Computer Speech and Language (1998) **12,** 229–246
Article No. la980047

Model parameter estimation for mixture density polynomial segment models

T. Fukada, K. K. Paliwal and Y. Sagisaka

ATR Interpreting Telecommunications Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Abstract

In this paper, we propose parameter estimation techniques for mixture density polynomial segment models (MDPSMs) where their trajectories are specified with an arbitrary regression order. MDPSM parameters can be trained in one of three different ways: (1) segment clustering; (2) expectation maximization (EM) training of mean trajectories; and (3) EM training of mean and variance trajectories. These parameter estimation methods were evaluated in TIMIT vowel classification experiments. The experimental results showed that modelling both the mean and variance trajectories is consistently superior to modelling only the mean trajectory. We also found that modelling both trajectories results in significant improvements over the conventional HMM.

1. Introduction

To date, one of the most successful approaches for large vocabulary continuous speech recognition has been based on the hidden Markov model (HMM). Although HMMs will continue to play an important role in most recognition systems for a long time to come, many alternative models have been proposed in recent years that enable some of the shortcomings of HMMs to be addressed. Broadly speaking, there are two HMM limitations that various models have tried to address: (1) weak duration modelling and (2) assumption of the conditional independence of observations given the state sequence. The first problem, where an HMM state duration model is implicitly given by a geometric distribution, has been addressed by introducing semi-Markov models with explicit state duration distributions. The second problem has been widely acknowledged to be more serious, and a number of alternative solutions that address this problem have been studied (Ostendorf & Roukos, 1989; Deng, 1992; Gales & Young, 1993; Ghitza & Sondhi, 1993; Gish & Ng, 1993; Paliwal, 1993; Goldenthal & Glass, 1994; Gong & Haton, 1994; Robinson, Hochberg & Renals, 1994; Holmes & Russell, 1995). Delta parameters offer the simplest way of representing the time dependency of observations, and have been shown to tremendously boost performance. Other alternatives are more elegant in representing the time dependency. The polynomial

0885–2308/98/030229 + 18 $30.00/0

segment modelling, proposed by Gish and Ng (1993) is one such technique for relaxing the independence assumption. This modelling technique, however, has a serious shortcoming; it assumes the variance to be time invariant within a segment. This will be disadvantageous with respect to the conventional HMMs which can represent variance changes in a segment by dividing the segment into a number of states with different variances.

In this paper, we consider the case where both mean and covariance are varying with time. We present a model parameter estimation method for mixture density polynomial segment models (MDPSMs) to deal with this type of time-varying case. The model parameters of the MDPSM are the mean trajectory coefficients, the covariance coefficients and the mixture weights. In our segmental modelling approach, higher order regression models are used not only for mean trajectory modelling but also for time-varying covariance modelling. The paper proposed by Gish and Ng (1993) can be viewed as a special case (i.e. 0-th order regression for modelling the covariance coefficients) if our method is considered. Recently, a similar approach was also proposed by Gish and Ng (1996). However, they restricted the time variation of the covariance coefficients to be limited to having three different covariance matrices existing over a segment, while there is no restriction of this type in our modelling.

The paper is organized as follows. Section 2 starts with an overview of single Gaussian segment modelling, goes on to describe two methods of model parameter estimation for MDPSMs with time-invariant variance, and finally provides model parameter estimation formulation for the time-variant variance case.[1] To confirm the performance of the three kinds of MDPSM, preliminary classification experiments are performed. These are described in Section 3. Section 4 concludes the paper.

2. Derivation of model parameter estimation formulas

2.1. Polynomial segment model (Gish & Ng, 1993)

Consider an L (in frames) length sequence of observation vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_L\}$, where $\mathbf{y}_t = [y_{t,1}, y_{t,2}, \ldots, y_{t,D}]$ is a D-dimensional observation (e.g. cepstrum) vector at time t. This sequence defines a segment which can be expressed in the form of an $L \times D$ matrix

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,D} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,D} \\ \vdots & \vdots & & \vdots \\ y_{L,1} & y_{L,2} & \cdots & y_{L,D} \end{bmatrix}. \tag{1}$$

In the polynomial segment model, this segment is represented by an R-th order trajectory model as follows:

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E}, \tag{2}$$

where $\mathbf{Z}$ is an $L \times (R+1)$ design matrix defined by

[1] In this paper, we provide the MDPSM formulation for diagonal covariance matrices only. However, it can be easily extended to the full covariance case.

$$\mathbf{Z}=\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & \frac{1}{L-1} & \dots & \left(\frac{1}{L-1}\right)^R \\ \vdots & \vdots & & \vdots \\ 1 & \frac{t-1}{L-1} & \dots & \left(\frac{t-1}{L-1}\right)^R \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}, \tag{3}$$

B is an $(R+1)\times D$ trajectory parameter matrix

$$\mathbf{B}=\begin{bmatrix} b_{0,1} & b_{0,2} & \dots & b_{0,D} \\ b_{1,1} & b_{1,2} & \dots & b_{1,D} \\ \vdots & \vdots & & \vdots \\ b_{R,1} & b_{R,2} & \dots & b_{R,D} \end{bmatrix}, \tag{4}$$

and **E** is an $L\times D$ residual error matrix

$$\mathbf{E}=\begin{bmatrix} e_{1,1} & e_{1,2} & \dots & e_{1,D} \\ e_{2,1} & e_{2,2} & \dots & e_{2,D} \\ \vdots & \vdots & & \vdots \\ e_{L,1} & e_{L,2} & \dots & e_{L,D} \end{bmatrix}. \tag{5}$$

Design matrix **Z** deals with normalizing different length of segments uniformly between times 0 and 1.

