ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language xxx (2003) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A mixed-level switching dynamic system for continuous speech recognition ☆

Jeff Ma [a], Li Deng [b,*]

[a] BBN Technologies, Cambridge, MA 02138, USA
[b] Microsoft Research, One Microsoft Way, Redmond WA 98052, USA

Received 21 July 2000; received in revised form 29 May 2003; accepted 11 June 2003

Abstract

A two-level mixture linear dynamic system model, with frame-level switching parameters in the observation equation and with segment-level switching parameters in the target-directed state equation, is developed and evaluated. The main contributions of this work are: (1) the new framework for dealing with mixed-level switching in the dynamic system and (2) the novel use of piecewise linear functions, enabled by the introduction of frame-level switching, to approximate the nonlinear function between the hidden vocal-tract-resonance space and the observable acoustic space. The approximation is accomplished by the frame-dependent switching parameters in the observation equation. In this paper, in a self-contained manner, we highlight the key algorithm differences from the earlier model having only single segment-level switching that is synchronous between the state and observation equations. A series of speech recognition experiments are carried out to evaluate this new model using a subset of Switchboard conversational speech data. The experimental results show that the approximation accuracy is improved with an increased number of switching-parameter values. The speech recognizer built from the new mixed-level switching dynamic system model using an N-best re-scoring evaluation paradigm show moderate word error rate reduction compared with using either single-level switching or no switching parameters.

☆ The main portion of the work was done while the authors were with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont., Canada N2L 3G1.

* Corresponding author. Tel.: +1-425-706-2719; fax: +1-425-936-7329.
E-mail address: deng@microsoft.com (L. Deng).

0885-2308/03/$ - see front matter
doi:10.1016/S0885-2308(03)00031-7

YCSLA 232
DISK / 7/7/03
ARTICLE IN PRESS
No. of pages: 17
DTD 4.3.1/SPS

1. Introduction

In recent years, a new approach to the challenging problem of conversational speech recognition has emerged, holding a promise to overcome some fundamental limitations of the conventional Hidden Markov Model (HMM) approach (e.g., Bridle et al., 1998; Deng, 1999; Deng and Ma, 2000; Ma and Deng, 2003; Picone et al., 1999). This new approach is a radical departure from the current HMM-based statistical modeling approaches. Rather than using a large number of unstructured Gaussian mixture components to account for the tremendous variation in the observable acoustic data of highly coarticulated spontaneous speech, the new speech model that we have developed provides a rich structure for the partially observed (hidden) dynamics in the domain of vocal-tract-resonances (VTRs) (Deng and Ma, 1999, 2000). In the design of the speech recognizer reported in Deng and Ma (1999, 2000), we use a statistical nonlinear dynamic system to describe the physical process of spontaneous speech production where knowledge of the VTR dynamic behavior in speech production is naturally incorporated into the model training and decoding.

In the previous work documented in Deng (1999), Deng and Ma (1999, 2000) and Ma and Deng (2003), a long-span coarticulatory model, one for each phone segment, was formulated in mathematical terms as a constrained state-space nonlinear system. The state equation represents a stochastic linear system, where the state variable represents the VTR hidden dynamics. The target-directed or asymptotic behavior of the dynamics is established by forcing the system to enter the asymptotic regime after large time steps. The measurement equation in the state-space model is a static nonlinear mapping from the hidden dynamic (VTR) space to the observable acoustic space (e.g., MFCC).

In our earlier work (Deng and Ma, 1999, 2000), due to the use of the nonlinear mapping function, $h(Z)$, two approximations had to be made in the model parameter learning process. The first approximation was $E[h(Z)] \approx h(E[Z])$. This amounts to making Taylor series expansion on $h(E[Z(k)])$ and then truncating all high-order terms above the linear term. This overcomes the difficulty in the calculation of the expectation of the nonlinear function, at the expense of an unknown degree of reduced accuracy. The second approximation arises from the use of the extended Kalman filter (EKF) (Kitagawa, 1987; Mendel, 1995; Tanizaki, 1996) (due to the presence of the nonlinear function), which was known to be non-optimal. In order to minimize the loss of computational accuracy in parameter learning and likelihood calculation due to these approximations (but at the expense of a possible loss of modeling accuracy), the work reported in Ma and Deng (2003) developed a mixture linear dynamic system model, where several (mixture) linear regression mapping functions in the measurement equation were used to approximate the single nonlinear mapping function. The switching between one mixture component to another occurred at the phone-segment level [in a manner analogous to the segmental models described in Gish and Ng (1993), Ostendorf et al. (1996)]; that is, the same mixture component was sustained across the entire segment and the new mixture component may be switched to only at the new segment boundary. Also, the same segmental constraint was applied to the switching of the parameters (targets and time constants) in the state equation. Since the parameter switchings in the state equation and in the measurement equation are synchronous, we call that model (Ma and Deng, 2003) as the single-level switching dynamic system.

In this paper, we introduce the mixed-level switching dynamic system where the parameter switchings in the state equation and in the measurement equation are not synchronous. The switching in the state equation remains at the segmental level, but that in the measurement equation becomes instead at the frame level. Making the switching of mixture (i.e., multiple) components in linear mapping functions of the measurement equation at the frame level has advantages over the previous segment-level switching described in Ma and Deng (2003) as follows. For the frame-level switching, the multiple linear mappings become a piecewise linear approximation, at each frame, to the nonlinear function that defines the true mapping between the VTR and observation spaces. At different frames, a different "piece" of the linear approximation function may be optimally selected. This greatly increases the flexibility of the model in approximating the original complex nonlinear function. In contrast, for the segment-level switching as developed in Ma and Deng (2003), only one linear approximation is made to the nonlinear function for a given segment (consisting of many frames), although there is an inventory (mixture) of possible linear functions available for each segment.

The organization of this paper is as follows. In Section 2, a description of the mixed-level switching dynamic system model is provided, including its complete parameterization. In Section 3, an Expectation-Maximization-based algorithm is presented for parameter estimation of this model. Some technical details will be referred to the earlier model of Ma and Deng (2003) with single-level switching, which shares some similar steps of the algorithm derivation, and key algorithm changes will be summarized. In Section 4, we report speech recognition experiments, aimed to evaluate the new mixed-level switching model on the Switchboard database under the N-best list re-scoring paradigm.

