ELSEVIER

Signal Processing 57 (1997) 65–79

SIGNAL PROCESSING

Maximum likelihood in statistical estimation of dynamic systems: Decomposition algorithm and simulation results

Li Deng*, Xuemin Shen

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Received 23 June 1995; revised 10 April 1996 and 10 October 1996

Abstract

In this paper, we describe an efficient decomposition algorithm for parameter estimation of linear dynamical systems with the state-space formulation which contain a "drive" term as a free, unknown system parameter. The dynamical system can be viewed as a natural extension from the discrete-state hidden Markov model to its continuous-state counterpart. The focus of this paper is on unified techniques for efficient estimation of the parameters of such a model. The Expectation-Maximization (EM) algorithm is developed, in conjunction with the conventional Kalman smoothing estimators, for estimating the system parameters by maximum likelihood. The algorithm developed is applicable to either stationary or non-stationary version of the dynamic system. In particular, a decomposition technique is described in detail and is shown to reduce effectively the computational load in parameter estimation for high-dimensional systems. Simulation results are presented which demonstrate the accuracy of the proposed parameter estimation technique. © 1997 Elsevier Science B.V.

Zusammenfassung

Dieser Artikel beschreibt einen effizienten Zerlegungsalgoritmus zur Parameterschätzung linearer dynamischer Systeme. Es wird eine Zustandsraumdarstellung gewählt, die einen "Anregungsterm" als freien, unbekannten Systemparameter enthält. Das dynamische System kann als eine natürliche Erweiterung der zustandsdiskreten Modellierung mit verdeckten Markow Prozessen (HMM) zum zustandskontinuierlichen Gegenstück betrachtet werden. Ein EM-Algorithmus zur Maximum-Likelihood-Schätzung der Systemparameter wird in Verbindung mit konventionellen Schätzern mittels Kalman–Glättung entwickelt. Das Verfahren ist sowohl auf eine stationäre, wie nichtstationäre Version des dynamischen Systems anwendbar. Eine Zerlegungstechnik wird detailliert beschrieben und es wird gezeigt, daß diese den Rechenaufwand zur Parameterschätzung bei hochdimensionalen Systemen erheblich reduziert. Es werden ferner Simulationsergebnisse präsentiert, die die Genauigkeit der vorgestellten Parameterschätzungsmethode demonstrieren. © 1997 Elsevier Science B.V.

Résumé

Dans cet article, nous décrivons un algorithme efficace de décomposition d'estimation des paramètres de systèmes linéaires, ayant une formulation dans l'espace d'état qui contient un terme "d'entraînement" comme paramètre inconnu et libre du système. Le système dynamique peut être vu comme une extension naturelle du modèle de Markov caché avec états discrets à sa contre-partie avec états continus. Dans cet article, l'accent est mis sur les techniques unifiées pour l'estimation efficace des paramètres d'un tel modèle. L'algorithme de Maximisation d'Espérance est développé, en conjonction avec les estimateurs d'adoucissement de Kalman conventionnels, pour estimer les paramètres du système à l'aide du maximum de vraisemblance. L'algorithme développé peut être appliqué aux versions stationnaires aussi bien que

*Corresponding author. Tel.: 519-885-1211 ext 6041; fax: 519-746-3077; e-mail: deng@crg5.uwaterloo.ca.

PII S0165-1684(96)00186-7

non-stationnaires du système dynamique. En particulier, on décrit en détail une technique de décomposition et on montre qu'elle réduit efficacement la charge de calcul lors de l'estimation des systèmes de dimension élevée. Des résultats de simulations sont présentés, qui montrent le précision de la technique d'estimation de paramètres proposée. © 1997 Elsevier Science B.V.

Keywords: Maximum likelihood; Decomposition algorithm; Kalman filter; Dynamical system

1. Introduction

The focus of the study reported in this paper is to develop a unified technique for parameter estimation of linear dynamic systems. In particular, we are interested in efficient methods for parameter estimation for high-dimensional systems. The interest in the dynamic systems and in the associated parameter estimation problem has arisen from several different disciplines. In statistics, linear regression techniques can be generalized to include temporal evolution of the input variable obeying laws governed by an autoregressive process. This generalization of the regression model unfolding in time gives rise to a linear dynamic system model. In control theory, the dynamic system has been widely used as a model for noisy observations on stochastic linearly behaved physical plants. In signal processing, a dynamic system can be viewed as a continuous-state counterpart of the conventional discrete-state hidden Markov model; the latter has found successful applications in speech processing and other signal processing areas (cf. [22]).

The success of the dynamic system model, when applied to the various disciplinary areas, depends strongly on the degree to which the model represents the true physical process being modeled. The linear form of the dynamic system model has the advantage of the mathematical tractability and (arguably) is a reasonably good representation of the global characteristics of many physical processes. Once the structure of a dynamic system is fixed (linear form in many practical applications), then the only factor determining how well the model represents a particular physical process being modeled is the parameters of the system.

Because of the significance of the parameter estimation problem discussed above and due to its well-known difficulties, we devote the current study to a unified maximum-likelihood-based approach to the problem's solution. First, we develop in this study an Expectation-Maximization (EM) algorithm, which includes Kalman smoothing estimators as an integrated component, and then we further develop a computationally efficient technique based on the principle of decomposing a high-dimensional system into several low-dimensional ones. The decomposition technique is demonstrated to have significantly reduced the computational requirement in parameter estimation in our simulation experiments reported in this paper.

2. Linear dynamic system and the associated EM algorithm

We study the linear dynamic system given by

$$x_{k+1} = Ax_k + T + w_k, \qquad y_k = Cx_k + v_k, \qquad Ex_0 = \mu_0, \qquad E(x_0 x_0^{\mathrm{T}}) = P_0, \tag{1}$$

where $x_k \in \mathscr{R}^n$ is a state vector, $y_k \in \mathscr{R}^m$ is an output vector, w_k and v_k are uncorrelated zero-mean Gaussian vectors with covariances $E(w_k w_k^{\mathrm{T}}) = Q$ and $E(v_k v_k^{\mathrm{T}}) = R$, respectively. Parameter T can be interpreted as an abstract 'drive' force of or control input to the system.

In parameter estimation problems based on maximum likelihood, it is usually more convenient to work with the negative of the logarithm of the likelihood function [21]. It is possible to do so because the logarithm is a monotonic function. The cost function for the system (1) with respect to θ is

$$J(\theta) = -\log L(x, y, \theta), \tag{2}$$

where θ is a vector of unknown parameters in A, T, C, Q, R.