2.2. Single Gaussian segment model (Gish & Ng, 1993)

The likelihood of the segment **Y**, given that it is generated by a label a, can be expressed as

$$\mathrm{P}(\mathbf{Y}|a)=\prod_{t=1}^{L} f(\mathbf{y}_t). \tag{6}$$

In this equation, $f(\mathbf{y}_t)$ is the likelihood of the feature vector $\mathbf{y}_t$ conditioned on the label a and is given by

$$f(\mathbf{y}_t)=\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_a|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\mathbf{z}_t\mathbf{B}_a)^T\mathbf{\Sigma}_a^{-1}(\mathbf{y}_t-\mathbf{z}_t\mathbf{B}_a)\right\}, \tag{7}$$

where $\mathbf{B}_a$ and $\mathbf{\Sigma}_a$ are the parameters of the single Gaussian segment model describing the label a. In Equation (7), the vector $\mathbf{z}_t$ is given by

$$\mathbf{z}_t = \left[1, \frac{t-1}{L-1}, \ldots, \left(\frac{t-1}{L-1}\right)^R\right]. \tag{8}$$

Assuming that we are given K segments $\mathbf{Y}_1$, $\mathbf{Y}_2$, ..., $\mathbf{Y}_K$ in the training data for label a, we want to compute model parameters $\mathbf{B}_a$ and $\mathbf{\Sigma}_a$ for the single Gaussian segment model. Probability of these segments given $\mathbf{B}_a$ and $\mathbf{\Sigma}_a$ is given as

$$\begin{aligned} \mathrm{P}(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K | \mathbf{B}_a, \mathbf{\Sigma}_a) &= \prod_{k=1}^{K} \mathrm{P}(\mathbf{Y}_k | \mathbf{B}_a, \mathbf{\Sigma}_a) \\ &= \prod_{k=1}^{K} \prod_{t=1}^{L} f(\mathbf{y}_{k,t}). \end{aligned} \tag{9}$$

These model parameters can be obtained by maximizing this probability with respect to $\mathbf{B}_a$ and $\mathbf{\Sigma}_a$. Their estimates can be computed as follows:

$$\bar{\mathbf{B}}_a = \left[\sum_{k=1}^{K} \mathbf{Z}_k^T \mathbf{Z}_k\right]^{-1} \left[\sum_{k=1}^{K} \mathbf{Z}_k^T \mathbf{Y}_k\right], \tag{10}$$

$$\bar{\mathbf{\Sigma}}_a = \frac{\sum_{k=1}^{K} (\mathbf{Y}_k - \mathbf{Z}_k \bar{\mathbf{B}}_a)^T (\mathbf{Y}_k - \mathbf{Z}_k \bar{\mathbf{B}}_a)}{\sum_{k=1}^{K} L_k}. \tag{11}$$

From now on, we omit the subscript a from the model parameters $\mathbf{B}_a$ and $\mathbf{\Sigma}_a$ for simplification.

2.3. *Mixture density polynomial segment model (Gish & Ng, 1993)*

The discussion in the previous section was concerned with single Gaussian segment modelling. In this section, we extend it to a mixture density case. In this case, the likelihood $f(\mathbf{y}_t)$ is represented by a mixture of M Gaussians; i.e.

$$f(\mathbf{y}_t) = \sum_{m=1}^{M} w_m f_m(\mathbf{y}_t), \tag{12}$$

where

$$f_m(\mathbf{y}_t)=\frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_m|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\mathbf{z}_t\mathbf{B}_m)^T\mathbf{\Sigma}_m^{-1}(\mathbf{y}_t-\mathbf{z}_t\mathbf{B}_m)\right\}, \tag{13}$$

and w_m is the weight of the m-th mixture component. The mixture components satisfy the relation $\Sigma_{m=1}^{M} w_m=1$. The model parameters $\mathbf{B}_m$, $\mathbf{\Sigma}_m$, and w_m in Equation (12) can be estimated by segment clustering or by EM training. These methods are described in detail in subsections 2.3.1 and 2.3.2, respectively. These methods are developed here under the assumption that the covariance matrices $\{\mathbf{\Sigma}_m,\ m=1,\ 2,\ \ldots,\ M\}$ for the M mixture components are diagonal; i.e.

$$\begin{aligned}\mathbf{\Sigma}_m&=\mathrm{diag}[c_{m,1},\ c_{m,2},\ \ldots,\ c_{m,D}]\\&=\begin{bmatrix}c_{m,1} & 0 & \ldots & 0\\ 0 & c_{m,2} & \ldots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \ldots & c_{m,D}\end{bmatrix}.\end{aligned} \tag{14}$$

Under this assumption, Equation (13) becomes

$$f_m(\mathbf{y}_t)=\frac{1}{(2\pi)^{D/2}(\Pi_{d=1}^{D} c_{m,d})^{1/2}}\exp\left\{-\sum_{d=1}^{D}\frac{(y_{t,d}-\mathbf{z}_t\mathbf{b}_{m,d})^2}{2c_{m,d}}\right\}, \tag{15}$$

where $\bar{\mathbf{b}}_{m,d}=[\bar{b}_{m,0,d},\ \bar{b}_{m,1,d},\ \ldots,\ \bar{b}_{m,R,d}]^T$ is the d-th dimensional mean trajectory parameters of the m-th mixture. We describe below different methods for estimating the model parameters $\mathbf{B}_m$, $\mathbf{\Sigma}_m$, and w_m under the assumption that $\mathbf{\Sigma}_m$ is diagonal. However, these methods can be easily extended to the full-covariance case.