2. Model formulation

The mixed-level, switching dynamic system model developed in this study is a linear combination of standard linear dynamic models (a total of M). The mth linear dynamic model with the first-level, segmental switching has the following form:

$$Z(k) = \Phi_m Z(k-1) + (I - \Phi_m)T_m + W_m(k-1), \tag{1}$$

$$O(k) = \dot{H}_m^{(l)}(k)\dot{Z}(k) + V_m(k), \tag{2}$$

where $\dot{H}_m^{(l)} = [a_m^{(l)}, H_m^{(l)}]$ and $\dot{Z}(k) = [1, Z(k)']'$. The state equation, Eq. (1), for each phone segment is associated with M sets of parameters (Φ_m, T_m, Q_m), but they are not switched from one set to another until at the end of the current segment (hence segment-level switching). That is, the parameter switching takes place at the boundary between two adjacent phones. In the measurement equation, Eq. (2), however, the parameters $\dot{H}_m^{(l)}(k)$ are allowed to switch at each time frame k. This gives rise to the second-level switching at the frame level. This frame-level switching is constrained to take values from fixed finite sets: $\{\dot{H}_m^{l'}, l' = 1, 2, \ldots, L\}$ for each frame. Note that the subscript m in Eq. (2) indicates that different segment-level mixture components correspond to different sets of $\dot{H}^{(l)}$ values. Also note that when $L = 1$, the above model is reduced to the single-level switching model described in Ma and Deng (2003).

To summarize, the complete set of model parameters consists of

$$\Theta = \{\pi_m, \Phi_m, T_m, Q_m, R_m, \dot{H}_m^l, \gamma_{m,l},\ m = 1, 2, \ldots, M,\ l = 1, 2, \ldots, L\},$$

where π_m represents the mixture weight probability $P(m|\Theta)$ and $\gamma_{m,l}$ represents the $\dot{H}_m^l$-value weight probability $P(l|m,\Theta)$. As discussed before, two levels of parameter switching have been designed. First, the mixture component indexed by m switches at the segment (phone) level. Second, the $\dot{H}_m^l$ value indexed by l switches at the frame level. The first level of switching corresponds to the target property of the VTR dynamics, which therefore is reasonably placed at the segment level. The second level of parameter switching is designed to provide the flexibility for using multiple linear regression functions to approximate the nonlinear relationship between the VTR and the measurement variables. It is desirable that the parameter switching happens at the frame level because the relationship approximated by the linear mapping can change at different frames. In this paper, we also call the former switching *mixture switching* and the later one *H-value switching*. As will be seen in Section 5, use of the additional H-value switching gives improved speech recognition performance over the model of Ma and Deng (2003) which has the mixture switching only.

In order to implement the mixture switching, it is necessary to impose the *mixture-path constraint*, where for each sequence of the acoustic observation associated with a phone, the observation is restricted to be produced from a fixed mixture component, m, of the model. This means that the target of the VTR in a phone is not permitted to switch from one mixture component to another at the frame level. [1] The constraint of such a type is motivated by the physical nature of the speech model – the target that is correlated with the phonetic identity is defined at the segment (phone) level, not at the frame level. This constraint is imposed both the model training and on the model scoring in implementing the speech recognizer. For the H-value switching, no such constraints are in place.

3. Learning model parameters

Due to the unobserved nature of the state in the model presented in the above section, Expectation-Maximization (EM) algorithm has been developed for model parameter estimation (Dempster et al., 1977; Deng, 1993; Digalakis et al., 1993; Ostendorf et al., 1996). The approach we are taking here has been inspired by that in Streit and Luginbuhl (1998), with substantial modifications to suit our specific mixed-level, switching dynamic model of speech.

Before formally describing the EM algorithm, we first define a discrete random variable X, which provides the observation-to-mixture assignment for a sequence of observations. For example, for a given sequence of observations of a phone, when $X = m,\ (1 \leqslant m \leqslant M)$, it means that the mth mixture component of the model is responsible for generating that observation sequence. We need one additional discrete variable to represent the H-value switching on the measurement equation. This is denoted by $Y = \{y_1, y_2, \ldots, y_K\}$ (K is the length of the observation), where $y_k (1 \leqslant k \leqslant K)$ is a discrete random variable indicating which one of the $\dot{H}_m^l s (1 \leqslant l \leqslant L)$ is switched onto at time frame k. For example, when $y_k = i$, it means the ith value, $\dot{H}_m^i$, is chosen at time k. (We assume that $y_1, y_2, \ldots, y_K$ are independent random variables.) Finally, we define a discrete-variable pair, $S = \{X, Y\}$, which represents the combination of X and Y. Note that in the single-level switching model of Ma and Deng (2003), only one discrete random variable X was introduced.

[1] This same mixture-path constraint has been imposed earlier on the mixture-trended HMM; see Deng and Aksmanovic (1997).

In order to impose the mixture-path constraint discussed earlier, the joint variable pair is defined at the segment level. Suppose we have N training tokens for a phone. We in this case define the joint variable as

$$\{O,Z,S\}^N = \{(O^1,Z^1,S^1),(O^2,Z^2,S^2),\ldots,(O^n,Z^n,S^n),\ldots,(O^N,Z^N,S^N)\},$$

where $O^n = \{O^n(1),O^n(2),\ldots,O^n(K_n)\}$ is the nth observation, Z^n is the corresponding hidden state sequence, and $S^n = \{X^n,Y^n\}$ denotes the corresponding observation-to-mixture assignment by X^n as well as the switching behavior of the measurement equation represented by

$$Y^n = \{y_1^n,y_2^n,\ldots,y_{K_n}^n\}.$$

Here, X^n and Y^n are combined to jointly determine how the observation sequences are generated.

The following assumptions are made in the development of the learning algorithm:

- The N tokens are independent of each other. That is, S^n $(1 \leqslant n \leqslant N)$ are independent of each other.
- The random variables X^n, $(1 \leqslant n \leqslant N)$ have identical (discrete) distributions.
- $y_1^n,y_2^n,\ldots,$ and $y_{K_n}^n$ are independent of each other. That is, the parameter switching does not depend on the history, nor on the future.
- y_k^n $(1 \leqslant k \leqslant K_n)$ have identical (discrete) distributions.

With these assumptions, the PDF of the joint variable $\{O,Z,S\}^N$, with the fixed model parameter set Θ, can be decomposed into

$$\begin{aligned} &p(\{O,Z,S\}^N|\Theta) \\ &= \prod_{n=1}^{N}\left\{p(Z_0^n|\Theta)\left[\prod_{k=1}^{K_n}p(Z_k^n|Z_{k-1}^n,X^n,\Theta)p(O_k^n|Z_k^n,X^n,y_k^n,\Theta)P(y_k^n|X^n,\Theta)\right]P(X^n|\Theta)\right\}. \end{aligned} \quad (3)$$

This result is generalization of a corresponding result for the single-level switching model presented and derived in Ma and Deng (2003).

Further, to compute the auxiliary Q-function in the E-step of the EM algorithm, we need several other PDFs.

