ELSEVIER

Signal Processing 39 (1994) 263–275

SIGNAL PROCESSING

# State-dependent time warping in the trended hidden Markov model

D.X. Sun[a], L. Deng[b,*], C.F.J. Wu[c]

[a] *State University of New York at Stony Brook, NY 11794-3600, USA*
[b] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*
[c] *University of Michigan, Ann Arbor, MI 48109-1027, USA*

Received 16 July 1992; revised 11 March 1994

**Abstract**

In this paper we present an algorithm for estimating state-dependent polynomial coefficients in the nonstationary-state hidden Markov model (or the trended HMM) which allows for the flexibility of linear time warping or scaling in individual model states. The need for the state-dependent time warping arises from the consideration that due to speaking rate variation and other temporal factors in speech, multiple state-segmented speech data sequences used for training a single set of polynomial coefficients often vary appreciably in their sequence lengths. The algorithm is developed based on a general framework with use of auxiliary parameters, which, of no interests in themselves, nevertheless provide an intermediate tool for achieving maximal accuracy for estimating the polynomial coefficients in the trended HMM. It is proved that the proposed estimation algorithm converges to a solution equivalent to the state-optimized maximum likelihood estimate. Effectiveness of the algorithm is demonstrated in experiments designed to fit a single trended HMM simultaneously to multiple sequences of speech data which are different renditions of the same word yet vary over a wide range in the sequence length. Speech recognition experiments have been performed based on the standard acoustic-phonetic TIMIT database. The speech recognition results demonstrate the advantages of the time-warping trended HMMs over the regular trended HMMs measured about 10 to 15% improvement in terms of the recognition rate.

**Zusammenfassung**

In dieser Arbeit stellen wir einen Algorithmus zur Schätzung zustandsabhängiger Polynomkoeffizienten beim nichtstationären Hidden Markov Model (THMM) vor, der eine flexible Zeitskalierung der individuellen Modellzustände gestattet. Der Grund für eine zustandsabhängige Zeitänderung folgt aus der Beobachtung, daß wegen der Sprechratenänderung und anderen zeitabhängigen Faktoren in der Sprache verschiedene Datenfolgen, die man als Lernfolgen eines Satzes von Polynomkoeffizienten verwendet, oft beträchtlich in ihrer Länge variieren. Der entwickelte Algorithmus verwendet Hilfsparameter die zwar selbst keine unmittelbare Bedeutung besitzen, aber nichtsdestotrotz geeignet sind, die maximale Genauigkeit der Schätzung der Polynomkoeffizienten des THMM zu erzielen. Es wird gezeigt, daß der vorgeschlagene Schätzalgorithmus gegen eine Lösung konvergiert, die der zustandsoptimierten Maximum-Likelihood-Schätzung äquivalent ist. Die Brauchbarkeit des Algorithmus wird anhand von Beispielen gezeigt, bei denen ein einzelnes THMM gleichzeitig für mehrere Folgen, die das gleiche Wort mit stark unterschiedlichen

[*] Corresponding author.

Elsevier Science B.V.
*SSDI* 0165-1684(94)00054-4

Folgenlängen darstellen, entworfen wird. Untersuchungen zur Spracherkennung wurden mit Hilfe der akustisch-phonetischen Datenbank TIMIT durchgeführt. Sie zeigen die Vorteile der zeitskalierten THMM gegenüber den regulären THMM von ca. 10–15% Verbesserung hinsichtlich der Erkennungsrate.

**Résumé**

Dans cet article, nous présentons un algorithme pour l'estimation des coefficients du polynôme dépendant de l'état dans le modèle de Markov à état non-stationnaire caché (ou HMM) qui permet une déformation ou un changement d'échelle flexible linéaire temporellement dans des états de modèles individuels. Les besoins pour des déformations temporelles dépendant de l'état surviennent lorsque l'on prend en considération que, due aux variations de débit et à d'autres facteurs temporels en parole, les séquences de données de parole segmentées utilisées pour suivre un seul ensemble de coefficients polynomiaux varient souvent de manière appréciable dans leurs longeurs. L'algorithme est développé en se basant sur un schéma général avec l'utilisation de paramètres auxiliaires, lesquelles, bien que n'ayant pas d'intérêt par eux-mêmes, procure néanmoins un outil intermédiaire pour atteindre une précision maximale pour l'estimation des coefficients du polynôme du HMM. Il est prouvé que l'algorithme d'estimation proposé converge vers une solution équivalente à l'estimation du maximum de vraisemblance à état optimisé. L'efficacité de l'algorithme est démontré par des expérimentations concues pour utiliser un seul HMM simultanément sur des séquences multiples de données de la parole ayant des rendus différents du même mot et dont la longeur de la séquence varie dans une grande mesure. Des expérimentations de reconnaissance de la parole ont été réalisées sur la base de données standards acoustique-phonétique TIMIT. Les résultats de reconnaissance de la parole démontrent que les avantages des HMMs à déformation temporelle sur les approches HMM normales sont de 10 à 15% en terme de taux de reconnaissance.

*Key words*: Speech signal; Acoustic transition; Scaling; Hidden Markov model; Nonstationarity; Time warping; Auxiliary parameter; Viterbi algorithm

## 1. Introduction

The standard hidden Markov model (HMM) developed in [1, 10] and widely in use for speech recognition [11] contains the mathematical structure of a (hidden) Markov chain with each state associated with a distinct independent and identically distributed (i.i.d.) or a *stationary* random process. The model is used as a type of data-generator for speech signals and approximates the near continuously varying speech signals in a piece-wise constant manner. Such an approximation would be a reasonably good one when each state is intended to represent only a short portion of sonorant sounds. However, since the acoustic patterns of continuously spoken speech sounds are almost never stationary in nature (c.f. [12]), it is desirable to improve this piece-wise constant approximation. Some recent work has been intended to achieve improvement of the approximation accuracy via use of piecewise polynomials, which was developed within a general framework of the nonstationary-state HMM (or the trended HMM) [4]. In that model, polynomial trend functions (or regression functions on time) are used as time-varying means in the output Gaussian distributions in the HMM states. The observation vector sequences, $O_t$, $t = 1, 2, \ldots, T$, are generated from the model according to

$$O_t = \sum_{m=0}^{M} \boldsymbol{B}_i(m) t^m + R_t(\Sigma_i), \qquad (1)$$

where the first term is the state-dependent polynomial regression function of order $M$, the second term is the residual noise assumed to be the output of an i.i.d., zero-mean Gaussian source with a state-dependent covariance matrix $\Sigma_i$, and state $i$ at a given time $t$ is determined by evolution of the underlying Markov chain in the HMM.