2.3.1. Clustering method

One simple way of estimating MDPSM parameters $\mathbf{B}_m$, $\mathbf{\Sigma}_m$, and w_m is based on segment clustering. That is, the training segments $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\ldots$, $\mathbf{Y}_K$ for the label a are partitioned into M regions using the K-means clustering algorithm. The K-means clustering algorithm requires a definition of distance between a segment $\mathbf{Y}$ and cluster m. The distance measure used here during the clustering is a "multivariate Gaussian distance measure":

$$\begin{aligned}\mathrm{Dist}(\mathbf{Y},m)&=\frac{1}{2}LD\log 2\pi+\frac{1}{2}L\sum_{d=1}^{D}\log c_{m,d}\\&\quad+\frac{1}{2}\sum_{t=1}^{L}\sum_{d=1}^{D}\frac{(y_{t,d}-\mathbf{z}_t\mathbf{b}_{m,d})^2}{c_{m,d}}.\end{aligned} \tag{16}$$

Estimates of $\mathbf{B}_m$ and $\mathbf{\Sigma}_m$ can be obtained in the same way as in the single mixture case using the segments assigned to cluster m. w_m is calculated as the relative frequency of the segments:

$$w_m = \frac{N_m}{\sum_{j=1}^{M} N_j}, \quad m = 1, \ldots, M, \tag{17}$$

where N_m is the number of segments for cluster m.

2.3.2. EM method

Let $\mathbf{Y}_1^K = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K\}$ be a set of training segments belonging to the label a. Our aim here is to estimate the model parameters $\mathbf{B}_m$, $\mathbf{\Sigma}_m$, and w_m from these training segments using the expectation–maximization (EM) algorithm. For this, we derive re-estimation formulas by maximizing $\mathrm{P}(\mathbf{Y}_1^K|a)$ based on the EM algorithm. This can be done by maximizing the following auxiliary function Q with the mixture components being the hidden variables:

$$\begin{aligned} Q(\bar{\Phi}|\Phi) &= E[\log \mathrm{P}(\mathbf{Y}_1^K, m|\bar{\Phi})|\mathbf{Y}_1^K, \Phi] \\ &= \sum_{m=1}^{M} \frac{\mathrm{P}(\mathbf{Y}_1^K, m|\Phi)}{\mathrm{P}(\mathbf{Y}_1^K|\Phi)} \log \mathrm{P}(\mathbf{Y}_1^K, m|\bar{\Phi}), \end{aligned} \tag{18}$$

where Φ and $\bar{\Phi}$ are the sets of the current model parameters and the re-estimated model parameters, respectively. m denotes the index of a mixture component. Since $\log \mathrm{P}(\mathbf{Y}_1^K, m|\bar{\Phi})$ in Equation (18) can be rewritten as

$$\log \mathrm{P}(\mathbf{Y}_1^K, m|\bar{\Phi}) = \log \mathrm{P}(\mathbf{Y}_1^K|m, \bar{\Phi}) + \log \mathrm{P}(m|\bar{\Phi}), \tag{19}$$

maximizing Equation (18) is equivalent to individually maximizing the following two functions:

$$Q_1(\bar{\Phi}|\Phi) = \sum_{m=1}^{M} \frac{\mathrm{P}(\mathbf{Y}_1^K, m|\Phi)}{\mathrm{P}(\mathbf{Y}_1^K|\Phi)} \log \mathrm{P}(\mathbf{Y}_1^K|m, \bar{\Phi}), \tag{20}$$

with respect to $\mathbf{B}_m$ and $\mathbf{\Sigma}_m$, and

$$Q_2(\bar{\Phi}|\Phi) = \sum_{m=1}^{M} \frac{\mathrm{P}(\mathbf{Y}_1^K, m|\Phi)}{\mathrm{P}(\mathbf{Y}_1^K|\Phi)} \log \mathrm{P}(m|\bar{\Phi}), \tag{21}$$

with respect to w_m.

Let the probability $\mathrm{P}(\mathbf{Y}_1^K, m|\Phi)/\mathrm{P}(\mathbf{Y}_1^K|\Phi)$ in Equation (20) and Equation (21) be denoted as $\Sigma_{k=1}^{K} \Sigma_{t=1}^{L_k} \gamma_{k,m,t}$ using the current model parameters Φ. Then, we can estimate $\gamma_{k,m,t}$ efficiently as

$$\gamma_{k,m,t}=\begin{cases}\dfrac{\alpha_{k,t}\beta_{k,t+1}w_m f_m(\mathbf{y}_{k,t+1})}{\mathrm{P}(\mathbf{Y}_1^K|\Phi)}, & t=1,\ldots,L_k-1,\\ \dfrac{\alpha_{k,T}}{\mathrm{P}(\mathbf{Y}_1^K|\Phi)}, & t=L_k,\end{cases} \tag{22}$$

where $\alpha_{k,t}$ and $\beta_{k,t}$ are obtained recursively:

$$\alpha_{k,t}=\begin{cases}f(\mathbf{y}_{k,1}), & t=1\\ \alpha_{k,t-1}f(\mathbf{y}_{k,t}), & t=2,\ldots,L_k,\end{cases} \tag{23}$$

$$\beta_{k,t}=\begin{cases}1, & t=L_k\\ \beta_{k,t+1}f(\mathbf{y}_{k,t+1}), & t=L_k-1,\ldots,1.\end{cases} \tag{24}$$

