly less accurate. To examine the properties of the algorithms further, the synthesized model outputs are plotted in Fig. 1. The model output plots closely match the observation sequence for all three algorithms.

TABLE I
ESTIMATION RESULTS FOR PARAMETER SET $\Theta = (\phi, T)$ AND KNOWN W AFTER 100 ITERATIONS OF THE EM, EKFZ AND EKFP ALGORITHMS

	Untrained	EM	EKFZ	EKFP
ϕ (% deviation)	28.93%	4.83%	0.574%	1.16%
T (% deviation)	25.96%	4.30%	0.379%	2.68%
Obs MSE	29.67	0.0979	0.0632	0.0870
Obs (% deviation)	99.89%	6.34%	5.41%	5.92%
CPU time	-	1955 s	890 s	712 s

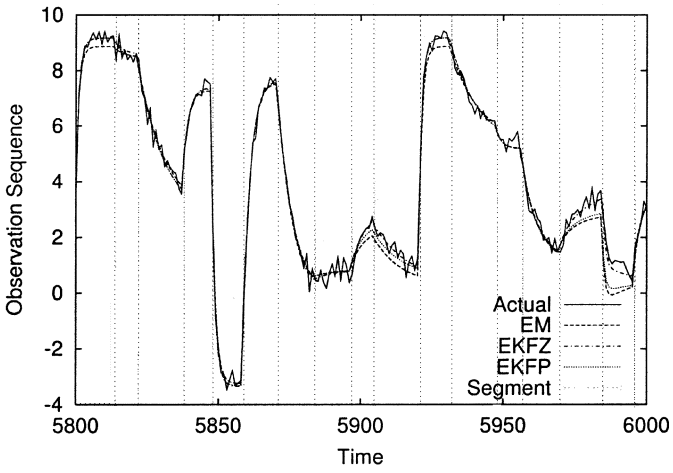

Fig. 1. Plot of the one component of the observation vector sequence $O(k)$ and synthesized EM, EKFZ, and EKFP HDM outputs after 100 iterations from frame 5800 to 6000 for parameter set $\Theta = (\phi, T)$ and known W.

TABLE II
ESTIMATION RESULTS FOR PARAMETER SET $\Theta = (\phi, T, W)$ AFTER 100 ITERATIONS OF THE EM, EKFZ, AND EKFP ALGORITHMS

	Untrained	EM	EKFZ	EKFP
ϕ (% deviation)	28.93%	23.00%	31.33%	25.92%
T (% deviation)	25.96%	42.88%	43.43%	32.12%
W (% deviation)	265.43%	290.94%	830.18%	858.04%
Obs MSE	29.72	1.53	0.617	1.010
Obs (% deviation)	100.98%	22.12%	13.78%	17.98%
CPU time	-	9577 s	185860 s	94609 s

This result follows from the convergence of the parameters to their true values.

Unlike our previous work based on the ECM algorithm [37], the use of a simple gradient descent method did not result in any convergence problems with the EM algorithm.

B. Parameter Set $\Theta = (\phi, T, W)$

From the results in Table II, it is evident that the combined state and observation parameter set $\Theta = (\phi, T, W)$ failed to converge to the true values, and hence, this system is not identifiable for either the EM or EKF-based algorithms. However, the observation MSE and percentage deviation for all the three algorithms is significantly lower than the Untrained performance. Furthermore, from Fig. 2, the synthesized model output converges to the observation sequence for all three algorithms. These results show that the system parameters are not uniquely specified, and incorrect values can yield the same model output performance. The underlying cause is the