First, we have

$$p(\{O,S\}^N|\Theta) = \prod_{n=1}^{N}\left[\prod_{k=1}^{K_n}p(O^n(k)|O_{1,k-1}^n,X^n,y_k^n,\Theta)P(y_k^n|X^n,\Theta)\right]P(X^n|\Theta). \quad (4)$$

Second, we decompose $p(\{O,X\}^N|\Theta)$ into

$$\begin{aligned} p(\{O,X\}^N|\Theta) &= \sum_{\{Y\}^N}p(\{O,S\}^N|\Theta) \\ &= \sum_{Y^1}\sum_{Y^2}\cdots\sum_{Y^N}\prod_{n=1}^{N}\left[\prod_{k=1}^{K_n}p(O^n(k)|O_{1,k-1}^n,X^n,y_k^n,\Theta)P(y_k^n|X^n,\Theta)\right]P(X^n|\Theta) \\ &= \prod_{n=1}^{N}\left[\prod_{k=1}^{K_n}\sum_{y_k^n=1}^{L}p(O^n(k)|O_{1,k-1}^n,X^n,y_k^n,\Theta)P(y_k^n|X^n,\Theta)\right]P(X^n|\Theta), \end{aligned} \quad (5)$$

where independence between tokens and between switchings at different times has been used.

Third, $p(\{O\}^N|\Theta)$ is computed as

$$p(\{O\}^N|\Theta) = \sum_{\{X\}^N} p(\{O,X\}^N|\Theta) = \prod_{n=1}^{N} \sum_{X^n=1}^{M} \left[\prod_{k=1}^{K_n} \sum_{y_k^n=1}^{L} p(O^n(k)|O^n_{1,k-1}, X^n, y_k^n, \Theta) P(y_k^n|X^n, \Theta) \right] P(X^n|\Theta). \tag{6}$$

Then, the conditional PDF $p(\{X\}^N)|\{O\}^N, \Theta)$ can be derived to be

$$p(\{X\}^N)|\{O\}^N, \Theta) = \frac{p(\{O,X\}^N|\Theta)}{p(\{O\}^N|\Theta)} = \prod_{n=1}^{N} \omega_m^n, \tag{7}$$

where

$$\omega_m^n = \frac{\left[\prod_{k=1}^{K_n} \sum_{l=1}^{L} p(O^n(k)|O^n_{1,k-1}, m, l, \Theta) P(l|m, \Theta) \right] P(m|\Theta)}{\sum_{m=1}^{M} \left[\prod_{k=1}^{K_n} \sum_{l=1}^{L} p(O^n(k)|O^n_{1,k-1}, m, l, \Theta) P(l|m, \Theta) \right] P(m|\Theta)}. \tag{8}$$

In the above, because X^ns have identical distributions, they are replaced by a common variable m for notational simplicity. For the same reason, y_k^n is replaced by the common variable l.

Finally, we compute the conditional PDF of

$$p(\{Y\}^N|\{X\}^N, \{O\}^N, \Theta) = \frac{p(\{O,S\}^N|\Theta)}{p(\{O,X\}^N|\Theta)} = \prod_{n=1}^{N} \prod_{k=1}^{K_n} \xi^n_{k,m,l}, \tag{9}$$

where

$$\xi^n_{k,m,l} = \frac{p(O^n(k)|O^n_{1,k-1}, m, l, \Theta) P(l|m, \Theta)}{\sum_{l=1}^{L} p(O^n(k)|O^n_{1,k-1}, m, l, \Theta) P(l|m, \Theta)}. \tag{10}$$

Using the independence assumption among tokens, we obtain

$$p(Y^n|X^n, O^n, \Theta) = \prod_{k=1}^{K_n} \xi^n_{k,m,l}. \tag{11}$$

The independence assumption among switchings at different times further gives

$$p(y_k^n|X^n, O^n, \Theta) = \xi^n_{k,m,l}. \tag{12}$$

3.1. EM algorithm: E-step

Given all the joint PDF and conditional PDF computations discussed above, we describe the EM algorithm. Since both $\{Z\}^N$ and $\{S\}^N$ are missing data, we compute integration over both of them to obtain the Q-function:

$$Q(\Theta|\bar{\Theta}) = \sum_{\{S\}^N} \int \log p(\{O,Z,S\}^N|\Theta) \cdot p(\{Z\}^N|\{O,S\}^N, \bar{\Theta}) \, \mathrm{d}\{Z\}^N p(\{S\}^N|\{O\}^N, \bar{\Theta}), \tag{13}$$

where $\bar{\Theta}$ denotes the parameter set at the immediately previous step of the EM algorithm.

Generalizing some derivation steps in Ma and Deng (2003), we can show that the auxiliary function consists of three parts:

$$Q(\Theta|\bar{\Theta}) = Q_Z + Q_Y + Q_X. \tag{14}$$

Also, following the same steps as in Ma and Deng (2003), these three terms can be simplified to (as before, we use m and l to denote X^n and y_k^n, respectively):

$$\begin{aligned} Q_Z = & \sum_{n=1}^{N}\sum_{m=1}^{M}\int\left[\sum_{k=1}^{K_n}\log p(Z_k^n|Z_{k-1}^n, m, \Theta)\right]p(Z^n|O^n, m, \bar{\Theta})\,\mathrm{d}Z^n\cdot\bar{\omega}_m^n \\ & + \sum_{n=1}^{N}\sum_{m=1}^{M}\int\left\{\sum_{k=1}^{K_n}\sum_{l=1}^{L}\log p(O_k^n|Z_k^n, m, l, \Theta)\cdot\bar{\xi}_{k,m,l}^n\cdot p(Z^n|O^n, m, l, \bar{\Theta})\right\}\mathrm{d}Z^n\cdot\bar{\omega}_m^n, \end{aligned} \tag{15}$$

$$Q_Y = \sum_{n=1}^{N}\sum_{m=1}^{M}\left[\sum_{k=1}^{K_n}\sum_{l=1}^{L}\log P(l|m, \Theta)\cdot\bar{\xi}_{k,m,l}^n\right]\bar{\omega}_m^n, \tag{16}$$

$$Q_X = \sum_{n=1}^{N}\sum_{m=1}^{M}\log P(m|\Theta)\bar{\omega}_m^n. \tag{17}$$

In the above the symbols $\bar{\omega}_m^n$ and $\bar{\xi}_{k,m,l}^n$ denote the corresponding variables of ω_m^n and $\xi_{k,m,l}^n$ (Eqs. (8) and (10), respectively) for the preceding EM iteration.