First, we consider obtaining the d-th dimensional mean trajectory parameters of the m-th mixture, $\bar{\mathbf{b}}_{m,d}=[\bar{b}_{m,0,d},\bar{b}_{m,1,d},\ldots,\bar{b}_{m,R,d}]^T$. These parameters can be obtained through differentiation of Equation (20) with respect to $\bar{b}_{m,r,d}$ and solving the equation:

$$\frac{\partial Q_1}{\partial \bar{b}_{m,r,d}}=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\frac{\partial \log \bar{f}_m(\mathbf{y}_{k,t})}{\partial \bar{b}_{m,r,d}}=0,\quad r=0,\ldots,R. \tag{25}$$

From Equation (15),

$$\frac{\partial \log \bar{f}_m(\mathbf{y}_{k,t})}{\partial \bar{b}_{m,r,d}}=\frac{(y_{k,t,d}-\mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})}{\bar{c}_{m,d}}\left(\frac{t-1}{L_k-1}\right)^r. \tag{26}$$

Substituting this equation into Equation (25) and noting that $\bar{c}_{m,d}$ is an independent constant of time t, we obtain

$$\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\left\{y_{k,t,d}-\sum_{u=0}^{R}\bar{b}_{m,u,d}\left(\frac{t-1}{L_k-1}\right)^u\right\}\left(\frac{t-1}{L_k-1}\right)^r=0,\quad r=0,\ldots,R. \tag{27}$$

Equation (27) can be rewritten as

$$\sum_{u=0}^{R}g_m(u+r)\bar{b}_{m,u,d}=v_{m,d}(r),\quad r=0,\ldots,R, \tag{28}$$

where

$$g_m(l)=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\left(\frac{t-1}{L_k-1}\right)^l,\quad l=0,\ldots,2R, \tag{29}$$

and

$$v_{m,d}(r)=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}y_{k,t,d}\left(\frac{t-1}{L_k-1}\right)^r,\quad r=0,\ldots,R. \tag{30}$$

Note that $g_m(l)$ is independent of the dimension d of the feature parameter and $v_{m,d}(r)$ is dependent of dimension d. The $(R+1)$ parameters, $\{\bar{b}_{m,u,d}, u=0,\ldots,R\}$, can be obtained by solving the set of $(R+1)$ simultaneous linear equations given by Equation (28). These equations can be written in a matrix form as follows:

$$\begin{bmatrix} g_m(0) & g_m(1) & \cdots & g_m(R) \\ g_m(1) & g_m(2) & \cdots & g_m(R+1) \\ \vdots & \vdots & \ddots & \vdots \\ g_m(R) & g_m(R+1) & \cdots & g_m(2R) \end{bmatrix}\begin{bmatrix} \bar{b}_{m,0,d} \\ \bar{b}_{m,1,d} \\ \vdots \\ \bar{b}_{m,R,d} \end{bmatrix}=\begin{bmatrix} v_{m,d}(0) \\ v_{m,d}(1) \\ \vdots \\ v_{m,d}(R) \end{bmatrix}. \tag{31}$$

In order to compute the d-th diagonal component of the covariance matrix $\bar{\boldsymbol{\Sigma}}_m$, we differentiate Equation (20) with respect to the $\bar{c}_{m,d}$ and solve the following equation:

$$\frac{\partial Q_1}{\partial \bar{c}_{m,d}}=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\frac{\partial \log \bar{f}_m(\mathbf{y}_{k,t})}{\partial \bar{c}_{m,d}}=0. \tag{32}$$

From Equation (15), we can derive

$$\frac{\partial \log \bar{f}_m(\mathbf{y}_{k,t})}{\partial \bar{c}_{m,d}}=-\frac{1}{2\bar{c}_{m,d}}+\frac{(y_{k,t,d}-\mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^2}{2\bar{c}_{m,d}^2}. \tag{33}$$

Using Equation (33), Equation (32) can be rewritten as

$$\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\{\bar{c}_{m,d}-(y_{k,t,d}-\mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^2\}=0. \tag{34}$$

Then, $\bar{c}_{m,d}$ can be obtained by

$$\bar{c}_{m,d}=\frac{\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}(y_{k,t,d}-\mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^2}{\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}}. \tag{35}$$

The weighting coefficient $\bar{w}_m$ can be obtained from Equation (21) by application of a Lagrange optimization using Lagrange multipliers:

$$\bar{w}_m = \frac{\sum_{k=1}^{K} \sum_{t=1}^{L_k} \gamma_{k,m,t}}{\sum_{k=1}^{K} \sum_{t=1}^{L_k} \sum_{j=1}^{M} \gamma_{k,j,t}}. \tag{36}$$

The results of the model parameters obtained from the clustering method described in subsection 2.3.1 can be used as initial model parameters for the EM algorithm.