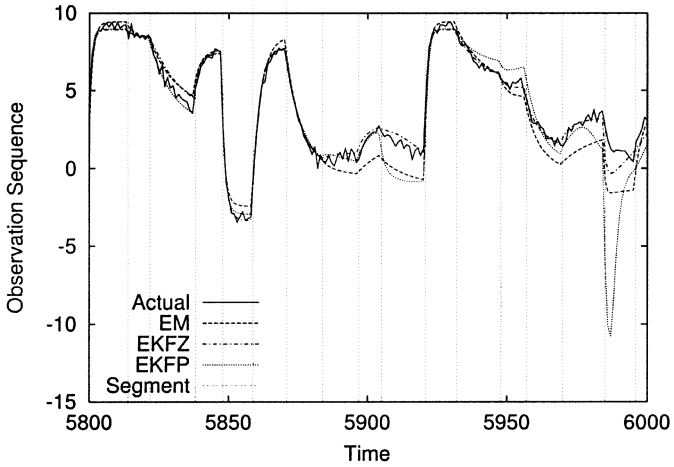

Fig. 2. Plot of the one component of the observation vector sequence $O(k)$ and synthesized EM, EKFZ, and EKFP HDM outputs after 100 iterations from frame 5800 to 6000 for parameter set $\Theta = (\phi, T, W)$.

TABLE III
ESTIMATION RESULTS FOR PARAMETER SET $\Theta = (\phi, W)$ WITH KNOWN T AFTER 100 ITERATIONS OF THE EM, EKFZ, AND EKFP ALGORITHMS

	Untrained	EM	EKFZ	EKFP
ϕ (% deviation)	28.93%	6.68%	0.750%	0.683%
W (% deviation)	265.43%	241.12%	314.69%	355.91%
Obs MSE	64.06	0.347	0.0684	0.0682
Obs (% deviation)	152.94%	11.24%	5.590%	5.59%
CPU time	-	9640 s	189714 s	93821 s

unconstrained nonlinear mapping implemented by the MLP, which can be freely adjusted to compensate for errors in the state parameter values. The insight gained from these results is that if the goal is to estimate physically plausible parameters (as we claim the HDM is), then the search space will need to be restricted for a unique solution.

Based on this insight, in the remaining two simulation experiments, one of the state parameters is assumed known in order to examine to what degree the system is uniquely specified under this constrained condition.

When comparing the computational load between the EM and EKF algorithms, the augmented state vector for the EKF, in particular the size of the MLP parameter set, increases the CPU time by almost two orders of magnitude, with the EKFZ exhibiting the worst overall computational performance. The main contributing factor is the multiplication of $(m+m^2+p)$ square matrices with $p \cong 140$ at each time step compared with the case when W is a known parameter and $p = 0$.

C. Parameter Set $\Theta = (\phi, W)$ With Known T

The results in Table III show that under the condition of the known target vector T, the time-constant ϕ converged to the true value for all three algorithms. The failure of the MLP weight parameters W to converge to the true values further strengthens the argument that the MLP network is too unrestricted, and it verifies the many-to-one mapping ability of the MLP. As there is no direct physical interpretation of the MLP weights, their convergence to the true values is not critical. However, convergence of the physically meaningful state parameters (ϕ, T) to the true

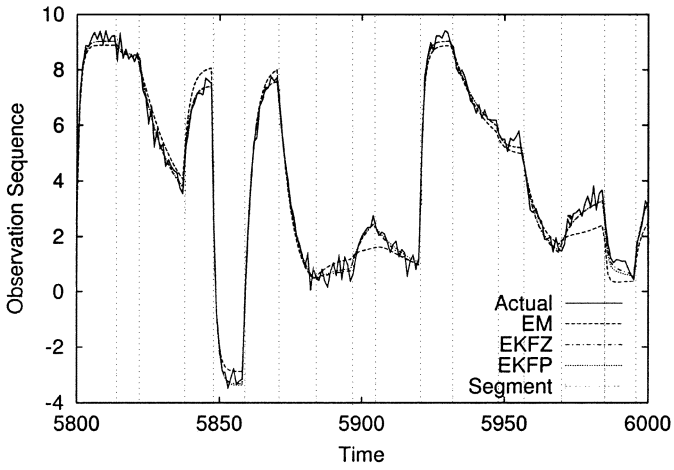

Fig. 3. Plot of the one component of the observation vector sequence $O(k)$ and synthesized EM, EKFZ, and EKFP HDM outputs after 100 iterations from frame 5800 to 6000 for parameter set $\Theta = (\phi, W)$ and known T.