By the definition of the model in Eqs. (1) and (2), $p(Z_k^n|Z_{k-1}^n, m, \Theta)$ is a Gaussian with mean: $\Phi_m Z^n(k-1) + (I - \Phi_m)T_m$ and with covariance: Q_m. And $p(O_k^n|Z_k^n, m, l, \Theta)$ is also a Gaussian with mean: $\dot{H}_m^l \dot{Z}^n(k)$ and with covariance: R_m. Therefore, Q_Z can be re-written as

$$\begin{aligned} Q_Z = & -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\{K_n\log|Q_m| + \sum_{k=1}^{K_n}E_m\left[e1_{k,m}^{n}{}'(Q_m)^{-1}e1_{k,m}^n\right]\right\}\cdot\bar{\omega}_m^n \\ & -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\{K_n\log|R_m| + \sum_{k=1}^{K_n}\sum_{l=1}^{L}E_{ml}\left[e2_{k,m,l}^{n}{}'(R_m)^{-1}e2_{k,m,l}^n\cdot\bar{\xi}_{k,m,l}^n\right]\right\}\cdot\bar{\omega}_m^n + \text{const.}, \end{aligned} \tag{18}$$

where $e1_{k,m}^n$ and $e2_{k,m,l}^n$ are defined as

$$e1_{k,m}^n = Z^n(k) - \Phi_m Z^n(k-1) - (I - \Phi_m)T_m,$$

$$e2_{k,m,l}^n = O^n(k) - \dot{H}_m^l \dot{Z}^n(k).$$

In Eq. (18), $E_m[\cdot]$ denotes the conditional expectation $E[\cdot|O^n, m, \bar{\Theta})]$ and $E_{ml}[\cdot]$ denotes the conditional expectation $E[\cdot|O^n, m, l, \bar{\Theta})]$. These conditional expectations, $E_m[\cdot]$ and $E_{ml}[\cdot]$, are computed from the Kalman smoothing algorithm that will be discussed in detail later.

3.2. EM algorithm: M-step

With the Q-function computed above, we now go to the M-step of the EM-algorithm.

Estimate for π_m. The final form of the reestimate for π_m is

$$\hat{\pi}_m = \frac{\sum_{n=1}^{N} \bar{\omega}_m^n}{\sum_{n=1}^{N} \sum_{m=1}^{M} \bar{\omega}_m^n} \quad \text{for } 1 \leqslant m \leqslant M. \tag{19}$$

This is identical to the estimate for the single-level switching model of Ma and Deng (2003), where a derivation was provided.

Estimate for $\gamma_{m,l}$. Note in the Q-function, only Q_Y is related to the "switching" probability $\gamma_{m,l} = P(l|m, \Theta)$. With the constraint, $\sum_{l=1}^{L} \gamma_{m,l} = 1$, the Lagrangian is

$$L_Y = Q_Y + \lambda(1 - \sum_{l}^{L} \gamma_{m,l}).$$

Taking the derivative of L_Y with respect to $\gamma_{m,l}$ and set it to zero, we obtain

$$\frac{\partial L_Y}{\partial \gamma_{m,l}} = \sum_{n=1}^{N} \left[\sum_{k=1}^{K_n} \frac{\bar{\xi}_{k,m,l}^n}{\gamma_{m,l}} \right] \bar{\omega}_m^n - \lambda.$$

Solving the above for $\gamma_{m,l}$ and normalizing with the Lagrangian finally give the final form of the re-estimation formula

$$\hat{\gamma}_{m,l} = \frac{\sum_{n=1}^{N} [\sum_{k=1}^{K_n} \bar{\xi}_{k,m,l}^n] \bar{\omega}_m^n}{\sum_{n=1}^{N} K_n \bar{\omega}_m^n}. \tag{20}$$

Note that the single-level switching model of Ma and Deng (2003) did not have the parameter $\gamma_{m,l}$ to estimate.

Estimates for Φ_m, T_m *and* Q_m. Note that in the Q-function only the first term of Q_Z is related to these parameters in the state equation. We first introduce the following notations:

$$A0_m^n = \sum_{k=1}^{K_n} E_m[Z^n(k-1)Z^n(k-1)'], \quad A1_m^n = \sum_{k=1}^{K_n} E_m[Z^n(k)Z^n(k)'],$$

$$A2_m^n = \sum_{k=1}^{K_n} E_m[Z^n(k)Z^n(k-1)'], \quad C_m = (I - \hat{\Phi}_m)\hat{T}_m,$$

$$B0_m^n = \sum_{k=1}^{K_n} E_m[Z^n(k-1)], \quad B1_m^n = \sum_{k=1}^{K_n} E_m[Z^n(k)].$$

Then the final estimation formulas for these parameters are

$$\hat{T}_m = \frac{(I - \Phi_m)^{-1} \sum_{n=1}^{N} \{B1_m^n - \Phi_m B0_m^n\} \cdot \bar{\omega}_m^n}{\sum_{n=1}^{N} K_n \cdot \bar{\omega}_m^n}, \tag{21}$$

$$\hat{\Phi}_m = \left\{ \sum_{n=1}^{N} (A2_m^n - B1_m^n \hat{T}_m' - \hat{T}_m B0_m^{n\prime} + K_n \hat{T}_m \hat{T}_m') \cdot \bar{\omega}_m^n \right\} \cdot \left\{ \sum_{n=1}^{N} (A0_m^n - B0_m^n \hat{T}_m' - \hat{T}_m B0_m^{n\prime} + K_n \hat{T}_m \hat{T}_m') \cdot \bar{\omega}_m^n \right\}^{-1}, \tag{22}$$

$$\hat{Q}_m = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K_n} E_m[e1^n_{k,m} e1^{n\,\prime}_{k,m}] \cdot \bar{\omega}^n_m}{\sum_{n=1}^{N} K_n \bar{\omega}^n_m}. \tag{23}$$

Again, these results are identical to the estimates for the single-level switching model of Ma and Deng (2003), since these estimated parameters are in the state equation and do not subject to the frame-level switching.

Estimates for $\dot{H}^l_m$ and R_m. Finally, we derive the re-estimation formulas for the parameters, $\dot{H}_{m,l}$ and R_m, contained in the measurement equation. In the Q-function, only the second term in Q_Z is related to these parameters. Due to the frame-level switching involved in the measurement equation, the results presented below are different from the corresponding estimation formulas described in Ma and Deng (2003).