In order to develop a procedure for estimating the parameters in the state-space model defined by (1), we note first that the joint log likelihood of $x = [x_0 x_1, \ldots, x_N]$, $y = [y_0 y_1, \ldots, y_N]$ can be written in the form [25]

$$\log L(x, y, \theta) = -\frac{1}{2}\sum_{k=1}^{N} \log Q + (x_k - Ax_{k-1} - T)^{\mathrm{T}} Q^{-1}(x_k - Ax_{k-1} - T)$$
$$-\frac{1}{2}\sum_{k=0}^{N} \log R + (y_k - Cx_k)^{\mathrm{T}} R^{-1}(y_k - Cx_k) + \text{constant}. \tag{3}$$

We assume that there are no constraints on the structure of the system matrices. When the states are unobservable, the forecasting and smoothing forms of the Kalman filter will be used to estimate the unobserved (continuously valued) states x_k. The forecast and smoothed values in the Kalman filter estimator will depend on the initial values assumed for the system parameters. The new estimates for the system parameters can be obtained by an iterative technique using the Expectation-Maximum (EM) algorithm.

The EM algorithm is a two-step iterative scheme for maximizing the likelihood function. Each iteration of the EM algorithm involves two steps called the Expectation step (E-step) and the Maximization step (M-step). A formal introduction of the EM algorithm appeared in [4]. Since then, many research works, with the focus of applications in particular, have been published in the literatures [7, 8, 11, 17, 24, 3, 15]. The EM algorithm guarantees an increase (or strictly non-decrease) of the likelihood upon each iteration of the algorithm and guarantees convergence of the iteration to a stationary point for an exponential family [4]. In the E-step, conditional expectation of log joint likelihood between x and y, given observation y, is computed, and, in the case of hidden state x, sufficient statistics containing data series x_k, $k = 1, 2, \ldots, N$, are estimated which are conditioned also on y. The results of the E-step are then used to obtain a new estimate of $\theta(A, T, C, Q, R)$ in the so called M-step. The new estimate is then fed back to the E-step, and the E-step and the M-step iterate themselves until convergence. The convergence properties of the EM algorithm have been discussed in [2, 26]. We now apply the EM algorithm to our dynamic system model (1).

2.1. E-step

The E-step of the EM algorithm is to compute the conditional expectation of $\log L(x, y)$ given y. The expectation is defined over the joint space x and y. However, the conditioning on y eliminates the y space in the expectation, and hence the conditional expectation is effectively taken over (hidden) space x only.

The conditional expectation can be written as

$$U(x, y, \theta) = E\{\log L(x, y)|y\}$$
$$= -\frac{N}{2}\log Q - \frac{N+1}{2}\log R - \frac{1}{2}\sum_{k=1}^{N} E_N[e_{k1}^{\mathrm{T}} Q^{-1} e_{k1}|y] - \frac{1}{2}\sum_{k=0}^{N} E_N[e_{k2}^{\mathrm{T}} R^{-1} e_{k2}|y], \tag{4}$$

where $e_{k1} = x_{k+1} - Ax_k - T$, $e_{k2} = y_k - Cx_k$, and E_N denotes the conditional expectation based on N samples of data. The estimates Q and R must satisfy

$$\frac{\partial U(x, y, \theta)}{\partial Q} = 0, \qquad \frac{\partial U(x, y, \theta)}{\partial R} = 0. \tag{5}$$

By

$$\frac{\partial}{\partial Q}\log Q = Q^{-1}, \qquad \frac{\partial}{\partial R}\log R = R^{-1}, \qquad \frac{\partial}{\partial Q} e_{k1}^{\mathrm{T}} Q e_{k1} = e_{k1} e_{k1}^{\mathrm{T}}, \qquad \frac{\partial}{\partial R} e_{k2}^{\mathrm{T}} R e_{k2} = e_{k2} e_{k2}^{\mathrm{T}},$$

and after setting the partial derivatives of (4), with respect to Q and R, to zero, we have

$$\frac{\partial U}{\partial Q^{-1}} = -\frac{N}{2}\frac{\partial}{\partial Q^{-1}}\log Q^{-1} + \frac{1}{2}\sum_{k=1}^{N}\frac{\partial}{\partial Q^{-1}}E_N(e_{k1}^{\mathrm{T}}Q^{-1}e_{k1}|y)$$

$$= \frac{N}{2}Q - \frac{1}{2}\sum_{k=1}^{N}E_N(e_{k1}e_{k1}^{\mathrm{T}}|y) = 0,$$

$$\frac{\partial U}{\partial R^{-1}} = -\frac{N+1}{2}\frac{\partial}{\partial R^{-1}}\log R^{-1} + \frac{1}{2}\sum_{k=0}^{N}\frac{\partial}{\partial R^{-1}}E_N(e_{k2}^{\mathrm{T}}R^{-1}e_{k2}|y)$$

$$= \frac{N+1}{2}R - \frac{1}{2}\sum_{k=0}^{N}E_N(e_{k2}e_{k2}^{\mathrm{T}}|y) = 0.$$

Then, the estimates Q and R become

$$\bar{Q} = \frac{1}{N}\sum_{k=1}^{N}E_N(e_{k1}e_{k1}^{\mathrm{T}}|y), \qquad \bar{R} = \frac{1}{N+1}\sum_{k=0}^{N}E_N(e_{k2}e_{k2}^{\mathrm{T}}|y). \tag{6}$$

Further, using $a^{\mathrm{T}}Va = \mathrm{trace}\{Va^{\mathrm{T}}a\}$, we have

$$\begin{aligned}
\sum_{k=1}^{N}E_N(e_{k1}^{\mathrm{T}}\bar{Q}^{-1}e_{k1}|y) &= \sum_{k=1}^{N}E_N\left\{e_{k1}^{\mathrm{T}}\left[\frac{1}{N}\sum_{k=1}^{N}E_N(e_{k1}e_{k1}^{\mathrm{T}}|y)\right]^{-1}e_{k1}|y\right\}\\
&= \mathrm{trace}\left\{\left[\frac{1}{N}\sum_{k=1}^{N}E_N(e_{k1}e_{k1}^{\mathrm{T}}|y)\right]^{-1}\left[\sum_{k=1}^{N}E_N(e_{k1}e_{k1}^{\mathrm{T}}|y)\right]\right\}\\
&= N\,\mathrm{trace}\,(I) = \text{constant},\\
\sum_{k=0}^{N}E_N(e_{k2}^{\mathrm{T}}\bar{R}^{-1}e_{k2}|y) &= \sum_{k=0}^{N}E_N\left\{e_{k2}^{\mathrm{T}}\left[\frac{1}{N}\sum_{k=1}^{N}E_N(e_{k2}e_{k2}^{\mathrm{T}}|y)\right]^{-1}e_{k2}|y\right\}\\
&= \mathrm{trace}\left\{\left[\frac{1}{N+1}\sum_{k=0}^{N}E_N(e_{k2}e_{k2}^{\mathrm{T}}|y)\right]^{-1}\left[\sum_{k=0}^{N}E_N(e_{k2}e_{k2}^{\mathrm{T}}|y)\right]\right\}\\
&= (N+1)\,\mathrm{trace}(I) = \text{constant}.
\end{aligned} \tag{7}$$