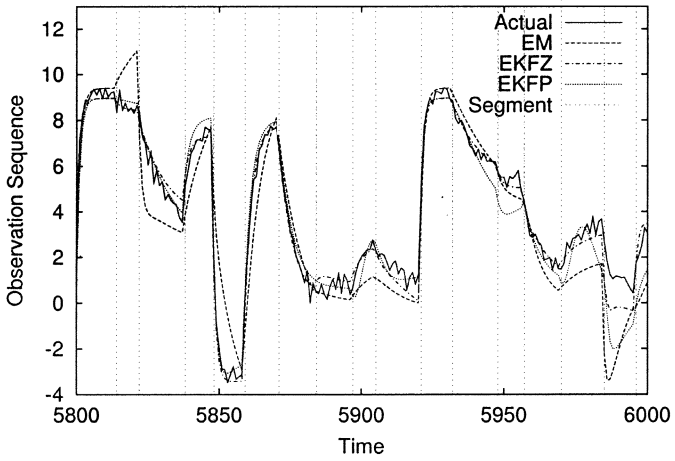

Fig. 4. Plot of the one component of the observation vector sequence $O(k)$ and synthesized EM, EKFZ, and EKFP HDM outputs after 100 iterations from frame 5800 to 6000 for parameter set $\Theta = (T, W)$ and known ϕ.

TABLE IV
ESTIMATION RESULTS FOR PARAMETER SET $\Theta = (T, W)$ WITH KNOWN ϕ AFTER 100 ITERATIONS OF THE EM, EKFZ, AND EKFP ALGORITHMS

	Untrained	EM	EKFZ	EKFP
T (% deviation)	28.93%	41.39%	47.31%	32.14%
W (% deviation)	265.43%	295.75%	929.23%	1476.43%
Obs MSE	30.05	2.76	0.288	1.29
Obs (% deviation)	101.90%	29.36%	10.66%	21.17%
CPU time	-	9626 s	191037 s	95501 s

values is significant. Thus, estimation of $\Theta = (\phi, W)$ with known T is a feasible alternative to estimation of the complete parameter set since ϕ can be uniquely identified. Moreover, from the observation MSE and deviation results in Table III and the results in Fig. 3, we again see that the synthesized model output sequence closely matches the observation sequence for all three algorithms.

As the parameter set includes the MLP weights W, the measured CPU time for EKFZ and EKFP is one to two orders of magnitude more than that for the EM algorithm. However, the EKFZ and EKFP exhibit superior observation MSE and deviation performance than the EM algorithm.

D. Parameter Set $\Theta = (T, W)$ With Known ϕ

The results in Table IV show that under the condition of known time-constant ϕ, the target vector T still failed to converge to the true value. This is in sharp contrast to the results in Table III and illustrates an asymmetric relation between the parameters ϕ and T. In addition, as expected, the MLP weights W also failed to converge to the true values. However, the observation MSE and deviation results and Fig. 4 indicate that for all three algorithms, the model output sequence was converging to the observation sequence. Thus, this system is not uniquely identifiable and will not produce physically meaningful estimates of T.

VI. SUMMARY AND DISCUSSION

Three different EM and EKF-based algorithms for state and parameter estimation in the HDM have been proposed and were evaluated on simulated data generated using a typical setup for applying the HDM to speech modeling. We presented and analyzed the experimental results in the following three aspects:

1) convergence or divergence of the estimated parameters to the known "true" values;
2) convergence or divergence of the synthesized model output to the given observation sequence;
3) comparative computational costs of the three algorithms.

Among the three algorithms evaluated, the EKFZ exhibited the best convergence. However, in the experiments involving the MLP weights W, both the EKFZ and EKFP algorithms experienced a one-to-two-orders-of-magnitude increase in computational cost compared with the EM algorithm. Thus, the EM algorithm is preferred in cases where the MLP weights W need to be estimated.