2.4. Variance trajectory model

In the previous Subsection, we have discussed the mixture density polynomial segment model (MDPSM), where the mean feature vector of the m-th mixture is varying with time, but the covariance matrix $\mathbf{\Sigma}_m$ remains time invariant. In this subsection, we modify the segment model so that both the mean vector and the covariance matrix can vary with time within a segment. We believe that this will allow more precise modelling. In order to characterize the time-varying behaviour of the mean vector and the covariance matrix, we represent their trajectories by a polynomial segment model. In the case of the covariance matrix $\mathbf{\Sigma}_{m,t} = \text{diag}[c_{m,t,1}, \ldots, c_{m,t,D}]$, this is represented as an R-th order polynomial:

$$\mathbf{c}_{m,t} = \mathbf{z}_t \mathbf{S}_m, \tag{37}$$

where

$$\mathbf{c}_{m,t} = [\mathbf{c}_{m,t,1}, \ldots, \mathbf{c}_{m,t,D}], \tag{38}$$

$$\mathbf{z}_t = \left[1, \frac{t-1}{L-1}, \ldots, \left(\frac{t-1}{L-1}\right)^R\right], \tag{39}$$

and

$$\mathbf{S}_m = \begin{bmatrix} s_{m,0,1} & s_{m,0,2} & \cdots & s_{m,0,D} \\ s_{m,1,1} & s_{m,1,2} & \cdots & s_{m,1,D} \\ \vdots & \vdots & & \vdots \\ s_{m,R,1} & s_{m,R,2} & \cdots & s_{m,R,D} \end{bmatrix}. \tag{40}$$

Note that we have represented the trajectories of the mean vector and the covariance matrix by the polynomial segment models of the same order R, though they can, in principle, be different. With this model, the likelihood of the vector $\mathbf{y}_t$ is given by

$$f_m(\mathbf{y}_t) = \frac{1}{(2\pi)^{D/2}(\Pi_{d=1}^{D} c_{m,t,d})^{1/2}} \exp\left\{ -\sum_{d=1}^{D} \frac{(y_{t,d} - \mathbf{z}_t \mathbf{b}_{m,d})^2}{2c_{m,t,d}} \right\}. \tag{41}$$

In this model, estimates of the mean trajectory and weight parameters can be obtained in a way similar to that described in Subsection 2.3.2, except that $\bar{c}_{m,d}$ in Equation (26) is replaced by $\Sigma_{n=0}^{R} s_{m,n,d}(t-1/L_k-1)^n$ to reflect its time dependence. The computation of the mean trajectory and weight parameters can be obtained from Equation (31) and Equation (36), respectively. Note that $g_m(l)$ and $v_{m,d}(r)$ in Equation (31) and $\gamma_{k,m,t}$ in Equation (36) are different from the time-invariant case, because the likelihood is computed by Equation (41) instead of Equation (15). The computation of variance differs as follows. The ML estimates of the d-th diagonal component of the covariance matrix $\bar{\mathbf{\Sigma}}_m$ can be obtained through differentiation of Equation (20) with respect to $\bar{s}_{m,r,d}$ and solving the equation:

$$\frac{\partial Q_1}{\partial \bar{s}_{m,r,d}} = \sum_{k=1}^{K} \sum_{t=1}^{L_k} \gamma_{k,m,t} \frac{\log \bar{f}_m(\mathbf{y}_{k,t})}{\partial \bar{s}_{m,r,d}} = 0. \tag{42}$$

It gives

$$\sum_{k=1}^{K} \sum_{t=1}^{L_k} \gamma_{k,m,t} \frac{\left(\frac{t-1}{L_k-1}\right)^r}{\left\{\sum_{n=0}^{R} \bar{s}_{m,n,d}\left(\frac{t-1}{L_k-1}\right)^n\right\}^2} \left\{ \sum_{u=0}^{R} \bar{s}_{m,u,d}\left(\frac{t-1}{L_k-1}\right)^u - (y_{k,t,d} - \mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^2 \right\} = 0. \tag{43}$$

This is a non-linear equation in $\bar{s}_{m,n,d}$. In order to make it linear, we use an approximation assuming $\bar{s}_{m,n,d}$ in the denominator is replaced by the current value, $s_{m,n,d}$, as

$$\sum_{k=1}^{K} \sum_{t=1}^{L_k} \gamma_{k,m,t} \frac{\left(\frac{t-1}{L_k-1}\right)^r}{\left\{\sum_{n=0}^{R} s_{m,n,d}\left(\frac{t-1}{L_k-1}\right)^n\right\}^2} \left\{ \sum_{u=0}^{R} \bar{s}_{m,u,d}\left(\frac{t-1}{L_k-1}\right)^u - (y_{k,t,d} - \mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^2 \right\} = 0. \tag{44}$$

Now equation (44) can be rewritten as

$$\sum_{u=0}^{R} h_{m,d}(u+r)\bar{s}_{m,u,d} = x_{m,d}(r), \quad r = 0, \ldots, R, \tag{45}$$

where

$$h_{m,d}(l)=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\frac{\left(\dfrac{t-1}{L_k-1}\right)^{l}}{\left\{\displaystyle\sum_{n=0}^{R}s_{m,n,d}\left(\frac{t-1}{L_k-1}\right)^{n}\right\}^{2}},\quad l=0,\ldots,2R, \tag{46}$$

and

$$x_{m,d}(r)=\sum_{k=1}^{K}\sum_{t=1}^{L_k}\gamma_{k,m,t}\frac{\left(\dfrac{t-1}{L_k-1}\right)^{r}(y_{k,t,d}-\mathbf{z}_{k,t}\bar{\mathbf{b}}_{m,d})^{2}}{\left\{\displaystyle\sum_{n=0}^{R}s_{m,n,d}\left(\frac{t-1}{L_k-1}\right)^{n}\right\}^{2}},\quad r=0,\ldots,R. \tag{47}$$

The $(R+1)$ parameters, $\{\bar{s}_{m,u,d}, u=0,\ldots,R\}$, can be obtained by solving the set of $(R+1)$ simultaneous linear equations given by Equation (45). These equations can be written in matrix form as follows:

$$\begin{bmatrix} h_{m,d}(0) & h_{m,d}(1) & \ldots & h_{m,d}(R) \\ \mathrm{h}_{m,d}(1) & h_{m,d}(2) & \ldots & h_{m,d}(R+1) \\ \vdots & \vdots & \ddots & \vdots \\ h_{m,d}(R) & h_{m,d}(R+1) & \ldots & h_{m,d}(2R) \end{bmatrix}\begin{bmatrix} \bar{s}_{m,0,d} \\ \bar{s}_{m,1,d} \\ \vdots \\ \bar{s}_{m,R,d} \end{bmatrix}=\begin{bmatrix} x_{m,d}(0) \\ x_{m,d}(1) \\ \vdots \\ x_{m,d}(R) \end{bmatrix}. \tag{48}$$

Note that both $h_{m,\mathrm{d}}(l)$ and $x_{m,d}(r)$ are dependent on dimension in this equation.