The derivative of Q_Z with respect to $\dot{H}^l_m$ is

$$\frac{\partial Q_Z}{\partial \dot{H}^l_m} = -R^{-1}_m \left\{ \sum_{n=1}^{N} \left(\sum_{k=1}^{K_n} E_{ml}[(\dot{H}^l_m \dot{Z}^n(k) - O^n_k)(\dot{Z}^n(k))'] \bar{\xi}^n_{k,m,l} \right) \bar{\omega}^n_m \right\}. \tag{24}$$

Setting the derivative to zero, we obtain

$$\widehat{\dot{H}^l_m} = \left\{ \sum_{n=1}^{N} \bar{\omega}^n_m \sum_{k=1}^{K_n} O^n(k) E_{ml}[\dot{Z}^n(k)]' \cdot \bar{\xi}^n_{k,m,l} \right\} \left\{ \sum_{n=1}^{N} \bar{\omega}^n_m \sum_{k=1}^{K_n} E_{ml}[\dot{Z}^n(k)(\dot{Z}^n(k))'] \cdot \bar{\xi}^n_{k,m,l} \right\}^{-1}, \tag{25}$$

where $E_{ml}[\dot{Z}^n(k)] = [1, E_{ml}[Z^n(k)]']'$ and

$$E_{ml}[\dot{Z}^n(k)(\dot{Z}^n(k))'] = \begin{bmatrix} 1 & E_{ml}[Z^n(k)]' \\ E_{ml}[Z^n(k)] & E_{ml}[Z^n(k)(Z^n(k))'] \end{bmatrix}.$$

The derivative of Q_Z with respect to R^{-1}_m is

$$\frac{\partial Q_Z}{\partial R^{-1}_m} = \frac{1}{2} \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K_n} \sum_{l=1}^{L} \left(R_m - E_{ml}[e2^n_{k,m,l}(e2^n_{k,m,l})'] \right) \cdot \bar{\xi}^n_{k,m,l} \right\} \cdot \bar{\omega}^n_m. \tag{26}$$

Setting this to zero, we obtain the estimate for R_m:

$$\hat{R}_m = \frac{\sum_{n=1}^{N} \left(\sum_{k=1}^{K_n} \sum_{l=1}^{L} E_{ml}[e2^n_{k,m,l}(e2^n_{k,m,l})'] \cdot \bar{\xi}^n_{k,m,l} \right) \cdot \bar{\omega}^n_m}{\sum_{n=1}^{N} K_n \bar{\omega}^n_m}, \tag{27}$$

where $E_{ml}[e2^n_{k,m,l}(e2^n_{k,m,l})']$ is calculated according to

$$\begin{aligned} E_{ml}[e2^n_{k,m,l}(e2^n_{k,m,l})'] = {} & O^n(k)O^n(k)' - E_{ml}[\dot{Z}^n(k)](\widehat{\dot{H}^l_m})' \\ & - \widehat{\dot{H}^l_m} E_{ml}[\dot{Z}^n(k)] + \widehat{\dot{H}^l_m} E_{ml}[\dot{Z}^n(k)\dot{Z}^n(k)'](\widehat{\dot{H}^l_m})'. \end{aligned} \tag{28}$$

3.3. *Sufficient statistics computed by Kalman smoother*

As we showed earlier, in order to obtain the re-estimates for the model parameters, a set of conditional expectations, which form the sufficient statistics for the estimation problem, need to be calculated during the M-step of the EM algorithm. These sufficient statistics include $E_m[Z^n(k)]$, $E_m[Z^n(k)Z^n(k)']$, $E_m[Z^n(k)Z^n(k-1)']$, $E_{ml}[Z^n(k)]$ and $E_{ml}[Z^n(k)Z^n(k)']$.

The conditional expectation $E_m[\cdot] = E[\cdot|O^n, m, \bar{\Theta}]$ is the Kalman smoother of the mth mixture component for the nth observation. However, the conventional Kalman smoother can not be directly applied here because the current model has parameter switching occurring in the measurement equation. This situation is similar to that presented in Shumway and Stoffer (1991), where the filtering and smoothing algorithms were derived. Based on the solution given in Shumway and Stoffer (1991), the conditional expectation $E_{ml}[\cdot] = E[\cdot|O^n, m, l, \bar{\Theta}]$ in our problem becomes the smoother of the mth mixture component for the nth observation under an extra condition of $\dot{H}_m(k) = \dot{H}_m^l$ (recall that we use l to represent $y_k^n = l$).

The basic theory of Kalman filtering and smoothing can be found in Tanizaki (1996), Kitagawa (1987), Mendel (1995), etc. In the following we list the filtering and smoothing algorithms for the new mixed-level switching model. These algorithms generalize the ones presented in Ma and Deng (2003), which became the special cases when the number (L) of frame-level or H-value switching possibilities is reduced to one.

Forward recursion (or filtering):

$$\hat{Z}^n_{k|k-1,m} = \Phi_m \hat{Z}^n_{k-1|k-1,m} + (I - \Phi_m) T_m, \tag{29}$$

$$\Sigma^n_{k|k-1,m} = \Phi_m \Sigma^n_{k-1|k-1,m} \Phi_m + Q_m, \tag{30}$$

$$\tilde{O}^n_{k,m,l} = O^n(k) - \dot{H}^l_m \dot{\hat{Z}}^n_{k|k-1,m}, \quad l = 1, 2, \ldots, L, \tag{31}$$

$$\Sigma^n_{\tilde{O}_{k,m,l}} = H^l_m \Sigma^n_{k|k-1,m} H^{l'}_m + R_m, \tag{32}$$

$$K_{k,m,l} = \Sigma^n_{k|k-1,m} H^{l'}_m \left(\Sigma^n_{\tilde{O}_{k,m,l}} \right)^{-1}, \tag{33}$$

$$\hat{Z}^n_{k|k,m,l} = \hat{Z}^n_{k|k-1,m} + K_{k,m,l} \tilde{O}^n_{k,m,l}, \tag{34}$$

$$\Sigma^n_{k|k,m,l} = \Sigma^n_{k|k-1,m} - K_{k,m,l} \Sigma^n_{\tilde{O}_{k,m,l}} K'_{k,m}, \tag{35}$$

$$\psi_{k,m,l} = \frac{\gamma_{m,l} \cdot \mathcal{N}\left(O^n(k); \dot{H}^l_m \dot{\hat{Z}}^n_{k|k-1,m}, \Sigma^n_{\tilde{O}_{k,m,l}} \right)}{\sum_{l=1}^{L} \gamma_{m,l} \cdot \mathcal{N}\left(O^n(k); \dot{H}^l_m \dot{\hat{Z}}^n_{k|k-1,m}, \Sigma^n_{\tilde{O}_{k,m,l}} \right)}, \tag{36}$$

$$\hat{Z}^n_{k|k,m} = \sum_{l=1}^{L} \psi_{k,m,l} \cdot \hat{Z}^n_{k|k,m,l}, \tag{37}$$

$$\Sigma^n_{k|k,m} = \sum_{l=1}^{L} \psi_{k,m,l} \cdot \Sigma^n_{k|k,m,l}, \tag{38}$$

where $\hat{Z}^n_{k|k-1,m}$ is the predictor and $\Sigma^n_{k|k-1,m}$ its error covariance. $\hat{Z}^n_{k|k,m}$ is the filter and $\Sigma^n_{k|k,m}$ its error covariance. $\mathcal{N}(O^n(k); \dot{H}^l_m \hat{Z}^n_{k|k-1,m}, \Sigma^n_{\tilde{O}_{k,m,l}})$ is a Gaussian density with mean $\dot{H}^l_m \hat{Z}^n_{k|k-1,m}$ and covariance $\Sigma^n_{\tilde{O}_{k,m,l}}$. This is the density of the innovation sequence at time k.