Finally, substitution of (6) and (7) back into (4) for $U(x, y, \theta)$ leads to the following result:

$$U(x, y, \theta) = -\frac{N}{2}\log\left\{\frac{1}{N}\sum_{k=1}^{N}E_N[(x_k - Ax_{k-1} - T)^2|y]\right\}$$

$$-\frac{N+1}{2}\log\left\{\frac{1}{N+1}\sum_{k=0}^{N}E_N[(y_k - Cx_k)^2|y]\right\} + \text{constant}.$$

2.2. M-step

Given the conditional expectation above, the M-step aims at minimization of the following two separate quantities (both being the expected values of a standard least-squares objective function):

$$U_1(A,T) = \frac{1}{N}\sum_{k=1}^{N} E_N[(x_k - Ax_{k-1} - T)^2|y],$$
$$U_2(C) = \frac{1}{N+1}\sum_{k=0}^{N} E_N[(y_k - Cx_k)^2|y]. \tag{8}$$

Since the order of expectation and differentiation can be interchanged, we obtain the parameter estimates by solving

$$\frac{\partial U_1(A,T)}{\partial A} = -\sum_{k=1}^{N} E_N\left[\frac{\partial}{\partial A}(x_k - Ax_{k-1} - T)^2|y\right] = 0,$$
$$\frac{\partial U_1(A,T)}{\partial T} = -\sum_{k=1}^{N} E_N\left[\frac{\partial}{\partial T}(x_k - Ax_{k-1} - T)^2|y\right] = 0, \tag{9}$$
$$\frac{\partial U_2(C)}{\partial C} = -\sum_{k=1}^{N} E_N\left[\frac{\partial}{\partial C}(x_k - Ax_{k-1})^2|y\right] = 0,$$

or

$$\bar{A}\sum_{k=1}^{N} E_N(x_{k-1}x_{k-1}^{\mathrm{T}}|y) + \bar{T}\sum_{k=1}^{N} E_N(x_{k-1}^{\mathrm{T}}|y) = \sum_{k=1}^{N} E_N(x_k x_{k-1}^{\mathrm{T}}|y),$$
$$\bar{A}\sum_{k=1}^{N} E_N(x_{k-1}|y) + N\bar{T} = \sum_{k=1}^{N} E_N(x_k|y), \tag{10}$$
$$-\bar{C}\sum_{k=0}^{N} E_N(x_k x_k^{\mathrm{T}}|y) + \sum_{k=0}^{N} E_N(y_k x_k^{\mathrm{T}}|y) = 0.$$

In the matrix form, (10) becomes

$$[\bar{A}\quad \bar{T}] = \left[\sum_{k=1}^{N} E_N(x_k x_{k-1}^{\mathrm{T}}|y) \sum_{k=1}^{N} E_N(x_k|y)\right]\begin{bmatrix} \sum_{k=1}^{N} E_N(x_{k-1}x_{k-1}^{\mathrm{T}}|y) & \sum_{k=1}^{N} E_N(x_{k-1}|y) \\ \sum_{k=1}^{N} E_N(x_{k-1}^{\mathrm{T}}|y) & N \end{bmatrix}^{-1},$$
$$\bar{C} = \left[\sum_{k=0}^{N} E_N(y_k x_k^{\mathrm{T}}|y)\right]\left[\sum_{k=0}^{N} E_N(x_k x_k^{\mathrm{T}}|y)\right]^{-1}. \tag{11}$$

Substituting Eq. (11) into Eq. (6), we obtain

$$\bar{Q} = \frac{1}{N}\sum_{k=1}^{N} E_N[(x_k - \bar{A}x_{k-1} - \bar{T})^2|y]$$
$$= \frac{1}{N}\sum_{k=1}^{N} E_N\{(x_k x_k^{\mathrm{T}}|y) - [\bar{A}\quad \bar{T}][x_k x_{k-1}^{\mathrm{T}}|y\quad x_k|y]^{\mathrm{T}}\},$$

$$\bar{R} = \frac{1}{N+1}\sum_{k=0}^{N} E_N[(y_k - \bar{C}x_k)^2|y]$$

$$= \frac{1}{N+1}\left[\sum_{k=0}^{N} E_N(y_k y_k^{\mathrm{T}}|y) - \bar{C}\left(\sum_{k=0}^{N} E_N(y_k x_k^{\mathrm{T}}|y)\right)^{\mathrm{T}}\right]. \tag{12}$$

To perform each iteration of the EM algorithm, it remains to evaluate all the conditional expectations in Eqs. (11) and (12). Such evaluation is described in the next section for the stationary version of the dynamic system.

3. Stationary stochastic system

Case 1. Perfect observation for state x. If the data series x_k, $k = 0, 1, \ldots, N$, i.e., the states of the system (1) are completely observable, then $E_N(x_k|y) = x_k$, $E_N(x_k x_k^{\mathrm{T}}|y) = x_k x_k^{\mathrm{T}}$ and $E_N(y_k x_k^{\mathrm{T}}|y) = y_k x_k^{\mathrm{T}}$. Eqs. (11) and (12) can then be directly applied to estimate the parameters A, T, C, Q and R.

Case 2. Incomplete or unobservable state x. In the case when the states x_k of the system (1) are unobservable and the log likelihood depends on the unobserved data series x_k, $k = 0, 1, \ldots, N$, we need to evaluate the various conditional expectations on the observed series y_k, $k = 0, 1, \ldots, N$, which appear in Eqs. (11) and (12) and are sufficient statistics for the parameter estimation.