The trended HMM takes a significant leap from the standard HMM in its generality and in its economical use of model parameters for approximating highly dynamical patterns of the speech signal. But despite these desirable properties, the trended HMM has nevertheless introduced its own specific problem in speech modeling and recognition applications. That is, speaking rate variation from one speech data sequence (which we call 'token') to another, given the same underlying phonetic representation for the HMM states, must be normalized. Because, unlike the standard stationary-state HMM, the polynomial trend function for each state of the trended HMM is varying with time, significant variability is necessarily introduced when using the same, single trend function to describe speech data from multiple (state-segmented) tokens from the same word with varying token durations. The varying token durations are the results of speaking rate variations and of other temporal factors in speech (e.g. [6, 9]).

To alleviate this difficulty, we have developed an algorithm which implements time warping in the state-dependent polynomial regression functions. Since the source of the difficulty is that each state in the model in (1) is not flexible enough to fit simultaneously to multiple tokens of speech data (which are different renditions of the same word yet vary over the sequence length), we introduce the token-dependent auxiliary time-warping parameter to normalize out the effect of the token duration. The time-warping parameters are called auxiliary ones because they are not considered intrinsic parameters of the model, but, rather, they are used only as a tool to improve the accuracy in estimating the intrinsic parameters – state-dependent polynomial coefficients $\boldsymbol{B}_i(m)$ in (1). These auxiliary parameters work by linearly adjusting the state duration using one separate 'optimal' scale for each individual token. The sole function of the auxiliary parameters is to group all the tokens available for training in an optimal way such that the variability of the state duration does not affect estimation of the intrinsic parameters of the model. The proposed algorithm is a two-stage iterative optimization procedure where estimation of the auxiliary parameters and the polynomial regression parameters is carried out alternatively. To simplify the complexity introduced by the Markov chain in the trended HMM, the proposed algorithm is embedded within each step of the global segmental $K$-means-like algorithm [8]. We prove in this paper that the proposed two-stage iterative algorithm converges to a solution which is equivalent to the state-optimized maximum likelihood estimates.

A related work for modeling the nonstationary features of speech signals is the dynamic-programming-based template matching algorithm proposed in [7]. The major difference is that the time-warping trended HMMs are parametric models while the dynamic-programming-based template matching algorithm is nonparametric in nature.

The organization of the paper is as follows. In Section 2, we give the formal formulation of the trended HMM incorporating the state-dependent time-warping mechanism. The two-stage iterative optimization algorithm, as a kernel step in the global segmental $K$-means-like algorithm, is presented in Section 3. This section also provides the convergence proof of the algorithm. Experimental results on fitting the trended HMMs to speech data, which are different renditions of the same word but vary significantly over the speech token length, are shown in Section 4. Comparisons between the data-fitting results with and without using the time-warping mechanism illustrate the need for time warping in the trended HMM and show effectiveness of the proposed algorithm for achieving the desired time warping. In Section 5, we present the results from the speech recognition experiments based on the standard acoustic-phonetic TIMIT database. The recognition results demonstrate the advantages of the time-warping trended HMMs over the regular trended HMMs.

## 2. Model formulation

The trended HMM with state-dependent time warping studied in this paper can be viewed as a data-generative type of model and be formally defined in terms of the following form for data generation:

$$O_t = g_i\left(\frac{t - \tau_i}{\lambda_i}\right) + R_t(\Sigma_i), \quad t = 1, \dots, T, \tag{2}$$

where $O_t$, $t = 1, \dots, T$, is the observation data sequence generated by the model (possibly vector-valued such as the cepstral vectors computed from the speech waveform), $i$ is the label of the state in the HMM (at time frame $t$), and $g_i(\cdot)$ is a deterministic function of time $t$ and indexed by state label $i$. The form of $g_i(\cdot)$ is chosen in this study to be polynomial functions since they are not only computationally simple, but also provide good approximations to most arbitrary functions. (For speech data in the spectral domain, relatively low-order polynomial functions are expected to suffice for acceptably good approximations.) In (2), the time-shift parameter $\tau_i$ registers the time when state $i$ in the HMM is just entered before the function $g_i(\cdot)$ becomes effective; i.e., $(t - \tau_i)$ represents the sojourn time in state $i$ at time $t$. $\lambda_i$ is the time-warping parameter associated with state $i$, which transforms the time points within state $i$ to a canonical scale (see Section 3 for detail). When multiple tokens are used in the training step, parameters $\tau_i, \lambda_i$ are also made dependent on each individual token. Note that parameters $\tau_i, \lambda_i$ are considered as auxiliary parameters, which will be discarded at the termination of the training step.

Now given $K$ tokens in the training data and given that the generic regression function $g_i(\cdot)$ takes a polynomial form up to order $M$, the following specific data-generative trended HMM (generating the $K$ training tokens) is considered in our discussion:

$$O_{r,t} = \sum_{m=0}^{M} B_{i,m}\left(\frac{t-\tau_{r,i}}{\lambda_{r,i}}\right)^m + R_{r,t}(\Sigma_i),$$
$$t = 1,\ldots,n_r,\ r = 1,\ldots,K, \tag{3}$$

where $O_{r,t}$ denotes the $r$th token of the training data at time $t$, $n_r$ is the length of the $r$th token, and $B_{i,m}$'s are the polynomial coefficients, considered as the intrinsic parameters of the model, to be estimated. Altogether, the parameters of this trended HMM are summarized as consisting of the following four sets:

1. $A = (a_{ij})$, $i,j = 1,\ldots,N$, is the transition probability matrix of the underlying Markov chain with a total of $N$ states.
2. $B = (B_{i,m})$, $i = 1,\ldots,N$, $m = 0,\ldots,M$, are the polynomial coefficients, of order $m$ and associated with state $i$, in the state-dependent deterministic regression function of time.
3. $(\tau_{r,i}, \lambda_{r,i})$, $i = 1,\ldots,N$, $r = 1,\ldots,K$, are the auxiliary parameters in the polynomial regression function associated with state $i$ for token $r$.
4. $\Sigma_i$, $i = 1,\ldots,N$, are the covariance matrices, associated with state $i$, of the Gaussian i.i.d. residual $R_{r,t}(\Sigma_i)$.

Note that, in the above, only $(A, B, \Sigma)$ are the intrinsic parameters, which are independent of the training token, of the trended HMM; while the time-shift and warping parameters $(\tau_{r,i}, \lambda_{r,i})$ serve only as auxiliaries whose role is to adjust the length of each training token for obtaining accurate estimates of the intrinsic model parameters. In the speech recognition step, the auxiliary parameters for the unknown utterance are estimated independently so as to adjust the duration of this new token to its own optimal scale for matching the trended HMM obtained in the training step.

## 3. Algorithm for parameter estimation

In this section, we present an algorithm for estimating the parameters in the trended HMM containing the state-dependent time-warping mechanism. This algorithm is embedded within each step of a global iteration, whose goal is to reduce the training complexity involving a multiple-state Markov chain to essentially that involving no Markov chain (or single-state Markov chain). (This is in the same spirit as the segmental $K$-means algorithm [8].) This global segmental $K$-means-like algorithm involves two iterative steps: the segmentation step and the optimization step. The parameters $A = (a_{ij})$ and $\tau_{r,i}$'s are readily determined from the result of the segmentation step (for fixed $\lambda_{r,i}$'s and $B_{i,m}$'s), while $\lambda_{r,i}$'s and $B_{i,m}$'s are estimated in the optimization step (for fixed $A = (a_{ij})$ and $\tau_{r,i}$'s). The segmentation step can be carried out by a Viterbi-like algorithm [11], with a slight modification in incorporating an additional optimization loop for the state sojourn time. We mainly focus on the optimization step in this paper.