3. Experiments

3.1. Conditions

To investigate the relative effectiveness of the three kinds of MDPSM, we perform experiments on a speaker-independent 16-vowel classification task using the TIMIT corpus. Sixteen vowels include 13 monothongs /aa, ae, ah, ao, eh, er, ey, ih, iy, ow, uh, uw, ux/ and three diphthongs /aw, ay, oy/. A total of 462 speakers (41 014 tokens) are employed for context-independent MDPSM training and 168 speakers (14 981 tokens) are employed for testing. The regression order of the mean trajectories and the time-varying variance trajectories are set to 2. We generate MDPSMs with diagonal covariance matrices from 10-dimensional mel-frequency cepstral coefficients (MFCCs) and their derivatives with a 5 ms frame rate. The experimental conditions are listed in Table I. As for the initial variances for the variance trajectory model, the estimates obtained from the EM method as described in Subsection 2.3.2 are used. That is, $s_{m,1,d}$ and $s_{m,2,}$

TABLE I. Experimental conditions

Analysis	
Sampling frequency	16 kHz
Preemphasis	$1 - 0{\cdot}98\ z^{-1}$
Frame length	25·6 msec (Hamming window)
Frame period	5·0 msec
Feature vector	10-order MFCC + 10-order Δ MFCC
Training	
Number of speakers	462
Number of tokens	41014
MDPSM	Context independent models with diagonal covariance matrices
Regression order	2
EM iterations	20
Testing	
Number of speakers	168
Number of tokens	14981

$_d$ in Equation (40) are set to zero for the initial values. Segment $\mathbf{Y}$ is classified as phoneme $\hat{m}$ if

$$\hat{m} = \arg\max_m \log \mathrm{P}(\mathbf{Y}|m). \tag{49}$$

3.2. Effectiveness of variance trajectory modelling

Figure 1 shows the differences between the conventional constant variance MDPSM [Fig. 1(a)] and the variance trajectory model [Fig. 1(b)]. These trajectories are obtained from the model parameters estimated for the /ay/ vowel segments with single mixture. The solid lines show the trajectories μ_t of the first and the second MFCC values. The dotted lines show the trajectories $\bar{\mu}_t$ calculated as:

$$\bar{\mu}_t = \mu_t \pm \sigma_t. \tag{50}$$

where σ_t represents the standard derivation. Note that σ_t is constant throughout the segment for Figure 1(a) and σ_t is time variant for Figure 1(b). In general, variances of central parts of vowel segments are smaller than those of the beginning or the ending parts of them. We can see from these figures that the variance trajectory model can capture these phenomena.

Figure 2 shows log likelihood as a function of iterations on the /aa/ vowel segments (3054 segments in total). The solid line shows the log likelihood for three component MDPSMs with constant variance as described in Subsection 2.3.2. The dotted line represents the three component MDPSMs with the variance trajectory model (VTM).