Denote the conditional expectation

$$E_N\{x_k^{\mathrm{T}}|y\} = \hat{x}_{k/N}^{\mathrm{T}}. \tag{13}$$

Then, to evaluate $E_N[x_k x_k^{\mathrm{T}}|y]$, we use the standard formula

$$E_N\{[x_k - E_N(x_k)][x_k - E_N(x_k)]^{\mathrm{T}}|y\} = E_N[x_k x_k^{\mathrm{T}}|y] - E_N[x_k|y]E_N[x_k^{\mathrm{T}}|y] \tag{14}$$

or

$$P_{k/N} = E_N[x_k x_k^{\mathrm{T}}|y] - \hat{x}_{k/N}\hat{x}_{k/N}^{\mathrm{T}}$$

to obtain

$$E_N[x_k x_k^{\mathrm{T}}|y] = \hat{x}_{k/N}\hat{x}_{k/N} + P_{k/N}. \tag{15}$$

In using the EM algorithm to obtain maximum likelihood estimates of the parameters of Eq. (1), i.e., to evaluate Eqs. (11) and (12), we require the following quantities at each iteration of the algorithm:

$$\begin{aligned}
&E\{x_k^{\mathrm{T}}|y\} = \hat{x}_{k/N}^{\mathrm{T}},\\
&E\{x_k x_k^{\mathrm{T}}|y\} = P_{k/N} + \hat{x}_{k/N}\hat{x}_{k/N}^{\mathrm{T}},\\
&E\{x_k x_{k-1}^{\mathrm{T}}|y\} = P_{k,k-1/N} + \hat{x}_{k/N}\hat{x}_{k-1/N}^{\mathrm{T}},\\
&E\{y_k x_k^{\mathrm{T}}|y\} = y_k\hat{x}_{k/N}^{\mathrm{T}},\\
&E\{y_k y_k^{\mathrm{T}}\} = y_k y_k^{\mathrm{T}}.
\end{aligned} \tag{16}$$

That is, we can use the fixed interval smoothing form of the Kalman filter to compute the required statistics. It consists of a backward pass that follows the standard Kalman filter forward recursions. In addition, in both

the forward and the backward passes, we need some additional recursions for the computation of the cross covariance. All the necessary recursions are summarized in the following.

Forward recursions:

$$\begin{aligned}
&\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k e_k,\\
&\hat{x}_{k+1/k} = A\hat{x}_{k/k} + T,\\
&e_k = y_k - C\hat{x}_{k/k-1},\\
&K_k = P_{k/k-1}C^{\mathrm{T}}P_{e_k}^{-1},\\
&P_{e_k} = CP_{k/k-1}C^{\mathrm{T}} + R,\\
&P_{k/k} = P_{k/k-1} - K_k P_{e_k} K_k^{\mathrm{T}},\\
&P_{k,k-1/k} = (I - K_k C)AP_{k-1/k-1},\\
&P_{k+1/k} = AP_{k/k}A^{\mathrm{T}} + Q.
\end{aligned} \tag{17}$$

Backward recursions:

$$\begin{aligned}
&\Gamma_k = P_{k-1/k-1}A^{\mathrm{T}}P_{k/k-1}^{-1},\\
&\hat{x}_{k-1/N} = \hat{x}_{k-1/k-1} + \Gamma_k[\hat{x}_{k/N} - \hat{x}_{k/k-1}],\\
&P_{k-1/N} = P_{k-1/k-1} + \Gamma_k[P_{k/N} - P_{k/k-1}]\Gamma_k^{\mathrm{T}},\\
&P_{k,k-1/N} = P_{k,k-1/k} + [P_{k/N} - P_{k/k}]P_{k/k}^{-1}P_{k,k-1/k}.
\end{aligned} \tag{18}$$

Using Eqs. (16)–(18), the parameter estimates of system (1) for unobservable state x_k can be completely obtained by Eqs. (11) and (12).

4. Nonstationary stochastic system

We now consider the nonstationary version of the stochastic linear system

$$x_{k+1} = A(s)x_k + T(s) + w_k, \qquad y_k = C(s)x_k + v_k, \qquad E\{x_0\} = \mu_0, \qquad E\{x_o x_o^{\mathrm{T}}\} = P_0(s), \tag{19}$$

where $x_k \in \mathscr{R}^n$ is a state vector, $y_k \in \mathscr{R}^m$ is an output vector, w_k and v_k are uncorrelated, zero-mean Gaussian vectors with covariance $E\{w_k w_k^{\mathrm{T}}\} = Q(s)$ and $E\{v_k v_k^{\mathrm{T}}\} = R(s)$. The entries of the matrices $A(s)$, $T(s)$, $C(s)$, $Q(s)$ and $R(s)$ are random signals depending on the distinctive region (or distinctive mode, indexed by s).

Due to the nonstationary nature of the system, one needs multiple-run observations in order to have sufficient data to make reliable estimates of the system parameters. The extension from the single-run approach to the multiple-run one is straightforward: it involves simply summing the appropriate sufficient statistics over the different runs and the corresponding cost function is

$$\begin{aligned}
\log L(x, y, \theta) = &-\frac{1}{2}\sum_{l=1}^{r}\left\{\sum_{k=1}^{N_l}[\log Q + (x_k - Ax_{k-1} - T)^{\mathrm{T}}Q^{-1}(x_k - Ax_{k-1} - T)]\right\}\\
&-\frac{1}{2}\sum_{l=1}^{r}\left\{\sum_{k=0}^{N_l}[\log R + (y_k - Cx_k)^{\mathrm{T}}R^{-1}(y_k - Cx_k)]\right\} + \text{constant},
\end{aligned} \tag{20}$$

where r is the total number of runs and N_l is the number of observations in the lth run. It is assumed that the discrete region for each run $s(k)$ can be observed. The parameter estimation of the nonstationary stochastic

system is then

$$
\begin{aligned}
[\hat{A}\ \hat{T}]_s &= \left[\sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(x_{kl}x_{(k-1)l}^{\mathrm{T}} \sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(x_{kl}|y)\right]_s \\
&\quad \times \begin{bmatrix} \sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(x_{(k-1)l}x_{(k-1)l}^{\mathrm{T}}|y) & \sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(x_{(k-1)l}|y) \\ \sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(x_{(k-1)l}^{\mathrm{T}}|y) & \sum_{l=1}^{r} N_l \end{bmatrix}_s^{-1}, \\
\hat{C}_s &= \left[\sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N(y_{kl}x_{kl}^{\mathrm{T}}|y)\right]_s \left[\sum_{l=1}^{r}\sum_{k=0}^{N_l} E_N(x_{kl}x_{kl}^{\mathrm{T}}|y)\right]_s^{-1}, \\
\hat{Q}_s &= \frac{1}{\sum_{l=1}^{r} N_l}\sum_{l=1}^{r}\sum_{k=1}^{N_l} E_N[(x_{kl} - \hat{A}x_{(k-1)l} - \hat{T})^2|y]_s \qquad (21) \\
&= \frac{1}{\sum_{l=1}^{r} N_l}\sum_{l=1}^{r}\sum_{k=1}^{N_l}\{E_N(x_{kl}x_{kl}^{\mathrm{T}}|y) - [\hat{A}\ \hat{T}]\,[E_N(x_{kl}x_{(k-1)l}^{\mathrm{T}}|y)\,E_N(x_{kl}|y)]^{\mathrm{T}}\}_s, \\
\hat{R}_s &= \frac{1}{\sum_{l=1}^{r}(N_l+1)}\sum_{l=1}^{r}\sum_{k=0}^{N_l} E_N[(y_{kl} - \hat{C}x_{kl})^2|y]_s \\
&= \frac{1}{\sum_{l=1}^{r}(N_l+1)}\sum_{l=1}^{r}\left[\left(\sum_{k=0}^{N_l} y_{kl}y_{kl}^{\mathrm{T}}\right) - \hat{C}\sum_{k=0}^{N_l} E_N(y_{kl}x_{kl}^{\mathrm{T}}|y)^{\mathrm{T}}\right]_s,
\end{aligned}
$$