Backward recursion (or smoothing):

$$A^n_{k,m} = \Sigma^n_{k|k,m} \Phi'_m (\Sigma^n_{k|k-1,m})^{-1}, \tag{39}$$

$$\hat{Z}^n_{k|K_n,m} = \hat{Z}^n_{k|k,m} + A^n_{k,m}[\hat{Z}^n_{k+1|K_n} - \hat{Z}^n_{k+1|k,m}], \tag{40}$$

$$\Sigma^n_{k|K_n,m} = \Sigma^n_{k|k,m} + A^n_{k,m}[\Sigma^n_{k+1|K_n,m} - \Sigma^n_{k+1|k,m}]A'_k. \tag{41}$$

Note that the above smoothing is for the computation of $E[\cdot|O^n, m, \bar{\Theta}]$. For the computation of $E[\cdot|O^n, m, l, \bar{\Theta}]$ at the given time point k, $\hat{Z}^n_{k|k,m}$ and $\Sigma^n_{k|k,m}$ in the above smoothing algorithm are simply replaced by $\hat{Z}^n_{k|k,m,l}$ and $\Sigma^n_{k|k,m,l}$, respectively, and correspondingly, $\hat{Z}^n_{k|K_n,m}$ becomes $\hat{Z}^n_{k|K_n,m,l}$ and $\Sigma^n_{k|K_n,m}$ becomes $\Sigma^n_{k|K_n,m,l}$. At other points, smoothing remains unchanged.

Using the above Kalman smoothing results, the conditional expectations as sufficient statistics are computed by

$$E_m[Z^n(k)] = \hat{Z}^n_{k|K_n,m}, \tag{42}$$

$$E_m[Z^n(k)Z^n(k)'] = \Sigma^n_{k|K_n,m} + \hat{Z}^n_{k|K_n,m}(\hat{Z}^n_{k|K_n,m})', \tag{43}$$

$$E_m[Z^n(k)Z^n(k-1)'] = \Sigma^n_{k,k-1|K_n,m} + \hat{Z}^n_{k|K_n,m}(\hat{Z}^n_{k-1|K_n,m})', \tag{44}$$

$$E_{ml}[Z^n(k)] = \hat{Z}^n_{k|K_n,m,l}, \tag{45}$$

$$E_{ml}[Z^n(k)Z^n(k)'] = \Sigma^n_{k|K_n,m,l} + \hat{Z}^n_{k|K_n,m,l}(\hat{Z}^n_{k|K_n,m,l})', \tag{46}$$

where $\Sigma^n_{k,k-1|K_n,m}$ is recursively calculated by Shumway (1982)

$$\Sigma^n_{k,k-1|K_n,m} = \Sigma^n_{k|k,m} A^{n'}_{k-1,m} + A^n_{k,m}(\Sigma^n_{k+1,k|K_n,m} - \Phi_m \Sigma^n_{k|k,m}) A^{n'}_{k-1,m} \tag{47}$$

for $k = K_n, \ldots, 2$, where the end point is

$$\Sigma^n_{K_n,K_n-1|K_n,m} = \sum_{l=1}^{L} \psi_{K_n,m,l}(I - K_{K_n,m,l} H^l_m) \Phi_m \Sigma^n_{K_n-1|K_n-1,m,l}. \tag{48}$$

3.4. Computation of ω and ξ

To compute ω^n_m and $\xi^n_{k,m,l}$ according to Eqs. (8) and (10), respectively, it suffices to compute $p(O^n_k|O^n_{1,k-1}, m, l, \bar{\Theta})$. This is actually the PDF of the innovation sequence, and hence it is computed straightforwardly by

$$p(O^n_k|O^n_{1,k-1}, m, l, \bar{\Theta}) = (2\pi)^{-d/2} \left|\Sigma^n_{\tilde{O}_{k,m,l}}\right|^{-1/2} \exp\left\{ -\frac{1}{2}(\tilde{O}^n_{k,m,l})'[\Sigma^n_{\tilde{O}_{k,m,l}}]^{-1} \tilde{O}^n_{k,m,l} \right\}. \tag{49}$$

4. Likelihood computation

Efficient computation of the likelihood of a sequence of observation vectors using the model is a key requirement [2] for implementing the speech recognizer (in the recognizer testing phase). Such computation is discussed in this section.

We combine M different linear dynamic models (mixture models) using different weights to represent a phone's VTR dynamics. After the weights and all model parameters are trained as described in the preceding section, the likelihood of a phone for a sequence of observation vectors is computed by

$$\begin{aligned} l(O|\Theta) &= \sum_{S} p(O,S|\Theta) = \sum_{X}\sum_{Y} p(O|Y,X,\Theta)p(Y|X,\Theta)p(X|\Theta) \\ &= \sum_{m=1}^{M}\left\{\prod_{k=1}^{K}\sum_{l=1}^{L} p(O_k^n|O_{1,k-1}^n,m,l,\Theta)\cdot\gamma_{m,l}\right\}\cdot\pi_m. \end{aligned} \tag{50}$$

Note that when $L = 1$ (and hence $\gamma_{m,l} = 1$), Eq. (50) is reduced to the corresponding likelihood computation formula in Ma and Deng (2003).

5. Speech recognition experiments

In this section, we first introduce the experimental paradigm and design of the new recognizer built on the dynamic speech model presented so far. We then report the evaluation results of the new recognizer on the Switchboard data. The evaluation is conducted with reference to the conventional triphone HMM recognizer under the identical experimental conditions.

5.1. Experimental paradigm and recognizer design

In all the experiments reported in this section, we use the N-best re-scoring paradigm to evaluate the new recognizer on the Switchboard spontaneous speech data. The N-best lists (transcription hypotheses) and their phone-level segmentation (or alignment) are obtained from a conventional triphone-based HMM system, which also serves as the benchmark to gauge the recognizer's performance improvement via use of the new speech model. The benchmark HMM system has been described in detail in Bridle et al. (1998), Picone et al. (1999) and will not be described in this paper.