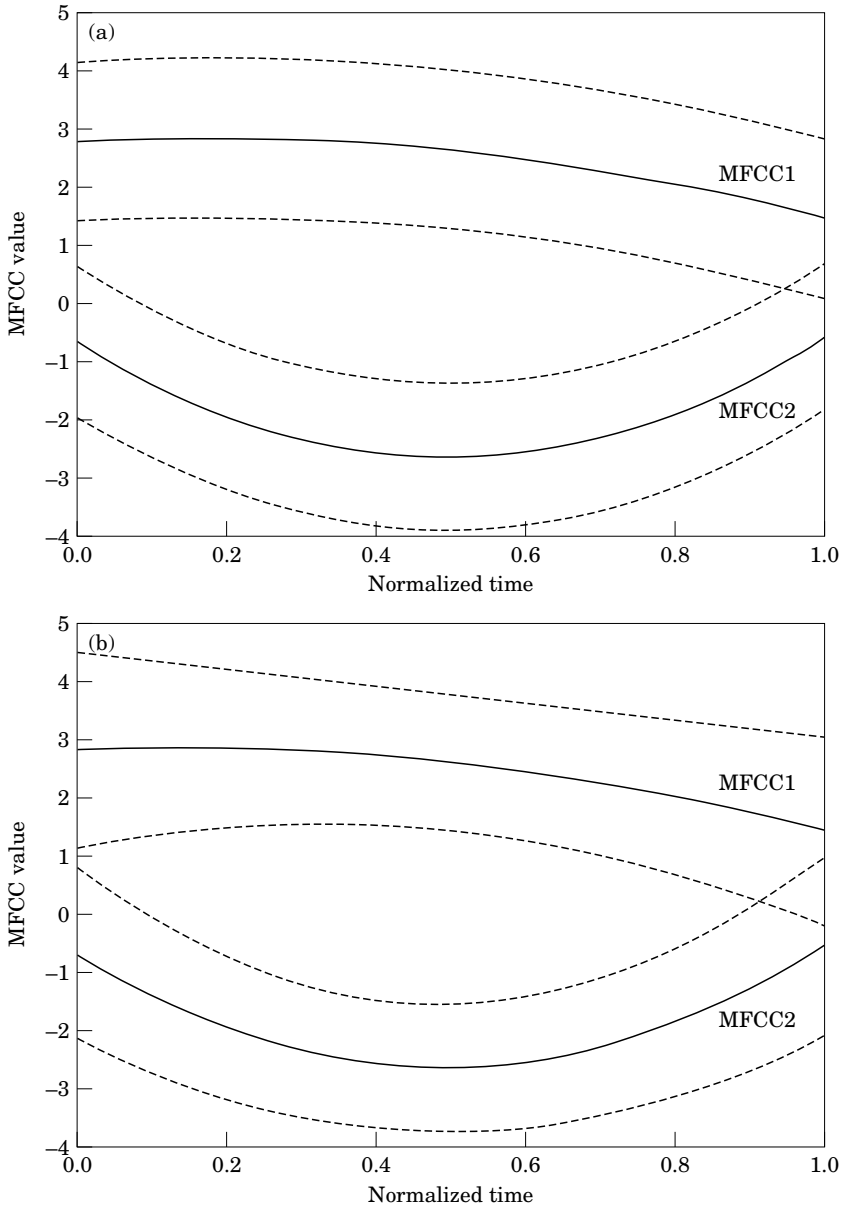

Figure 1. Comparison between (a) constant variance model and (b) variance trajectory model.

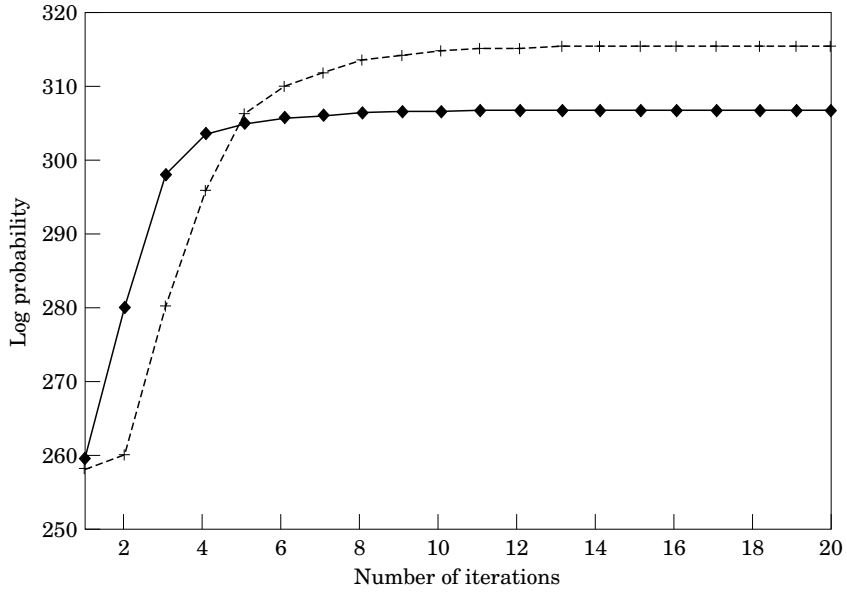

Figure 2. Log likelihood as function of iterations on training data (vowel/aa/). Constant varience (◆); Varience trajectory (+).

We can see from this figure that the VTM gives higher log likelihood than the constant variance model at more than five iterations.

3.3. Classification results

3.3.1. Baseline performances

The classification results based on Equation (49) are shown in Table II. In this table, clustering, EM and VTM stand for the MDPSMs described in Subsections 2.3.1, 2.3.2 and 2.4, respectively. Note that the clustering and EM methods for single mixture give the same performances. Duration probabilities are not considered. From these results, we can say that the variance trajectory model consistently outperformed clustering and

TABLE II. Classification rate (%) (without duration probability)

	Number of mixtures				
Method	1	3	5	7	9
Clustering	54·3	56·2	58·9	59·5	60·1
EM	54·3	59·6	61·4	63·0	63·2
VTM	56·7	60·7	62·6	63·6	63·8

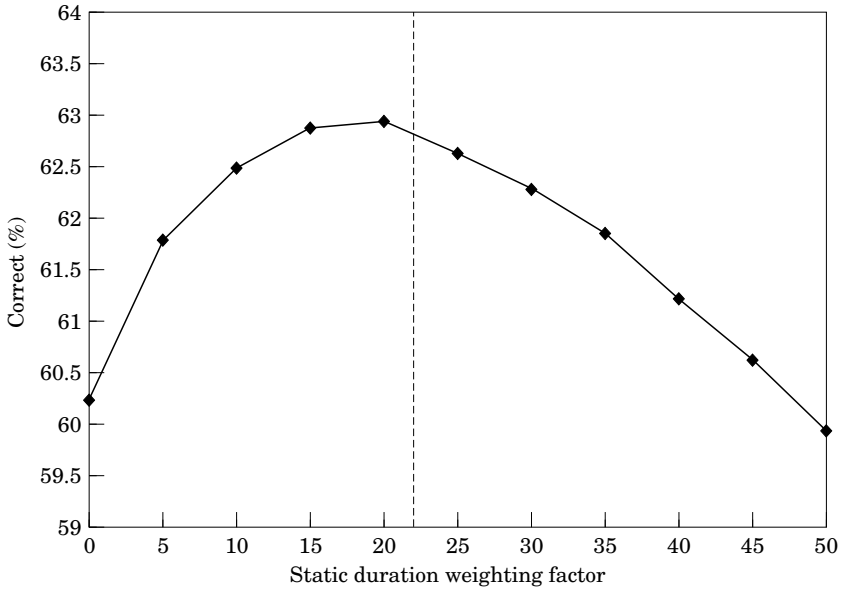

Figure 3. Classification performance as a function of static weighting factor α of duration probability. A vertical line is put at α = average segment duration (= 21·99 frames).