Once all the state boundaries are determined in the segmentation step, the entire process for parameter estimation of the trended HMM is broken down into several independent and essentially identical processes for estimating the parameters associated with each individual state. Therefore, for notational simplicity, we hereafter drop the state label $i$ and consider only the parameter estimation for one single state. Further, we assume that the covariance matrix $\Sigma_i$ is diagonal with the $d$th diagonal component $\sigma_d^2$. Hence the estimation for each dimension of the vector can be treated separately, and for simplicity in writing we consider scalar-valued data sequences only. Thus the $K$ tokens of the vector-valued observation speech data can be simplified into $K$ samples of one-dimensional sequences each with length $n_r$, $r = 1, \ldots, K$,

$$\begin{array}{ll} O_{1,1}, O_{1,2}, \ldots, O_{1,n_1} & \text{(token 1)} \\ O_{2,1}, O_{2,2}, \ldots, O_{2,n_2} & \text{(token 2)} \\ \vdots & \vdots \\ O_{K,1}, O_{K,2}, \ldots, O_{K,n_K} & \text{(token } K\text{).} \end{array}$$

After these simplifications (without loss of generality), we may write the likelihood function contributed by the data points within a state as

$$L(B_0, \ldots, B_M; \lambda_1, \ldots, \lambda_K; \sigma^2) \propto (\sigma)^{-\sum_{r=1}^{K} n_r} \exp\left\{\frac{1}{2\sigma^2} \times \sum_{r=1}^{K} \sum_{t=1}^{n_r} \left[O_{r,t} - \sum_{m=0}^{M} B_m \left(\frac{t}{\lambda_r}\right)^m\right]^2\right\}. \tag{4}$$

Note that in maximizing (4), the parameter $\sigma$ and the remaining parameters can be treated separately. An estimate of $\sigma$ can be obtained very easily from just the residuals based on the estimates for parameters $B_m$ and $\lambda_r$. On the other hand, maximization of (4) with respect to only $B_m$ and $\lambda_r$ is equivalent to minimization of the quadratic objective function

$$Q = \sum_{r=1}^{K} \sum_{t=1}^{n_r} \left[O_{r,t} - \sum_{m=0}^{M} B_m \left(\frac{t}{\lambda_r}\right)^m\right]^2. \tag{5}$$

Direct minimization of (5), unfortunately, is a multidimensional nonlinear regression problem, which would require intensive computation (and would also guarantee no global optimum in general). As an efficient solution to this multi-dimensional non-linear regression problem, we propose a two-stage alternating optimization method by taking computational advantages of linear regression and of efficient methods of root finding for one-dimensional polynomial functions.

### 3.1. Two-stage alternating algorithm

The basic idea behind this algorithm is to decompose the complex optimization problem for (5) into two separate stages as follows.

*Stage 1*: Given $\lambda_r$, $r = 1, \ldots, K$, first linearly scale times as $x_{r,t} = t/\lambda_r$. Then $\underline{B} = (B_0, \ldots, B_M)'$ can be estimated by the solution of

$$\min_{\underline{B}} \sum_{r=1}^{K} \sum_{t=1}^{n_r} \left[O_{r,t} - \sum_{m=0}^{M} B_m x_{r,t}\right]^2. \tag{6}$$

This can be easily solved by the ordinary linear regression method. The estimate can be written as the closed-form result:

$$\hat{\underline{B}}_{(M+1)\times 1} = (X'X)^{-1} X' \underline{O},$$

where

$$X_{(n_1 + \cdots + n_K)\times(M+1)} = \begin{array}{c} \left[\begin{array}{ccccc} 1 & x_{11} & x_{11}^2 & \cdots & x_{11}^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n_1} & x_{1,n_1}^2 & \cdots & x_{1,n_1}^M \\ 1 & x_{21} & x_{21}^2 & \cdots & x_{21}^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,n_2} & x_{2,n_2}^2 & \cdots & x_{2,n_2}^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{K,1} & x_{K,1}^2 & \cdots & x_{K,1}^M \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{K,n_K} & x_{K,n_K}^2 & \cdots & x_{K,n_K}^M \end{array}\right] \end{array} \begin{array}{l} \\ \text{token 1} \\ \\ \\ \text{token 2} \\ \\ \vdots \\ \\ \text{token } K \\ \\ \end{array}$$

and

$$\underline{O} = (O_{1,1}, \ldots, O_{1,n_1}, O_{2,1}, \ldots, O_{2,n_2}, \ldots, O_{K,1}, \ldots, O_{K,n_K})'.$$

*Stage 2*: Given $B_0, \ldots, B_M$ obtained from Stage 1, we then estimate $\lambda_r$ for each $r$, $r = 1, \ldots, K$, independently by minimizing the objective function for the $r$th token:

$$\hat{Q}_r = \sum_{t=1}^{n_r}\left[O_{r,t} - \sum_{m=0}^{M} B_m\left(\frac{t}{\lambda_r}\right)^{-m}\right]^2,$$

which, after removing optimization-independent terms, is equivalent to minimizing

$$\begin{aligned} Q_r &= \sum_{t=1}^{n_r}\left[\sum_{m=0}^{M}(\lambda_r)^{-m}\left(\sum_{i=0}^{m} B_i B_{m-i} t^m - 2O_{r,t} B_m t^m\right)\right. \\ &\quad \left. + \sum_{m=M+1}^{2M}(\lambda_r)^{-m}\left(\sum_{i=1}^{m-M} B_i B_{2M-m} t^m\right)\right] \\ &= \sum_{m=0}^{2M} C_m(\lambda_r)^{-m}, \end{aligned}$$

where

$$C_m = \begin{cases} \displaystyle\sum_{i=0}^{m} B_i B_{m-i} t^m - 2O_{r,t} B_m t^m, \\ \quad m = 0, \dots, M, \\ \displaystyle\sum_{i=1}^{m-M} B_i B_{2M-m} t^m, \\ \quad m = M+1, \dots, 2M. \end{cases}$$

Minimization of $Q_r$ can be achieved by setting

$$\frac{\mathrm{d}Q_r}{\mathrm{d}\lambda_r} = \sum_{m=1}^{2M}(-m)C_m(\lambda_r)^{-m-1} = 0,$$

or

$$\sum_{m=1}^{2M} mC_m(\lambda_r)^{2M-m-1} = 0,$$

which can be easily solved by invoking standard polynomial root finding subroutines.