where $s = 1, 2, \ldots, m$, is the sth invariant region of the model.

The EM algorithm for the nonstationary version of the dynamic system, firstly, involves at each iteration the computation of the sufficient statistics (i.e., all the conditional expectations in Eq. (21)) using the recursions described in Section 3 with the previous estimates of the model parameters (E-step). The new estimates for the system parameters are then obtained using the sufficient statistics according to Eq. (21) (M-step).

5. Decomposition algorithm for parameter estimation of high-dimensional system

In many practical applications, the dynamic system model for physical plants needs to be of relatively high dimensionality. For example, when a complete dynamic system model is used for describing articulatory motions in the human vocal tract, the desirable dimensionality would be of the order of 10 to 20. The numerical calculation for high-dimensional systems' parameter estimation may be impractical due to the extraordinary demand for the amount of computation. In order to reduce the high computational load, we propose a method that allows split of a large-scale estimation problem into a set of simpler subsystem problems. Many estimation algorithms for large-scale systems have appeared in the research literature. Here we adopt the one which appeared in [13, 12] as applied to the dynamic system we are currently studying.

The algorithm uses a hierarchical structure to perform successive orthogonalizations on the measurement subspaces of each subsystem in order to provide the optimal estimates. In building the subsystem estimators, we use a standard optimization scheme, treating the outputs of preceding estimators as known inputs to the subsequent estimator being optimized. The overall estimator obtained by this process as the union of the subsystem estimators is stable and suboptimal. With an additional computational effort, the degree of suboptimality can be computed with respect to the globally optimal estimator taken as a reference.

Rewriting Eq. (1) as N coupled subsystems, we have

$$\begin{aligned} x_i(k+1) &= A_{ii}x_i(k) + \sum_{j=1, j\neq i}^{N} A_{ij}x_j(k) + T_i + w_i(k), \\ y_i(k) &= C_{ii}x_i(k) + \sum_{j=1, j\neq i}^{N} C_{ij}x_j(k) + v_i(k). \end{aligned} \tag{22}$$

Define the adjusted measurement for the ith subsystem as

$$y_i^*(k) = C_{ii}x_i(k) + v_i^*(k), \tag{23}$$

where $v_i^*(k)$ is the adjusted zero-mean measurement noise:

$$v_i^*(k) = v_i(k) + \sum_{j=1, j\neq i}^{N} C_{ij}\,\tilde{x}_j(k|k-1), \tag{24}$$

with $\tilde{x}_j(k|k-1) = x_j(k) - \hat{x}(k|k-1)$ representing the predicted estimation error. The updated estimate $\hat{x}_i(k|k)$ is now given by

$$\hat{x}_i(k|k) = \hat{x}_i(k|k-1) + K_{ii}^*(k)e_i(k), \tag{25}$$

where K_{ii}^* is the gain matrix to be defined and $e_i(k)$ is the innovation sequence:

$$e_i(k) = C_{ii}x_i(k|k-1) + v_i^*(k) = y_i^*(k) - C_{ii}\,\hat{x}_i(k|k-1). \tag{26}$$

The vector $\hat{x}_i(k|k-1)$ is the predicted estimate, which can be expressed as

$$\hat{x}_i(k+1|k) = \sum_{j=1}^{N} A_{ij}\,\hat{x}_j(k|k) + T_i. \tag{27}$$

The overall estimator is then the simple union of the subsystem estimators

$$\hat{x}(k|k) = \hat{x}(k|k-1) + K_D^*(k)e(k), \tag{28}$$

where

$$K_D^*(k) = \mathrm{diag}[K_{11}^*(k), K_{22}^*(k), \ldots, K_{NN}^*(k)], \tag{29}$$

$$e(k) = y(k) - C\,\hat{x}(k|k-1). \tag{30}$$

The update and prediction errors can be obtained as

$$\tilde{x}(k|k) = (I - K_D^*(k)C)\,\tilde{x}(k|k-1) + K_D^*(k)v(k) \tag{31}$$

and

$$\tilde{x}(k+1|k) = A\,\tilde{x}(k|k) + w(k), \tag{32}$$

with the covariance

$$P^*(k|k) = \Phi^*(k)P^*(k|k-1)\Phi^{*\mathrm{T}}(k) + K_D^*(k)R_vK_D^{*\mathrm{T}}(k), \tag{33}$$

$$P^*(k+1|k) = AP^*(k|k)A^{\mathrm{T}} + R_w. \tag{34}$$

The matrix $\Phi^* = \{\Phi^*_{ij}\}$, which is defined by

$$\Phi^*(k) = I - K^*_D(k)C, \tag{35}$$

has the block structure

$$\Phi^*_{ij}(k) = \begin{cases} I_i - K^*_{ii}(k)C_{ii}, & i = j, \\ -K^*_{ii}(k)C_{ij}, & i \neq j. \end{cases} \tag{36}$$

Then, the submatrices $P^*_{ij}(k|k)$ and $P^*_{ij}(k|k-1)$ can be formulated as

$$P^*_{ij}(k|k) = \sum_{p=1}^{N}\sum_{q=1}^{N} \Phi^*_{ip}(k)P^*_{pq}(k|k-1)\Phi^{*\mathrm{T}}_{jq}(k) + K^*_{ii}(k)R^{ij}_v(k)K^{*\mathrm{T}}_{jj}(k), \quad i,j = 1,2,\ldots,N, \tag{37}$$

and

$$P^*_{ij}(k|k-1) = \sum_{p=1}^{N}\sum_{q=1}^{N} A_{ip}(k)P^*_{pq}(k-1|k-1)A^{\mathrm{T}}_{jq}(k) + Q^{ij}_w(k), \quad i,j = 1,2,\ldots,N. \tag{38}$$