In comparing the EKFZ and EKFP algorithms, we observe that the EKFZ exhibits better performance but at the cost of a greater computational load due to the larger state vector $\theta(k)$. The performance of the EKFP is at least as good as, if not better, than the EM algorithm. Since the EKFP only used static rather the dynamic derivatives, this may explain its inferior performance to the EKFZ. The EKFZ and EKFP algorithms both outperform the EM algorithm, and given the suboptimal nature of these algorithms, more advanced implementations of the EKF-based algorithms remain to be further investigated. These implementations include the decoupled EKF (DEKF) [14], [32] to reduce the computational complexity arising from estimating the MLP weights, the dual and joint forms of the SR-UKF [40] to further improve convergence, and the estimation of the MLP weights W by an offline batch-mode trained back-propagation algorithm, as is the case with the EM implementation. By training the MLP weights separately, the computational burden associated with using a large state vector with the EKF and SR-UKF algorithms is greatly reduced.

The results for the complete parameter set $\Theta = (\phi, T, W)$ show that the presence of the unrestricted, many-to-one, MLP mapping function prevents the state parameters (ϕ, T) from converging to their true values. Thus, estimation of the complete parameter set will fail to yield physically meaningful values for

the time-constant and target parameters unless these are carefully initialized. However, if the target T or the MLP weights W are assumed known, then the unknown system matrix parameters (i.e., the time constants) will converge to the true values.

In the case of acoustic modeling for speech, the targets T derived from the Klatt synthesiser setup [7] can be assumed to be reasonably close to the known "true" values, and thus, the parameter set to be estimated is reduced to $\Theta = (\phi, W)$. Alternatively, the Klatt-synthesizer derived targets can be used as the initial values for T with the complete parameter set $\Theta = (\phi, T, W)$ being estimated. It remains to be seen by future investigations whether convergence to the true values will hold.

From investigation of the acoustic-phonetics of speech, the mapping functions from the VTR states to the acoustic measurements may be established for different classes of speech and speaker characteristics. Thus, the MLP weight vector W can be treated as known, and the parameter set can be reduced to $\Theta = (\phi, T)$. Since this would not include the W parameter, the current implementation of the EKFZ algorithm or the more optimal SR-UKF can be efficiently used due to their superior convergence over the EM algorithm.

As has been shown by the results in Section V, the performance of the training algorithms and convergence of the parameters to their true values depends heavily on whether the MLP weights W are assumed known or are parameters that need to be estimated. In the justification for the current structure of the HDM [4], [7], [38], the MLP is chosen to represent the most general nonlinear mapping between the internal states and the observable acoustics. Alternative formulations with linear mappings and mixture of linear mappings have also been proposed [19], and more investigation is needed to determine whether a simpler mapping function than a feedforward MLP is feasible.

The performance of the EKF-based algorithms was found to be highly dependent on the initialization of the $P(0|0)$, as well as of the noise covariance matrices Q and R. In the work reported in this paper, Q and R were set to some empirically fixed values, which are not optimal. Furthermore, the effect of the parameter noise processes $w_\phi(k)$, $w_T(k)$, and $w_W(k)$ on the convergence of the EKF-based algorithms requires further investigation. For the synthetic experiments discussed in this paper, the use of fixed values did not result in any serious convergence problems or inconsistent results, but ideally, Q and R should be adapted during the training process, especially for more realistic problems. Possible improvements in this respect are 1) including Q and R in the parameter set (i.e., $\Theta = (\phi, T, W, Q, R)$) [5] or, in cases where that is too difficult, 2) annealing Q and R during the training process [14], and investigating 3) the addition of a small artificial noise process to Q and R to improve stability [23].