Putting this two-stage alternating algorithm into an iteration, we have the following complete procedure for estimating $B_m$ and $\lambda_r$:

(1) Set initial values for $\lambda_1^{(0)}, \dots, \lambda_K^{(0)}$ (e.g. set all $\lambda_1^{(0)}, \dots, \lambda_K^{(0)} = 1$); Set the iteration count $n = 0$;
(2) Pass the values of $\lambda_1^{(n)}, \dots, \lambda_K^{(n)}$ to Stage 1 and obtain estimates $B_0^{(n)}, \dots, B_M^{(n)}$;
(3) Estimate $\lambda_1^{(n+1)}, \dots, \lambda_K^{(n+1)}$, given $B_0^{(n)}, \dots, B_M^{(n)}$, through Stage 2;
(4) Check convergence: if convergence occur, finish; else set $n = n + 1$ and go to (2).

The algorithm is considered to be convergent when the difference between the $Q_r$ values of two successive iterations becomes smaller than a predetermined tolerance value.

### 3.2. Convergence properties of the algorithm

In this subsection, we prove convergence of the above proposed two-stage optimization algorithm and prove its equivalence to maximum likelihood estimation. The proofs follow the notations used in [3].

Let $\Theta = (B_0, \dots, B_M)$, $\Lambda = (\lambda_1, \dots, \lambda_K)$. Denote the estimates at the $n$th iteration of the algorithm by $\Theta^{(n)}, \Lambda^{(n)}$.

**Theorem 1.** *The likelihood function in Eq. (4) is nondecreasing over iterations of the algorithm, or*

$$L(\Theta^{(n+1)}, \Lambda^{(n+1)}) \geqslant L(\Theta^{(n)}, \Lambda^{(n)}). \tag{7}$$

**Proof.** We have

$$\begin{aligned} &L(\Theta^{(n+1)}, \Lambda^{(n+1)}) - L(\Theta^{(n)}, \Lambda^{(n)}) \\ &\quad = \underbrace{[L(\Theta^{(n+1)}, \Lambda^{(n+1)}) - L(\Theta^{(n)}, \Lambda^{(n+1)})]}_{\text{(I)}} \\ &\quad + \underbrace{[L(\Theta^{(n)}, \Lambda^{(n+1)}) - L(\Theta^{(n)}, \Lambda^{(n)})]}_{\text{(II)}}. \end{aligned}$$

Since Stage 1 and Stage 2 guarantee nonnegativeness of (I) and (II), respectively, we immediately prove the theorem. □

**Theorem 2.** *Suppose that the following 'identifiability condition' is satisfied:*

$$\begin{aligned} &L(\Theta^{(n+1)}, \Lambda^{(n+1)}) - L(\Theta^{(n)}, \Lambda^{(n)}) \\ &\quad \geqslant \eta \,\|(\Theta^{(n+1)}, \Lambda^{(n+1)}) - (\Theta^{(n)}, \Lambda^{(n)})\|, \end{aligned} \tag{8}$$

*where $\eta$ is a fixed constant and $\|\cdot\|$ is the Euclidean norm. Then $(\Theta^{(n)}, \Lambda^{(n)})$ converges to some $(\Theta^*, \Lambda^*)$ in the parameter space. (Note that the identifiability condition implies that if one set of parameters is different from another set, then their associated likelihoods must also be different; otherwise, the two sets of parameters would not be distinguishable.)*

**Proof.** From Theorem 1, the sequence $L(\Theta^{(n)}, \Lambda^{(n)})$ must converge to some value $L^* < \infty$. Hence, for any $\varepsilon > 0$, there exists an $n(\varepsilon)$ such that, for all $n \geqslant n(\varepsilon)$ and all $J \geqslant 1$,

$$\sum_{j=1}^{J} \{L(\Theta^{(n+j)}, \Lambda^{(n+j)}) - L(\Theta^{(n+j-1)}, \Lambda^{(n+j-1)})\}$$
$$= L(\Theta^{(n+J)}, \Lambda^{(n+J)}) - L(\Theta^{(n)}, \Lambda^{(n)}) < \varepsilon.$$

Applying the identifiability condition, we obtain

$$\varepsilon > \eta \sum_{j=1}^{J} \|(\Theta^{(n+j)}, \Lambda^{(n+j)}) - (\Theta^{(n+j-1)}, \Lambda^{(n+j-1)})\|$$
$$\geqslant \eta \left\| \sum_{j=1}^{J} (\Theta^{(n+j)}, \Lambda^{(n+j)}) - (\Theta^{(n+j-1)}, \Lambda^{(n+j-1)}) \right\|$$
$$\geqslant \eta \|(\Theta^{(n+J)}, \Lambda^{(n+J)}) - (\Theta^{(n)}, \Lambda^{(n)})\|.$$

This inequality implies that the sequence $(\Theta^{(n)}, \Lambda^{(n)})$ is a Cauchy sequence, and it must converge, say, to $(\Theta^*, \Lambda^*)$. ($(\Theta^*, \Lambda^*)$ would be the ultimate estimate to be obtained if we would let the algorithm iterate infinitely many times.) □

**Corollary.** *$(\Theta^*, \Lambda^*)$ is a stationary point of the likelihood function, i.e.,*

$$\frac{\mathrm{d}}{\mathrm{d}\Theta} L(\Theta^*, \Lambda^*) = 0, \qquad \frac{\mathrm{d}}{\mathrm{d}\Lambda} L(\Theta^*, \Lambda^*) = 0, \tag{9}$$

*where the derivatives are taken as*

$$\frac{\mathrm{d}}{\mathrm{d}\Theta} L(\Theta^*, \Lambda^*) = \frac{\mathrm{d}}{\mathrm{d}\Theta} L(\Theta, \Lambda^*)\bigg|_{\Theta = \Theta^*}.$$

**Proof.** Stages 1 and 2 described in the last subsection assure that

$$\frac{\mathrm{d}}{\mathrm{d}\Theta} L(\Theta^{(n+1)}, \Lambda^{(n+1)}) = 0,$$

$$\frac{\mathrm{d}}{\mathrm{d}\Lambda} L(\Theta^{(n)}, \Lambda^{(n+1)}) = 0. \tag{10}$$

By taking limits of both sides (let $n \to \infty$) and then applying Theorem 2, the corollary follows immediately. □

Our final task is to prove that $(\Theta^*, \Lambda^*)$ is in fact the maximum likelihood estimate. For this it suffices to show that the second-order derivative $(\mathrm{d}^2/\mathrm{d}(\Theta, \Lambda)^2) L(\Theta^*, \Lambda^*)$ is negative definite.