We built the switching dynamic system models for a total of 44 distinct phone-like symbols, including eight context-dependent phones. The phone-like symbol inventory are listed here:

Context-independent phones: aa ae ah eh ao ax er ih iy uh uw l el r w y en n s z sh zh th dh d t sil sp

Context-dependent phones: f v b g p k m ng _f_ _v_ _b_ _g_ _p_ _k_ _m_ _ng_

[2] In fact, this is the only requirement if the N-best re-scoring scheme is used to evaluate the recognizer, as is reported in this paper.

The VTR targets of the above eight context-dependent phones are affected by the anticipatory tongue position associated with the following phone. The targets of those phones with subscript "_" are conditioned on the following phones being "front" vowels (iy, ih, eh, ae and y) and their correspondences without subscript "_" conditioned on the following phones being the "non-front" phones.

Physically, silence (sil) and short pause (sp) have no VTR targets. In our experiments, we assigned pseudo-targets to them so that they can be distinguished from other phones.

5.2. Baseline system

An HMM triphone system was trained on the "train-ws97-a" Switchboard training data (about 160 h). It served as the baseline system for the Workshop'97, and hence we call it the "ws97-baseline" system. The performance of this HMM baseline system is listed as row 3 in Table 1, where "Ref + 100" and "100 best" mean 100-best hypotheses with and without reference included, respectively. We also add the "Oracle" and "By chance" performance into Table 1 to calibrate the recognizer's performance. The "Oracle" WER is calculated by always choosing the best one hypothesis and the "By chance" WER is computed by randomly selecting one out of all hypotheses.

5.3. Training and test sets

In our experiments, we used several sets of training data with an increasing size to train the parameters of the switching dynamic system models of speech. The smallest set of training data consist of one male speaker's data (speaker ID: 1028) extracted from the Switchboard training set "train-ws97-a". It contains several telephone conversations with a total of 30 min long. Due to the use of training data from only one speaker, we avoid the speaker normalization problem for both the VTR targets and for the MFCC observation. We name this set as "1/2 hour" training set to be used in describing the experimental results in the remaining of this section.

An HMM system has also been trained on this "1/2 hour" training set, which we call the "HMM-baseline" system. Its performance, measured by the percentage word error rate (WER) is listed in row 5 of Table 1.

To investigate how the amount of training data and the number of mixture components affect the recognizer performance, we also extracted multiple speakers' data from the same "train-ws97-a" training set to train the speech models. we first added another half an hour data to the original "1/2 hour" training set, where the new data came from 30 different speakers. This increased dataset is called "1 hour" training set. we then added one more hour of training data to the "1

Table 1
Performance (WER) of baseline HMM system

Systems	Ref + 100	100 Best
Oracle	0.0	32.5
By chance	59.6	60.2
ws97-baseline	56.2	56.9

hour" training set. This additional one hour of data came from 50 different speakers. This gives a new, expanded "2 hour" training set.

The test data in all of our experiments consist of all the male speakers from the WS'97 DevTest set. They comprise a total of 23 male speakers, 24 conversations, 1243 utterances, 9970 words, and 50 min of speech. All the 100-best hypotheses for each of the 1243 utterances were generated by the "ws97-baseline" HMM system (Bridle et al., 1998; Picone et al., 1999). All the experiments reported in this paper have used the VTR dynamic regimes derived sub-optimally from the phone alignments provided by this baseline HMM system.

5.4. Experiment I: models trained with one speaker's data

In these experiments, the "1/2 hour" training set is used for model training. First, we use a single mixture component (1-mix) for the mixture-linear dynamic model (MLDM) with switching parameters and we increase the number of H-switching values from one to three. The percentage-WER results are tabulated in Table 2. It is observed that use of H-switching (two or three H values) reduces errors compared with no use of H-switching (one H value only).

We then use two mixture components (2-mix) while again gradually increasing the number of H-switching values. The WER results are listed in Table 3. We observe a similar pattern of error reduction to the previous experiment while uniformly raising the overall recognizer performance level somewhat.

Under identical conditions, compared with the "HMM-baseline" system trained on the same data, the new switching dynamic system model with two mixture components and two H switching values achieves 2.3% absolute WER reduction on the "100-best" case and more than 10% relative WER reduction on the "Ref + 100" case.

5.5. Experiment II: models trained with multiple speakers' data

In these experiments, the amount of training data is increased from the earlier "1/2 hour" training set. The results obtained from use of "1 hour" training dataset are listed in Table 4. For

Table 2
Performance (WER) of MLDM with a single mixture component and different H values (trained with "1/2 hour" training set)

Systems	Ref + 100	100 Best
MLDM:1-mix, 1-H	55.7	58.9
MLDM:1-mix, 2-H	55.0	57.7
MLDM:1-mix, 3-H	55.1	57.2

Table 3
Performance (WER) of MLDM with two mixture components and different H values (trained with "1/2 hour" training set)

Systems	Ref + 100	100 Best
MLDM:2-mix, 1-H	50.7	57.0
MLDM:2-mix, 2-H	50.4	56.6
MLDM:2-mix, 3-H	50.5	56.8

Table 4
Performance (WER) of MLDM with various H values and with various mixture components (trained with "1 hour" training set)

Systems	Ref + 100	100 Best
MLDM:2-mix, 1-H	51.0	57.1
MLDM:2-mix, 2-H	50.1	56.4
MLDM:2-mix, 4-H	50.0	56.4
MLDM:4-mix, 1-H	49.8	56.6
MLDM:4-mix, 2-H	49.6	56.5

Table 5
Performance (WER) of MLDM with various H values and with fixed four mixture components (trained with "2 hour" training set)

Systems	Ref + 100	100 Best
MLDM:4-mix, 1-H	49.5	56.0
MLDM:4-mix, 2-H	48.8	55.8

the two mixture component (2-mix) case, we observe a WER reduction from the use of one H value to the use of more than one H values. Similar observations are made for the four mixture component (4-mix) case, although the WER reduction is of less magnitude.

For the four mixture component (4-mix) case, we further experimented with using "2 hour" training set. The WER results are shown in Table 5. A greater error reduction is observed moving from one H value to two H values when compared with the earlier result of Table 4 with use of fewer training data.

5.6. Some analysis of the model behavior

In this section, we provide some analysis on the behavior of the trained switching dynamic model of speech and on the experimental results. The analysis is based on the fact that the noise covariance matrix R (or noise variance if R is treated as diagonal as in our model implementation) is estimated according to Eq. (27), where $e2$ is the difference of the actual MFCC and the output of the $h(\cdot)$ function. Thus, if the function $h(\cdot)$ accurately describes the relation between the VTR space and the MFCC space, the estimated R will be small. Otherwise, the estimated R will be large. Therefore, the accuracy of approximating the physically nonlinear relation between the VTR space and the MFCC space using a piecewise linear function as implementing by switching-H values can be assessed by examining the size of the estimated noise variance, R.