The diagonal blocks of the gain matrix are computed by

$$K^*_{ii}(k) = \sum_{j=1}^{N} P^*_{ij}(k|k-1)C^{\mathrm{T}}_{ij}\left[\sum_{j=1}^{N} C_{ij}\sum_{l=1}^{N} P^*_{jl}(k|k-1)C^{\mathrm{T}}_{il} + R^{ii}_v(k)\right]^{-1}, \quad i = 1,2,\ldots,N, \tag{39}$$

and

$$P_{ii}(k|k) = \sum_{p=1}^{N}\sum_{q=1}^{N} \Phi_{ip}(k)P_{pq}(k|k-1)\Phi^{\mathrm{T}}_{iq}(k) + K_{ii}(k)R^{ii}_v(k)K^{\mathrm{T}}_{ii}(k). \tag{40}$$

Note that although the above algorithm and the global Kalman filter are algebraically equivalent, the numerical properties of the decomposed filter are significantly better since the filter calculation is performed on low-order blocks of subsystem equations. Further, for high-dimensional systems, the above decomposition algorithm gives substantial savings in computation time. The quantification of the computational saving from the decomposition algorithm as compared to that of the global Kalman filter has been discussed in detail in [12].

Now, given the estimates $\hat{x}(k|k)$ and $\hat{x}(k+1|k)$ (i.e., the conditional expectations as sufficient statistics) as efficiently computed above and using the backward recursions (Eq. (18)), the system parameters can be estimated from Eqs. (11) and (12) as in the case for the global Kalman filter.

6. Simulation results

In order to demonstrate the proposed method, the system (1) with the following system matrices is simulated using random number generators in our numerical experiments:

$$A = \begin{bmatrix} 0.9 & 0.0 & -0.3 \\ 0.0 & 0.7 & 0.0 \\ 0.2 & 0.3 & 0.6 \end{bmatrix}, \qquad T = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.5 \end{bmatrix},$$

$$C = I_{3\times 3}, \qquad Q = R = 0.1 \times I_{3\times 3}, \qquad x_0 = [1\ \ 1\ \ 1]^{\mathrm{T}},$$

where I is the identity matrix. After the observations, y, from the above system are made available, various parameter estimation methods described in the preceding sections are used to obtain the system parameters, allowing us to examine their deviation from the exact system parameters as a means to assess the accuracy of the estimation methods. A summary of the results for three cases is provided below.

Case (i). The state of the system x is observable. In this case, the true x obtained from the simulation is provided to the estimation algorithm and no hidden variables are presents. (Hence, no iteration is required in the estimation.) Estimates of all the system parameters can be obtained using Eq. (13) for single output run, where $N=100$ (data simulation length) is used in the model simulation. The system parameter estimates are

$$\bar{A}=\begin{bmatrix} 0.89 & 0.08 & -0.31 \\ -0.01 & 0.68 & -0.10 \\ 0.23 & 0.26 & 0.59 \end{bmatrix}, \quad \bar{T}=\begin{bmatrix} -0.04 \\ 0.02 \\ 0.53 \end{bmatrix}, \quad \bar{C}=\begin{bmatrix} 1.01 & 0.02 & 0.01 \\ 0.02 & 1.03 & -0.01 \\ 0.02 & -0.01 & 1.02 \end{bmatrix},$$

$$\bar{Q}=\begin{bmatrix} 0.094 & 0.02 & 0.01 \\ & 0.11 & -0.01 \\ & & 0.098 \end{bmatrix}, \quad \bar{R}=\begin{bmatrix} 0.09 & 0.003 & 0.002 \\ & 0.08 & -0.001 \\ & & 0.11 \end{bmatrix}.$$

These estimates are very close to the exact system parameters.

Case (ii). The state of the system x is unobservable. In order to apply the EM algorithm, initial values are required for the parameters. In the simulation, we set the initial system parameters as

$$A_0 = C_0 = 0.5 \times I_{3\times 3}, \qquad T_0 = [0.5\ 0.5\ 0.5]^{\mathrm{T}},$$
$$Q_0 = R_0 = 0.05 \times I_{3\times 3}, \qquad P_0 = 0.1 \times I_{3\times 3}.$$

In simulating the system, we generated a total of 150 data points as the observation data y. The parameter estimates of the system after one, two, three and four iterations are shown in Tables 1 and 2. Table 1 shows the estimates having all the 150 data points as one single run (with estimation procedure described in Section 3), and Table 2 shows the estimates with the 150 data points broken down into three separate runs (containing 40, 50 and 60 points for each run, and with estimation equations (21) described in Section 4). We first observe that in both cases, upon each iteration, the estimates of the parameters are approaching the exact parameters used to simulate the system. At the fourth iteration, the estimates of most of the parameters are reasonably close to the exact parameters, although not nearly as close as for Case (i) where the exact state information x is made available to the estimation algorithm. Second, we observe that the estimates of Table 2 (multiple runs) are slightly superior to those of Table 2 with single run (superiority in the sense of being closer to the exact parameters). This superiority is likely to result from the fact that breaking the output data points into several pieces gives the EM algorithm more chances to escape from the local optimum in its maximum likelihood estimation.

Case (iii). Decomposition approach (state x is unobservable). Given the simulated data points y which are identical to those used to obtain the results of Table 2 (multiple runs), the computationally more efficient decomposition algorithm (described in Section 5) is used to estimate the unobservable state x and then Eqs. (11) and (12) are used for the parameter estimation. In applying the decomposition algorithm, we treat each subsystem as a scalar system (with a total of three systems) and assume that the measurement data y for subestimators are synchronized with each other. The parameter estimates after one, two, three and four iterations of the EM algorithm are shown in Table 3. Comparing the results of Tables 2 and 3, with the only difference being the use of the global Kalman filter (Section 4) versus the use of the decomposition algorithm (Section 5) for state x estimation, we observe only a very slight degradation of the estimation quality from the former to the latter. For this simulation example, such degradation is negligible while significant computation efficiency has been gained.