**Theorem 3.** *Suppose that*

$$\frac{\mathrm{d}^2}{\mathrm{d}\Theta^2} L(\Theta^{(n)}, \Lambda^{(n)}), \qquad \frac{\mathrm{d}^2}{\mathrm{d}\Lambda^2} L(\Theta^{(n)}, \Lambda^{(n+1)})$$

*are both negative definite with eigenvalues bounded away from zero. Then*

$$\frac{\mathrm{d}^2}{\mathrm{d}(\Theta, \Lambda)^2} L(\Theta^*, \Lambda^*)$$

*is negative definite.*

**Proof.** We have

$$\frac{\mathrm{d}^2}{\mathrm{d}(\Theta, \Lambda)^2} L(\Theta^{(n)}, \Lambda^{(n)})$$
$$= \begin{bmatrix} \frac{\mathrm{d}^2}{\mathrm{d}\Theta^2} L(\Theta^{(n)}, \Lambda^{(n)}) & \frac{\mathrm{d}^2}{\mathrm{d}\Theta\,\mathrm{d}\Lambda} L(\Theta^{(n)}, \Lambda^{(n)}) \\ \frac{\mathrm{d}^2}{\mathrm{d}\Theta\,\mathrm{d}\Lambda} L(\Theta^{(n)}, \Lambda^{(n)}) & \frac{\mathrm{d}^2}{\mathrm{d}\Lambda^2} L(\Theta^{(n)}, \Lambda^{(n)}) \end{bmatrix}.$$

Since

$$\frac{\mathrm{d}^2}{\mathrm{d}\Lambda^2} L(\Theta^{(n)}, \Lambda^{(n)}) - \frac{\mathrm{d}^2}{\mathrm{d}\Lambda^2} L(\Theta^{(n)}, \Lambda^{(n+1)}) \to 0$$

and $(\mathrm{d}^2/\mathrm{d}\Lambda^2) L(\Theta^{(n)}, \Lambda^{(n+1)})$ is negative definite with eigenvalues bounded away from zero, we conclude that $(\mathrm{d}^2/\mathrm{d}\Lambda^2) L(\Theta^{(n)}, \Lambda^{(n)})$ is also negative definite with eigenvalues bounded away from zero. On the other hand, from (10) it is obvious that

$$\frac{\mathrm{d}^2}{\mathrm{d}\Theta\,\mathrm{d}\Lambda} L(\Theta^{(n)}, \Lambda^{(n)}) = 0,$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\Theta\,\mathrm{d}\Lambda} L(\Theta^{(n)}, \Lambda^{(n)}) = 0.$$

Putting together the above facts, we have that $(\mathrm{d}^2/\mathrm{d}(\Theta, \Lambda)^2) L(\Theta^{(n)}, \Lambda^{(n)})$ converges to a negative definite matrix, which is $(\mathrm{d}^2/\mathrm{d}(\Theta, \Lambda)^2) L(\Theta^*, \Lambda^*)$. □

In summary, Theorem 2 gives the convergence property of the two-stage iterative algorithm, and the corollary of Theorem 2, combined with

Theorem 3, proves that the proposed algorithm indeed leads to maximum likelihood estimates.

## 4. Results on fitting trended HMMs to speech data

In this section, we apply the state-dependent time-warping HMM, trained with the algorithm described in Section 3, to fit the acoustic-parameter sequences from different renditions of the same word which vary in the sequence duration. In particular, we compare goodness of the data fitting between the trended HMM incorporating the time warping discussed in this paper and that without the time-warping mechanism built in i.e. the model in [4].

The first set of speech data was taken from two tokens of the word *peek* /pi:k/ spoken by a native English speaker with intentionally different speaking rates. The second set of speech data were selected from the TIMIT acoustic-phonetic continuous speech corpus. The two tokens used for illustration were excised from the same word *bike* in the sentence 'sx332' uttered by two male speakers from dialect region 2 and region 7.

The raw speech data was in the form of digitally sampled signal at 16 kHz. The mel-frequency cepstral coefficients [2] were computed from the raw data with a Hamming window of duration 25.6 ms and with a frame rate of 10 ms. Trended HMMs with three states and with order three in the state-dependent polynomial regression functions on time are used to fit speech data from word *peek*. For speech data from the word *bike* (in TIMIT), four-state trended HMMs with order three in the regression functions are used. For the sake of space saving and for purposes of illustration, we show here only the data fitting results for the first-order cepstral coefficient C1 from word *peek* and for the third-order cepstral coefficient C3 from *bike*. (Similar results were obtained for other cepstral coefficients.)

The solid, less smoothed lines in both graphs of Fig. 1 are C1 data sequences of two tokens uttered by the same speaker from the same word *peek*. The vertical axis represents the magnitude of C1 data and the horizontal axis is the frame number (frame size 10 ms). Superimposed on the same graphs in Fig. 1 as dotted, more smoothed lines are the three sequentially advanced state-dependent polynomial regression functions in the previously developed trended HMM without time warping (i.e. the model in [4]). The polynomial coefficients in these regression functions were estimated from the two tokens using the algorithm described also in [4]. Given the model parameters, the process of fitting models to the data proceeded by first finding the optimal segmentation of the data into the HMM states and then fitting the segmented data using the regression functions associated with the corresponding states. (Optimal segmentation of the data was obtained by a Viterbi-like algorithm.) The point in time in each graph where the otherwise continuous regression line is broken is the frame at which the 'optimal' state transition occurs. We note that the shapes of the data sequences of these two tokens are largely the same except the initial portion of the data in token one (left graph) is nearly twice as fast as that in token two (right graph). (But token one slows down during the mid portion.) With no time-warping mechanism built into the regression function, a single set of polynomial coefficients trained from the two tokens having varying state-durations are not able to fit closely to both the tokens. The polynomial coefficients were trained in such a way that the fitting accuracy is compromised between the two tokens.

In Fig. 2, we show the results of fitting the same two C1 data sequences as in Fig. 1 but using the new trended HMM containing time-warping parameters. Again, the same two tokens were used to train the model according to the algorithm described in Section 3. In contrast to the results in Fig. 1, with use of the new model, the fitting accuracy is high for both of the two tokens despite their varying state durations. This simultaneous high accuracy is achieved through use of different values of the auxiliary time-warping parameters for the two tokens.

As another example, Figs. 3 and 4 are analogous to Figs. 1 and 2 except for use of C3 data sequences of two tokens from the TIMIT database. Again, use of time-warping parameters in the trended HMM produces more accurate fitting to the two data tokens simultaneously than without use of them.