As typical examples, the diagonal values of the estimated R for phone models "aa", "d", and "n" are shown in Tables 6–8, respectively. (Column three lists the average values of these diagonal elements.) We observe that these estimated variance values are strictly decreasing with the increasing number of H-switching values. This suggests that the approximation accuracy by using the piecewise linear function is improved with the use of more linear functions. When the number of H-switching values is increased from one to four, the average R value is decreased from 22.1 to 18.6 for "aa", from 22.2 to 19.6 for "d", and from 20.0 to 17.6 for "n", respectively.

Table 6
Values of diagonal elements of R noise variance for the phone model "aa" as a function of the number of H switching values.

No. of H	Diagonal elements of R	Average
1 H	10.0 14.6 17.1 22.5 21.7 18.9 32.3 28.4 29.4 23.5 24.4 21.8	22.1
2 H	8.87 13.4 14.6 20.0 21.3 20.1 29.7 27.5 30.9 21.7 23.4 21.8	21.1
4 H	6.92 11.5 13.4 18.6 17.8 19.2 26.9 19.7 27.9 23.4 17.6 19.7	18.6

Table 7
Values of diagonal elements of R noise variance for the phone model "d" as a function of the number of H switching values

No. of H	Diagonal elements of R	Average
1 H	8.9 17.5 18.8 16.9 19.8 25.6 27.4 27.2 30.2 27.2 26.7 20.7	22.2
2 H	5.6 16.8 16.4 15.1 18.8 24.9 26.4 23.4 31.4 23.6 24.4 19.8	20.6
4 H	5.2 17.0 16.7 15.0 17.0 24.8 24.9 22.9 27.4 21.8 24.7 19.6	19.6

Table 8
Values of diagonal elements of R noise variance for the phone model "n" as a function of the number of H switching values

No. of H	Diagonal elements of R	Average
1 H	6.0 15.9 16.8 15.5 18.7 20.7 18.9 27.2 31.0 26.3 21.4 22.0	20.0
2 H	6.1 15.6 15.0 16.6 18.1 21.1 17.8 26.9 24.1 24.4 20.8 19.6	18.8
4 H	5.5 14.5 13.9 16.4 18.4 17.9 16.0 26.1 20.9 23.9 18.7 19.0	17.6

It is interesting to note that the mild improvement in the linear piecewise approximation accuracy as reflected by the reduced R value is correlated with the mild WER reduction in the speech recognition results presented earlier in this section.

6. Conclusions

A new version of the switching dynamic model of speech, the linear dynamic system model with mixed-level (segment and frame) switching parameters, is presented in this paper. The segment-level switching parameters in the target-directed state equation is the same as that in the earlier single-level switching model (Ma and Deng, 2003), and the new frame-level switching parameters in the observation or measurement equation is introduced in the current model. The use of the frame-level switching parameters effectively provides piecewise linear functions to approximate the physically nonlinear function between the partially observable VTR space and the observable MFCC space in the observation equation.

A series of speech recognition experiments have been carried out to evaluate the new mixed-level switching model. The experimental results show that the approximation accuracy is improved with an increasing number of H-switching values (about a 10% reduction in the estimated measurement noise variances). The speech recognizer built from the mixed-level switching dynamic system model using the N-best rescoring evaluation paradigm also shows some varying degrees of word error rate reduction compared with using the single-level switching model. The new recognizer built with mixed-level switching parameters achieved a lower error rate than a baseline HMM system evaluated under identical experimental conditions.

Acknowledgements

We thank Prof. Mari Ostendorf and other reviewers for useful comments and suggestions that improve the technical contents and the presentation of this paper.

References

Bridle, J., Deng, L., Picone, J., Richards, H., Ma, J., Kamm, T., Schuster, M., Pike, S., Reagan, R., 1998. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, pp. 1–61.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the *EM* algorithm. J. Royal Statist. Soc. B-39, 1–38.

Deng, L., Ma, J., 1999. A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics. Proc. Eurospeech 4, 1499–1502.

Deng, L., Ma, J., 2000. Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics. J. Acoust. Soc. Am. 108 (6), 3036–3048.

Deng, L., 1993. A stochastic model of speech incorporating hierarchical nonstationarity. IEEE Trans. Speech Audio Process. 1, 471–474.

Deng, L., 1999. Computational models for speech production. In: Computational Models of Speech Pattern Processing (NATO ASI). Springer, Berlin, pp. 199–214.

Deng, L., Aksmanovic, M., 1997. Speaker-independent phonetic classification using hidden Markov models with state-conditioned mixtures of trend functions. IEEE Trans. Speech Audio Process. 5 (4), 319–324.

Digalakis, V., Rohlicek, J., Ostendorf, M., 1993. ML estimation of a stochastic linear system with the *EM* algorithm and its application to speech recognition. IEEE Trans. Speech Audio Process. 1, 431–442.

Gish, H., Ng, K., 1993. A segmental speech model with applications to word spotting. Proc. ICASSP II, 447–450.

Kitagawa, K., 1987. Non-Gaussian state-space modeling of nonstationary time series. J. Am. Stat. Assoc. 82, 1032–1041.

Ma, J., Deng, L., 2003. Target-directed mixture linear dynamic models for spontaneous speech recognition. IEEE Trans. Speech Audio Process. (to appear).

Mendel, J.M., 1995. Lessons in Estimation Theory for Signal Processing, Communications and Control. Prentice-Hall, Englewood Cliffs, NJ.

Ostendorf, M., Digalakis, V., Kimball, O., 1996. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Process. 4, 360–378.

Picone, J., Pike, S., Reagan, R., Kamm, T., Bridle, J., Deng, L., Ma, J., Richards, H., Schuster, M., 1999. Initial evaluation of hidden dynamic models on conversational speech. Proc. ICASSP I, 109–112.

Shumway, R.H., 1982. An approach to time series smoothing and forecasting using the EM algorithm. J. Time Series Anal. 3 (4), 253–264.

Shumway, R.H., Stoffer, D.S., 1991. Dynamic linear models with switching. J. Am. Stat. Assoc. 86, 763–769.

Streit, R., Luginbuhl, T., 1998. Probabilistic multi-hypothesis tracking. In: Studies in Probabilistic Multi-Hypothesis Tracking and Related Topics, pp. 1–51.

Tanizaki, H., 1996. Nonlinear Filters, second ed. Springer, Berlin.