An important problem arising from the switching state characteristic is the estimation of the switching state sequences $j(k)$ and $r(k)$. In practice, the majority of speech corpora are not phonetically transcribed. Furthermore, the available phonetic transcriptions may not perfectly align with the target-directed phonetic model structure of the HDM. An example of this is the anticipatory effect of the succeeding phone, which is usually empirically modeled by setting the midpoint of the phonetic transcription as the model boundary in the HDM. To overcome this arbitrariness, we will, in future work, investigate optimal segmentation of the HDM sequence in conjunction with the estimation algorithms discussed in this paper, i.e., expanding the parameter set to $\Theta = (\phi, T, W, j, r)$ [11].

ACKNOWLEDGMENT

The authors wish to thank Dr. J. Ma for help with the implementation of the EM algorithm described in Section III-A of this paper and both the editor and anonymous reviewers, whose suggestions and comments have significantly improved the presentation of this paper.

REFERENCES

[1] C. K. Chui and G. Chen, *Kalman Filtering With Real-Time Applications*. New York: Springer-Verlag, 1991, ch. 8, pp. 108–130.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. B 39, pp. 1–38, 1977.

[3] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, pp. 299–323, 1998.

[4] L. Deng and J. Ma, "A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," in *Proc. Eurospeech*, Sept. 1999, pp. 1499–1502.

[5] L. Deng and X. Shen, "Maximum likelihood in statistical estimation of dynamic systems: Decomposition algorithm and simulation results," *Signal Process.*, vol. 57, pp. 65–79, 1997.

[6] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2702–2719, 1994.

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

[8] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 431–442, Oct. 1993.

[9] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte-Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[10] L. A. Feldkamp and G. V. Puskorius, "A signal processing framework based on dynamic neural networks with application to problems in adaptation, filtering and classification," *Proc. IEEE*, vol. 86, pp. 2259–2277, Nov. 1998.

[11] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Comput.*, vol. 12, no. 4, pp. 831–864, 2000.

[12] Y. Gao, R. Bakis, J. Huang, and B. Ziang, "Multistage coarticulation model combining articulatory, formant and cepstral features," in *Proc. ICSLP*, Oct. 2000, pp. 25–28.

[13] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1999, ch. 4, pp. 156–252.

[14] ——, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1999, ch. 15, pp. 733–789.

[15] R. L. Kashyap, "Maximum likelihood identification of stochastic linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-15, pp. 25–34, Feb. 1970.

[16] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *J. R. Statist. Soc. Ser. B*, vol. 57, no. 2, pp. 425–437, 1995.

[17] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 36–50, Feb. 1979.

[18] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[19] Z. Ma, "Spontaneous speech recognition using statistical dynamic models for the vocal-tract-resonance dynamics," Ph.D. dissertation, Univ. Waterloo, Waterloo, ON, Canada, 2000.

[20] J. Ma and L. Deng, "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech," *Comput. Speech Language*, vol. 14, pp. 101–114, 2000.

[21] R. K. Mehra, "Approaches to adaptive filtering," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 693–698, Oct. 1972.