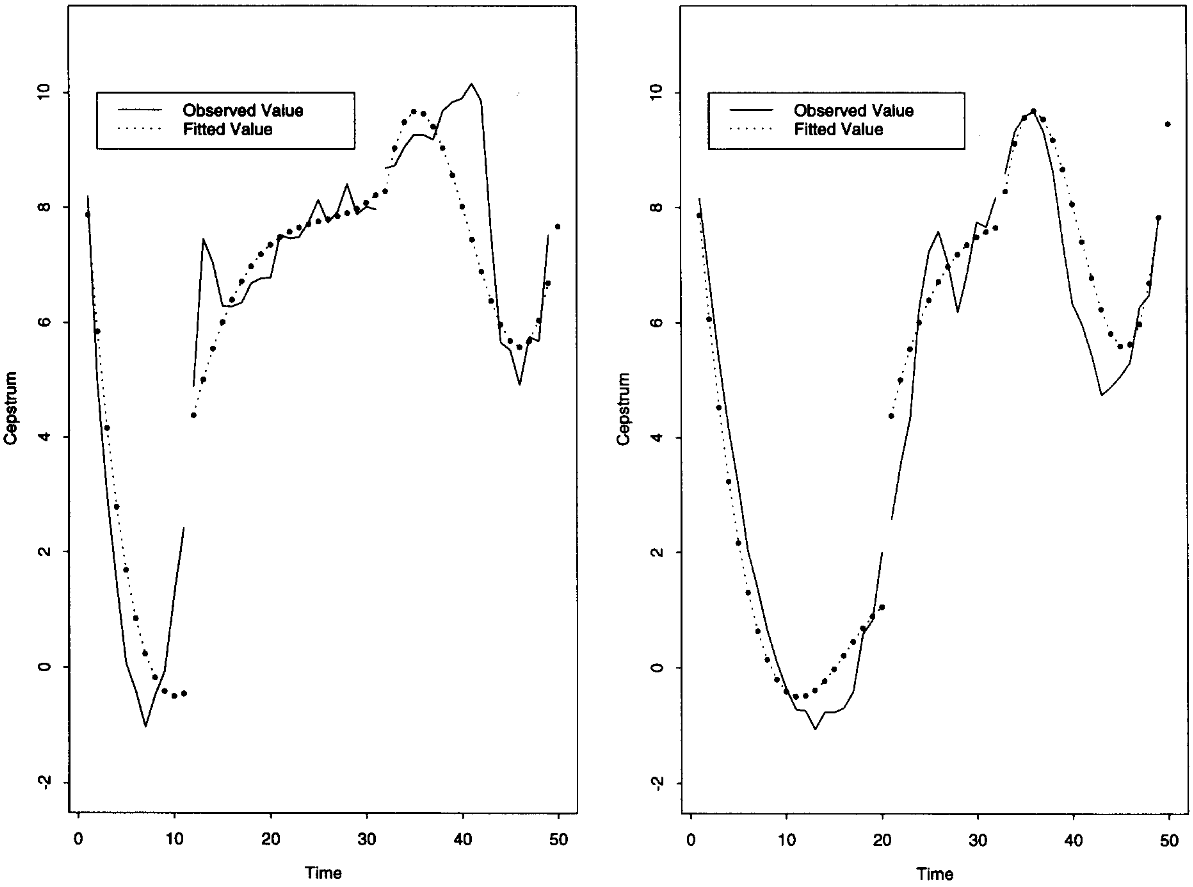

Fig. 1. The solid lines in the two graphs are C1 data sequences from two tokens, respectively, uttered by the same speaker from the same word *peek*. Vertical axis is the magnitude of C1; horizontal axis is the frame number. The dotted lines superimposed on the graphs are three sequentially arranged state-dependent polynomial regression functions with the polynomial coefficients optimally trained from the two data tokens and with no time warping incorporated (i.e. using the trended HMM in [4].) The point in time in each graph where the otherwise continuous regression line is broken is the frame at which optimal state transitions occur.

## 5. Speech recognition experiments

We choose the standard acoustic–phonetic TIMIT database for the evaluation experiments in this paper. The results on phonetic recognition are presented using time-warping trended HMMs in comparison with the regular trended HMMs. The advantage of the trended HMMs over the standard HMMs have been demonstrated in [5] and is not the focus of the recognition experiments in this paper.

Since trended HMMs are mostly effective for long-span, continuously varying patterns, we only consider vowel recognitions in the experiments. For consonant segments, it is not beneficial to employ the trended HMMs due to the extremely dynamic nature of the acoustic features. After some exploratory study, we notice that there is a large amount of variations among the tokens of the same vowel in the TIMIT database due to heavy co-articulation and varying speakers. The advantage of modeling accuracy using trended HMMs tends to be cancelled out by the merge of tokens with large variations, in particular the contextual variations. Therefore, we train multiple models for each vowel according to its left and right contexts. In the

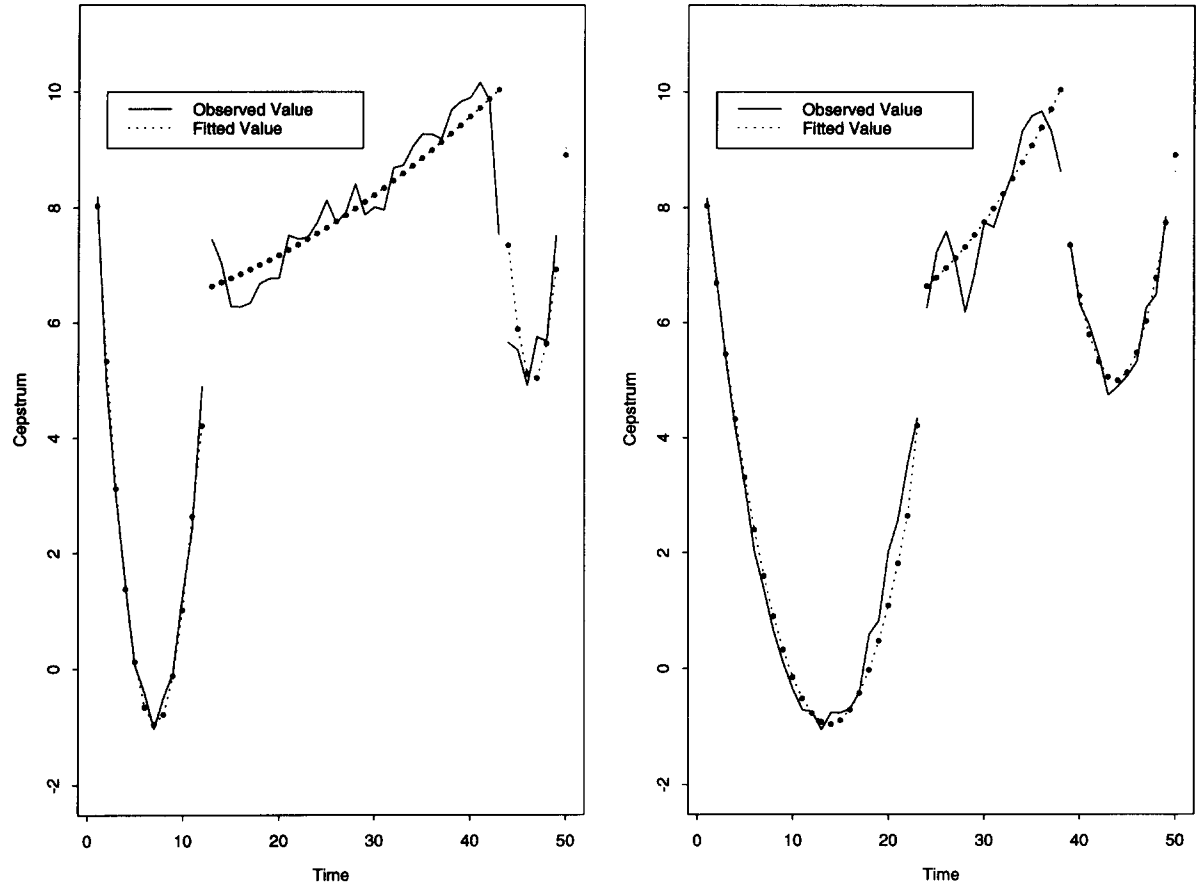

Fig. 2. The same data fitting as in Fig. 1 but using the trended HMM with the state-dependent time-warping mechanism incorporated.