Table 1
Single output run with unknown state x ($N = 100$)

	A			T			C			Q			R	
Exact	0.9	0.0	−0.30	0.0	1.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	
	0.0	0.7	0.0	0.0	0.0	1.0	0.0		0.1	0.0		0.1	0.0	
	0.2	0.3	0.6	0.5	0.0	0.0	1.0			0.1			0.1	
Initial	0.5	0.0	0.0	0.5	0.5	0.0	0.0	0.05	0.0	0.0	0.05	0.0	0.0	
	0.0	0.5	0.0	0.5	0.0	0.5	0.0		0.05	0.0		0.05	0.0	
	0.0	0.0	0.5	0.5	0.0	0.0	0.5			0.05			0.05	
Iteration 1	0.63	−0.11	−0.10	−0.30	1.86	0.09	−0.28	0.28	0.07	0.06	0.19	0.11	0.12	
	0.21	0.55	0.15	0.22	0.22	1.57	−0.29		0.21	0.22		0.24	0.09	
	0.11	0.13	0.35	0.33	0.18	−0.11	1.24			0.17			0.18	
Iteration 2	0.70	0.09	−0.15	−0.22	1.47	−0.05	−0.15	0.19	0.05	−0.05	0.15	0.07	0.08	
	0.03	0.60	0.03	0.17	0.15	1.35	−0.14		0.17	0.15		0.18	0.06	
	0.13	0.17	0.46	0.40	−0.10	−0.06	1.18			0.12			0.14	
Iteration 3	0.75	0.05	−0.21	−0.15	1.21	−0.02	0.10	0.16	−0.01	−0.03	0.13	0.04	0.07	
	0.02	0.62	0.03	0.12	0.09	1.07	−0.09		0.15	0.09		0.16	0.05	
	0.15	0.20	0.51	0.43	0.06	−0.05	1.15			0.09			0.13	
Iteration 4	0.78	−0.04	−0.25	−0.11	1.16	0.02	0.07	0.14	−0.03	−0.04	0.12	0.02	0.06	
	0.03	0.63	−0.02	0.09	0.06	1.07	−0.05		0.13	0.0		0.15	0.05	
	0.16	0.22	0.55	0.45	0.03	−0.04	1.13			0.11			0.13	

Table 2
Multiple output runs with unknown state x ($N_1 = 40$, $N_2 = 50$, $N_3 = 60$)

	A			T			C			Q			R	
Exact	0.9	0.0	−0.30	0.0	1.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	
	0.0	0.7	0.0	0.0	0.0	1.0	0.0		0.1	0.0		0.1	0.0	
	0.2	0.3	0.6	0.5	0.0	0.0	1.0			0.1			0.1	
Initial	0.5	0.0	0.0	0.5	0.5	0.0	0.0	0.05	0.0	0.0	0.05	0.0	0.0	
	0.0	0.5	0.0	0.5	0.0	0.5	0.0		0.05	0.0		0.05	0.0	
	0.0	0.0	0.5	0.5	0.0	0.0	0.5			0.05			0.05	
Iteration 1	0.83	−0.11	−0.22	−0.10	1.36	0.02	−0.18	0.09	0.01	0.01	0.07	0.01	0.01	
	0.09	0.57	0.05	0.22	0.10	1.25	−0.19		0.08	0.02		0.08	0.0	
	0.13	0.16	0.65	0.33	0.10	−0.01	1.17			0.07			0.08	
Iteration 2	0.85	−0.08	−0.25	0.07	1.07	−0.01	−0.02	0.093	0.0	−0.01	0.082	0.01	0.0	
	0.03	0.65	0.03	0.12	0.05	1.05	0.05		0.093	0.01		0.085	0.0	
	0.15	0.25	0.63	0.43	−0.10	−0.03	1.05			0.08			0.09	
Iteration 3	0.88	−0.06	−0.31	−0.03	1.01	−0.02	0.02	0.11	−0.01	−0.01	0.11	0.01	0.0	
	0.02	0.68	0.03	0.08	−0.02	1.03	−0.01		0.095	0.0		0.09	0.0	
	0.18	0.26	0.60	0.46	0.02	−0.04	1.03			0.091			0.11	
Iteration 4	0.88	−0.04	−0.29	−0.05	1.01	0.02	0.00	0.1	−0.01	0.0	0.11	0.02	0.01	
	0.03	0.72	−0.02	0.06	0.0	1.03	−0.01		0.097	0.0		0.1	0.0	
	0.19	0.27	0.59	0.49	0.01	−0.04	1.03			0.093			0.091	

Table 3
Multiple output runs with decomposition approach ($N_1 = 40$, $N_2 = 50$, $N_3 = 60$)

	A			T			C			Q			R	
Exact	0.9	0.0	−0.30	0.0	1.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	
	0.0	0.7	0.0	0.0	0.0	1.0	0.0		0.1	0.0		0.1	0.0	
	0.2	0.3	0.6	0.5	0.0	0.0	1.0			0.1			0.1	
Initial	0.5	0.0	0.0	0.5	0.5	0.0	0.0	0.05	0.0	0.0	0.05	0.0	0.0	
	0.0	0.5	0.0	0.5	0.0	0.5	0.0		0.05	0.0		0.05	0.0	
	0.0	0.0	0.5	0.5	0.0	0.0	0.5			0.05			0.05	
Iteration 1	0.65	−0.11	−0.10	−0.27	1.93	0.11	−0.28	0.31	0.11	0.09	0.26	0.11	0.12	
	0.21	0.55	0.15	0.23	0.24	1.65	−0.29		0.25	0.22		0.24	0.12	
	0.10	0.13	0.39	0.31	0.19	−0.13	1.23			0.21			0.18	
Iteration 2	0.74	−0.08	−0.15	−0.18	1.46	−0.05	−0.15	0.22	0.05	−0.04	0.18	0.07	0.08	
	0.16	0.60	0.08	0.17	0.15	1.35	−0.14		0.18	0.15		0.18	0.09	
	0.13	0.17	0.47	0.40	0.12	−0.06	1.18			0.15			0.17	
Iteration 3	0.80	−0.06	−0.20	−0.12	1.19	−0.02	0.06	0.16	0.01	−0.03	0.14	0.04	0.05	
	0.09	0.63	0.04	0.12	0.08	1.12	−0.08		0.14	0.09		0.16	0.07	
	0.16	0.22	0.53	0.45	0.06	−0.05	1.15			0.12			0.14	
Iteration 4	0.84	−0.04	−0.28	−0.08	1.15	0.01	0.02	0.13	−0.02	−0.02	0.11	0.02	0.04	
	0.05	0.65	−0.03	0.09	0.04	1.08	−0.04		0.11	0.05		0.13	0.06	
	0.18	0.25	0.56	0.47	0.03	−0.04	1.09			0.11			0.12	

7. Summary and discussion

A unified technique, based on the EM algorithm, for the parameter estimation of stationary (single output run) and non-stationary (multiple output runs) linear stochastic systems has been described in detail in this paper. This technique takes into account the cases where the system state x is either observable or unobservable. In order to reduce the computational load for parameter estimations of a high-dimensional dynamic system, a decomposition algorithm is developed which uses a hierarchical structure to perform successive orthogonalizations on the measurement subspaces of each subsystem. Since only low-order subsystem equations are manipulated at each stage, numerical inaccuracies are reduced and the subfilters decomposed from the global Kalman filter have better stability properties. With the inherent advantages of multiple processor computers in computation speed and in program modularity, the decomposition algorithm for parameter estimation described in this paper will show greater benefits when the algorithm is implemented in a multiple-processor platform. Our simulation results show that the decomposition approach performs almost as well as the global Kalman filtering approach while gaining substantial savings in computation time in executing state estimation.