[22] J. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995, ch. 24, pp. 384–394.
[23] ——, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995, ch. 18, pp. 259–276.
[24] ——, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995, ch. 21, pp. 317–344.
[25] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–288, 1993.
[26] R. van der Merwe, A. Doucet, N. de Freitas, and E. A. Wan, "The unscented particle filter," in *Advances in Neural Information Processing Systems (NIPS13)*, T. K. Leen, T. G. Dietrich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2000.
[27] R. van der Merwe and E. A. Wan, "Efficient derivative-free Kalman filters for online learning," in *Proc. Eur. Symp. Artificial Neural Networks*, Bruges, Belgium, Apr. 2001.
[28] L. W. Nelson and E. Stear, "The simultaneous on-line estimation of parameters and states in linear systems," *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 94–98, Feb. 1976.
[29] D. Nettleton, "Convergence properties of the EM algorithm in constrained parameter spaces," *Can. J. Statist.*, vol. 27, no. 3, pp. 639–648, 1999.
[30] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 360–378, Sept. 1996.
[31] J. Picone *et al.*, "Initial evaluation of hidden dynamic models on conversational speech," in *Proc. ICASSP*, Mar. 1999, pp. 109–112.
[32] G. V. Puskorius and L. A. Feldkamp, "Neurocontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 279–297, Apr. 1994.
[33] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.
[34] H. B. Richards and J. S. Bridle, "The HDM: A segmental hidden dynamic model of coarticulation," in *Proc. ICASSP*, Mar. 1999, pp. 357–360.
[35] R. H. Shumway and D. S. Stoffer, "Dynamic linear models with switching," *J. Amer. Statist. Assoc.*, vol. 86, no. 415, pp. 763–769, 1991.
[36] ——, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.
[37] R. Togneri, J. Ma, and L. Deng, "Parameter estimation of a target-directed dynamic system model with switching states," *Signal Process.*, vol. 81, no. 5, pp. 975–987, 2001.
[38] R. Togneri and L. Deng, "An EKF-based algorithm for learning statistical hidden dynamic model parameters for phonetic recognition," in *Proc. ICASSP*, vol. 1, May 2001, pp. 465–468.
[39] E. A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. Symp. AS-SPCC*, Lake Louise, AB, Canada, Oct. 2000.
[40] ——, *Kalman Filtering and Neural Networks*, S. Haykin, Ed. New York: Wiley, 2001, Dual EKF Methods, ch. 5, p. 40.
[41] ——, *Kalman Filtering and Neural Networks*, S. Haykin, Ed. New York: Wiley, 2001, The Unscented Kalman Filter, ch. 7, p. 50.

Roberto Togneri (M'89) received the B.E. and Ph.D. degrees from the University of Western Australia (UWA), Crawley, in 1985 and 1989, respectively.

He joined the School of Electrical, Electronic, and Computer Engineering, UWA, in 1988 as senior tutor and was appointed lecturer in 1992 and senior lecturer in 1997. He is a member of the Centre for Intelligent Information Processing Systems (CIIPS) and heads the Signals and Information Processing Group at UWA. In 1999, he was employed for three-months as a visiting professor with the Speech and Information Processing Group, University of Waterloo, Waterloo, ON, Canada, and in 1993, he held a six-month visiting position with the Computer Engineering Research Group, University of Toronto, Toronto, ON. His research activities include signal processing and feature extraction of speech signals, statistical and neural network models for speech recognition, applications of spoken language technology, and related aspects of information retrieval, distribution, and communications. He has published over 20 papers in refereed conference and journal publications in the areas of spoken language and information systems and recently published the book *Fundamentals of Information Theory and Coding Design* (Boca Raton, FL: Chapman & Hall/CRC, 2002).

Dr. Togneri was the conference secretariat for the 1994 Speech Science and Technology (SST'94) conference and deputy-chair on the technical programme committee of the Fifth International Conference on Spoken Language Processing (ICSLP98).

Li Deng (S'83–M'86–SM'91) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1982 and the M.S. and Ph.D. degrees from the University of Wisconsin-Madison in 1984 and 1986, respectively.

He worked on large vocabulary automatic speech recognition with INRS Telecommunications, Montreal, QC, Canada, from 1986 to 1989. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada as Assistant Professor, where he became a tenured Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998, he was with the ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher and is currently a principal investigator for the DARPA-EARS program and affiliate Professor of electrical engineering at the University of Washington, Seattle. His research interests include acoustic-phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, noise robust speech processing, statistical methods and machine learning, nonlinear signal processing, spoken language systems, multimedia signal processing, and multimodal human-computer interaction. In these areas, he has published over 200 technical papers and book chapters and has given keynote, tutorial, and other invited lectures worldwide. He recently completed the book *Speech Processing—A Dynamic and Optimization-Oriented Approach* (New York: Marcel Dekker, 2003).

Dr. Deng served on Education Committee and Speech Processing Technical Committee of the IEEE Signal Processing Society from 1996 to 2000 and is currently serving as Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.