experiments, we select ten vowels: /aa/, /ao/, /aw/, /ay/, /eh/, /er/, /iy/, /ow/, /oy/, /uw/, and for each vowel, we select the tokens in various CVC contexts, where the initial and the final consonants are selected from the stop consonants: /b/, /d/, /g/, /p/, /t/, /k/. The acoustic features for model estimation are only based on the vowel segments. In other words, the acoustic features of the consonant segments are not used.

The recognition experiments involve two types of recognition systems based on the regular trended HMMs and the time-warping trended HMMs. For each recognition system, we vary the degree of the polynomials in the trended HMMs and the number of states in the HMMs. For the ten vowels in CVC context, we created 126 HMMs for each recognition experiment (about 5 to 20 models per vowel). Each experiment is based on 351 training tokens selected from a set of 120 speakers in the training set and 518 test tokens selected from the 168 speakers in the test set. Table 1 lists the vowel recognition results of the recognizers.

These results demonstrated the advantages of the time-warping trended HMMs over the regular trended HMMs (10 to 15% differences in recognition rates). The major reason is that both training and test tokens in the TIMIT database are highly variable in their durations. While the regular trended HMMs are not capable of eliminating such variations, the time-warping trended HMMs effectively avoid the duration variations by 'on-line' estimation of auxiliary time-warping parameters. Based

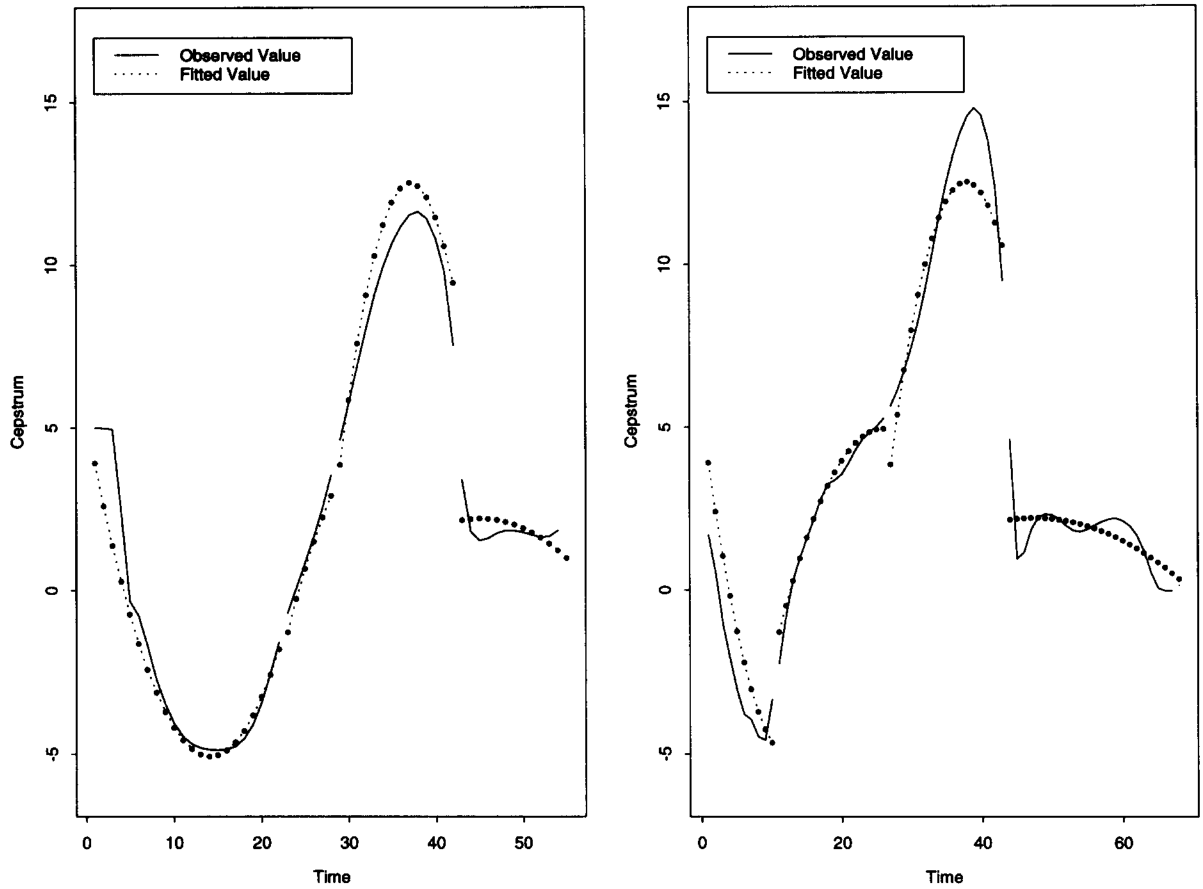

Fig. 3. Fitting the trended HMM with no time-warping mechanism to C3 data sequences from two tokens of word *bike* excised from the TIMIT database.

on the results of experiments, we also conclude that the number of states in both regular trended HMMs and time-warping trended HMMs does not appear to play an important role for improving the performance of the recognizers.

## 6. Summary and conclusion

In this study we propose an improved version of the nonstationary state, trended HMM published in [4] in its provision of the flexibility of time scaling in individual HMM states. We identify the need for incorporating this flexibility in the new model: Since multiple speech-data sequences (after state segmentation) used for training a single set of state-dependent polynomial coefficients usually do not have the same sequence length, the resulting polynomial trend function cannot simultaneously fit all these data sequences well. (Interestingly, this problem did not exist for the conventional stationary-state HMM, where all the state-dependent 'trend functions' are constant over time.)