Our initial motivation for this study was the desire to establish a statistical model for the speech signal that is capable of naturally and parametrically describing the correlation structure of the speech, and the decomposition algorithm has been developed for an intended use in the speech modeling context because of the high dimensionality required for the hidden state x underlying speech generation mechanisms. The dynamic system model as described in this paper, when viewed as a continuous-state HMM, appears to be superior to the discrete-state HMM for speech modeling. The discrete-state HMM, as originally formulated by Baum [1], is defined on an unobservable (or hidden) discrete-state Markov chain. When the HMM is applied as a model for speech, the real intention has been to account for the statistical properties of the surface acoustic signal by such a generative, 'hidden' stochastic process as the motion of the unobservable articulatory structure [20]. A natural consequence of this articulatory interpretation of the hidden states in the HMM is a proposal to

extend the discrete-state formulation of the HMM to its continuous-state counterpart, because the articulatory dynamics, assuming to be Markovian, is continuously valued.

In the recent past, a great deal of research has been devoted to improving the output distributions of the discrete-state HMM [16, 14, 18, 5–8]. That is, mainly the surface (acoustics) form of the speech signal has been focused on. This modeling approach leaves the increased sizes of the HMM state space and of the parameter set as the only possibility for increased modeling accuracy. Even in the recent limited research effort devoted to enhancing the capability of the underlying Markov chain for coarticulatory modeling [19, 9, 10], the discrete nature of the Markov state space remains unchanged, rendering inevitably insufficient precision for representing the articulatory dynamics that is intended to be described by the underlying hidden Markov process. The only continuous-state version of the Markov model proposed for use as a speech model is the stationary linear stochastic system described in [11]. The dynamic system model, as a mathematical abstraction, and the associated parameter estimation technique have been documented in the system theory and time series literatures for a long time [17, 24, 3, 15, 23]. The general technique described in this paper, including the efficient and accurate decomposition algorithm, for parameter estimation of the dynamic system (or the continuous-state HMM) has extended these previous works and, unlike these previous works, has presented details of simulation results demonstrating the effectiveness of the various versions of the estimation algorithm. Apparently, the ultimate success of the dynamic system model and the associated parameter estimation techniques developed in this paper in speech processing applications will require a great deal of future research. In particular, how to construct the continuous state x in the dynamic system model as an abstract representation of the articulatory dynamics in speech production and how to control the articulatory dynamics within the dynamic system framework will be the focus of future research.

Acknowledgements

We wish to acknowledge valuable discussions with V. Digalakis and M. Ostendorf, whose earlier work motivated this study. This work was supported by the NSERC of Canada.

References

[1] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, Vol. 3, 1972, pp. 1–8.

[2] R.A. Boyles, "On the convergence of the *EM* algorithm", *J. Roy. Statist. Soc. B*, Vol. 45, 1983, pp. 47–50.

[3] P.E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

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

[5] L. Deng, P. Kenny, M. Lennig, V. Gupta, F. Seitz and P. Mermelstein, "Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 39, 1991, pp 1677–1681.

[6] L. Deng, P. Kenny, M. Lennig and P. Mermelstein, "Modeling acoustic transitions in speech by state-interpolation hidden Markov models", *IEEE Trans. Signal Process.*, Vol. 40, 1992, pp. 265–272.

[7] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal", *Signal Processing*, Vol. 27, 1992, pp. 65–78.

[8] L. Deng, "A stochastic model of speech incorporating hierarchical nonstationarity", *IEEE Trans. Speech Audio Process.*, Vol. 1, 1993, pp. 471–474.

[9] L. Deng and K. Erler, "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units", *J. Acoust. Soc. Amer.*, Vol. 92, 1992, pp. 3058–3067.

[10] L. Deng and D. 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, May 1994, pp. 2702–2719.

[11] 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 Process.*, Vol. 1, 1993, pp. 431–442.

[12] M. Hassan, G. Salut, M.G. Singh and A. Titli, "A decentralized computational algorithm for the global Kalman filter", *IEEE Trans. Automat. Control*, Vol. AC-23, 1978, pp. 262–268.
[13] M. Hodzic and D.D. Siljak, "Iterative methods for parallel-multirate estimation", *Proc. IFAC Large Scale Systems: Theory and Applications*, Zurich, Switzerland, 1986.
[14] X. Huang and M. Jack, "Semi-continuous hidden Markov models for speech signal", *Comput. Speech Language*, Vol. 3, 1989, pp. 239–251.
[15] A.J. Isaksson, "Identification of ARX-models subject to missing data", *IEEE Trans. Automat. Control*, Vol. AC-38, 1993, pp. 813–819.
[16] B. Juang, S. Levinson and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chain", *IEEE Trans. Inform. Theory*, Vol. IT-32, 1986, pp. 307–309.
[17] R.L. Kashyap, "Maximum likelihood identification of stochastic linear systems", *IEEE Trans. Automat. Control*, Vol. AC-15, 1970, pp. 25–34.
[18] P. Kenny, M. Lennig and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 38, 1990, pp. 220–225.
[19] P. Kenny, R. Zhao, V. Gupta, M. Lennig and D. O'Shaughnessy, "Articulatory Markov models", *Proc. 1991 IEEE Workshop on Automatic Speech Recognition*, Arden House, Harriman, New York, 1991.
[20] S.E. Levinson, L.R. Rabiner and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", *Bell System Tech. J.*, Vol. 62, 1983, pp. 1035–1074.
[21] L. Ljung, *System Identification – Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[22] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, 1989, pp. 257–286.
[23] F.C. Schweppe, *Uncertain Dynamic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[24] R.H. Shumway and D.S. Stoffer, "An approach to time series smoothing and forecasting using the *EM* algorithm", *J. Time Series Anal.*, Vol. 3, 1982, pp. 253–264.
[25] D.A. Wilson and A. Kumar, "Derivative computations for the log likelihood function", *IEEE Trans. Automat. Control*, Vol. 27, 1982, pp. 230–232.
[26] C.F.J. Wu, "On the convergence properties of the *EM* algorithm", *Ann. Statist.*, Vol. 11, 1983, pp. 95–103.