After we addressed the importance of incorporating time warping in the trended HMM states, we provide an effective solution to this problem. The solution is based on use of token-dependent time scaling parameters, which we call auxiliary parameters to distinguish them from the intrinsic model parameters such as the polynomial coefficients in

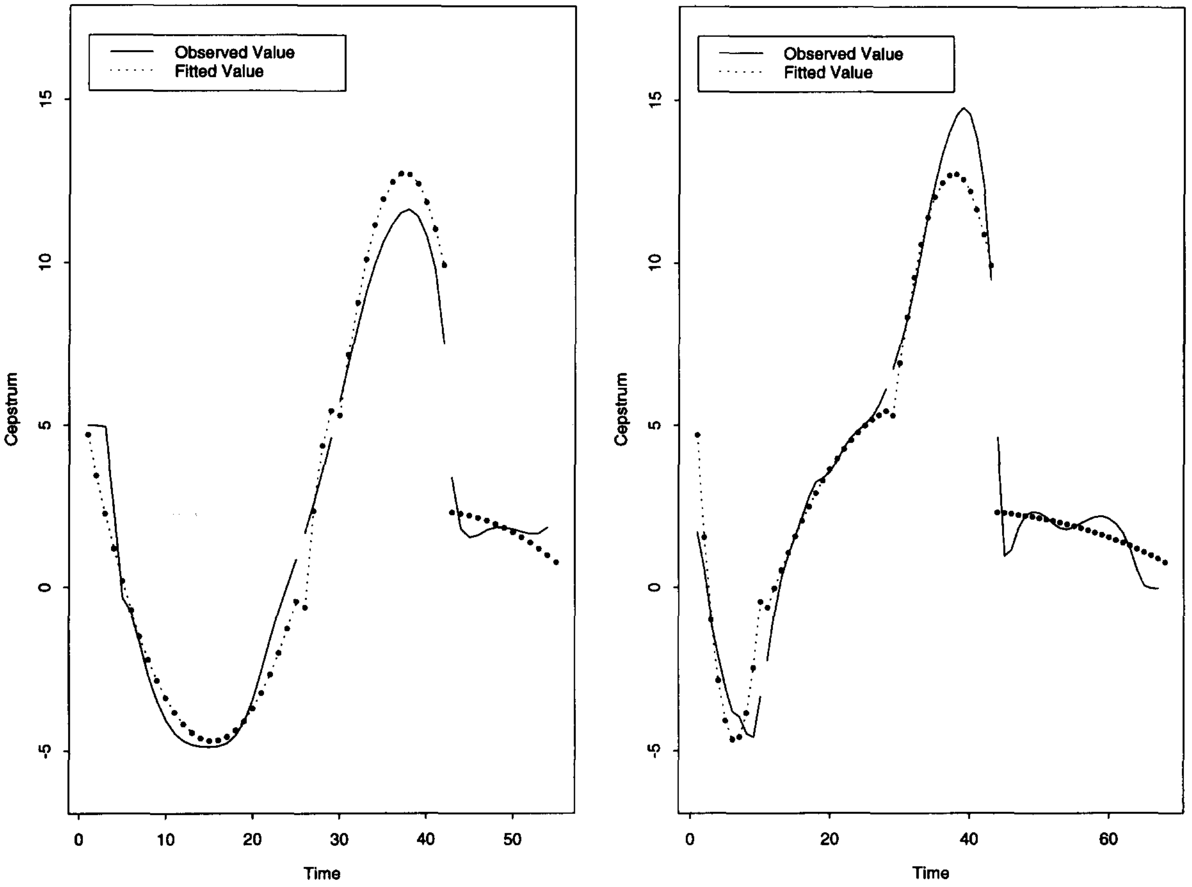

Fig. 4. The same data fitting as in Fig. 3 but using the trended HMM with the state-dependent time-warping mechanism incorporated.

Table 1
Speech recognition rates based on regular and time-warping trended HMMs

| No. of states | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Regular trend HMM (order 1) | 60.4 | 59.7 | 62.0 | 62.4 | 64.3 |
| Regular trend HMM (order 2) | 64.5 | 64.5 | 65.1 | 65.3 | 66.0 |
| Time-warping trend HMM (order 1) | 75.5 | 75.1 | 77.8 | 73.4 | 74.1 |
| Time-warping trend HMM (order 2) | 76.1 | 75.7 | 76.6 | 74.5 | 74.1 |

the trend functions. An efficient, iterative two-stage optimization algorithm is developed to accomplish maximum likelihood estimation of all model parameters, including both the auxiliary and intrinsic ones. Convergence of the algorithm is proved under most general conditions, so is its convergence to (state-optimized) maximum likelihood estimates.

Numerical experiments are designed to fit a single state-dependent trend function in the trended HMM simultaneously to multiple sequences of speech data which are different renditions of the same word yet vary in the sequence length. When no time-warping mechanism is built into the trend function, we demonstrate that a single set of (state-dependent) polynomial coefficients are not able to fit multiple tokens possessing different token lengths. The polynomial coefficients were trained such that all the tokens are fitted moderately well but none of them is closely fitted. In contrast, using the new model developed in this paper and applying the algorithm described in Section 3, the fitting accuracy becomes much higher for all the tokens regardless of their varying state durations. Such simultaneous high accuracy is achieved through use of different values of the auxiliary time-warping parameters for different tokens, which are automatically determined by the training algorithm, for the different tokens.

Incorporation of the mechanism for state-dependent time warping described in this paper is a necessary step for accurate speech recognition to be pursued in our future work. The warping automatically normalizes speaking rate variation from one speech token to another given the same phonetic or subphonetic content for the HMM state. Otherwise, this speaking rate variation would introduce unnecessary variability in the estimates for the state-dependent trend function's parameters and hence increase overlap (confusion) among different classes of speech sounds.

## Acknowledgments

We wish to thank Mr. M. Aksmanovic for recording various utterances with intentionally altered speaking rates. The support for this work is provided by Natural Sciences and Engineering Research Council of Canada and by University of Waterloo Interdisciplinary Grants Program.

## 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] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28, No. 4, 1980, pp. 357–365.

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

[4] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal", *Signal Processing*, Vol. 27, No. 1, April 1992, pp. 65–78.

[5] L. Deng, M. Aksmanovic, D.X. Sun and C.F.J. Wu, "Speech recognition and using hidden Markov models with polynomial regression functions as nonstationary states", *IEEE Trans. Speech Audio Process.*, October 1994, to appear.

[6] L. Deng, M. Lennig and P. Mermelstein, "Use of vowel duration information in a large vocabulary word recognizer", *J. Acoust. Soc. Amer.*, Vol. 86, August 1989, pp. 540–548.

[7] O. Ghitza and M. Sondhi, "Hidden Markov models with templates as nonstationary states: An application to speech recognition", *Comput. Speech Language*, Vol. 7, No. 2, 1993, pp. 101–119.

[8] B.H. Juang and L.R. Rabiner, "The segmental *k*-means algorithm for estimating parameters of hidden Markov models", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 38, 1990, pp. 1639–1641.

[9] I. Lehiste, *Readings in Acoustic Phonetics*, MIT Press, Cambridge, MA, 1967.

[10] L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources", *IEEE Trans. Inform. Theory*, Vol. 28, 1982, pp. 729–734.

[11] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, February 1989, pp. 257–285.

[12] V.W. Zue, *Speech Spectrogram Reading: An Acoustic Study of the English Language*, Lecture Notes, MIT, Cambridge, MA, August 1991.