EM-based models. It indicates that time-variant modelling of variances is effective for improving the classification performances.

3.3.2. Utilization of duration probabilities

In general, duration probabilities provide useful information for speech recognition. Therefore, we investigate here the use of the duration probabilities for improving speech recognition performance. These probabilities are computed from a histogram of the training segment durations.

In order to match the dynamic ranges of log $P(\mathbf{Y}|m)$ and log $P(L|m)$ and log $P(L|m)$, here we use the following two approaches. In the first approach, we use a fixed (or, static) weighting and perform classification using the following equation:

$$\hat{m} = \underset{m}{\text{argmax}} \{\log P(\mathbf{Y}|m) + \alpha \log P(L|m)\}, \tag{51}$$

where $P(L|m)$ is L-frame duration probability given phoneme m and α is a static weighting factor. In order to obtain the optimal value of the static weighting factor α, we conduct classification experiments, three mixture variance trajectory models are used. Figure 3 shows the classification performance as a function of α. We have computed the average segment length from all the segments in the training data and found it to be 21·99 frames. In Figure 3, we have put a vertical dotted line at $\alpha = 21{\cdot}99$. We can see from this figure that the classification performance initially increases with α, attains a maximum at $\alpha = 20$, and then starts decreasing. Thus, the best classification

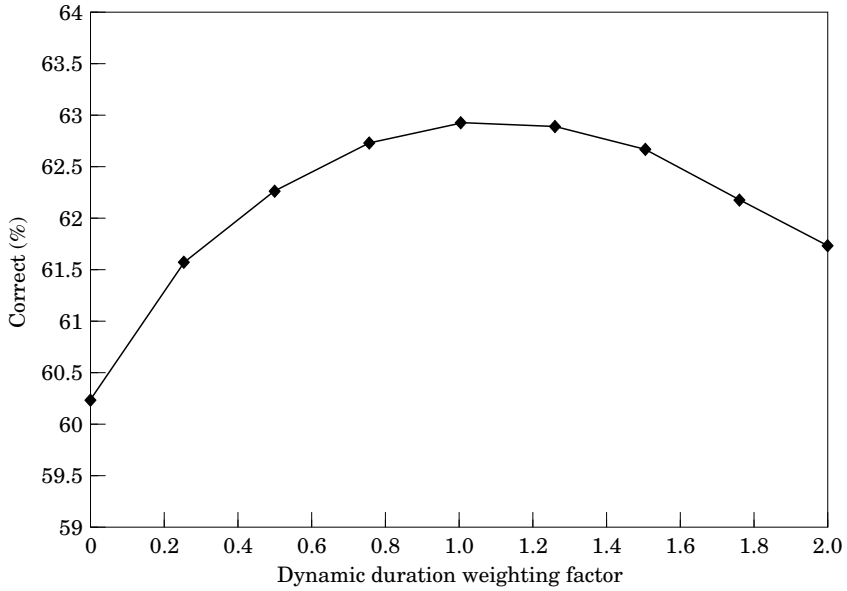

Figure 4. Classification performance as a function of dynamic weighting factor β of duration probability.

performance occurs when the weighting factor is approximately equal to the average segment length.

In the second approach, we use a variable (or, dynamic) weighting for matching the dynamic ranges of log $P(\mathbf{Y} \mid m)$ and log $P(L \mid m)$ and perform classification using the following equation:

$$\hat{m} = \underset{m}{\text{argmax}} \{\log P(\mathbf{Y} \mid m) + \beta L \log P(L \mid m)\}, \tag{52}$$

where βL can be considered as a segment-length dependent dynamic weighting factor. Figure 4 shows the classification performance on the training data as a function of dynamic weighting factor, β. It can be seen from this figure that we get the best classification performance when $\beta = 1{\cdot}0$. Thus, in both the approaches, the optimal value of the weight is approximately equal to the segment length. However, the dynamic weighting is intuitively more appealing, because duration probabilities are weighted depending on segment lengths. Moreover, it is easy to use, as we do not need to calculate the average segment length from the training data. We use the dynamic weighting with $\beta = 1$ in all the classification experiments reported hereafter.

The classification results on the test data using the duration probability (dynamic weighting and $\beta = 1{\cdot}0$) are shown in Table III. From this table, it can be seen that VTM gives consistently higher classification rates compared to constant variance models (EM).

3.3.3. Comparison between VTM and HMM

In order to compare the performance of VTM with a conventional HMM system, we investigate a three-state context independent HMM. Continuous density HMMs for

TABLE III. Classification rate (%) (with duration probability)

	Number of mixtures				
Method	1	3	5	7	9
Clustering	56·8	59·5	61·6	62·2	62·4
EM	56·8	62·3	63·8	65·4	65·9
VTM	58·7	63·4	65·0	66·0	66·2

TABLE IV. Classification results (%) for the VTM and HMM systems

	Number of mixtures				
Method	1	3	5	7	9
VTM (without duration)	56·7	60·7	62·6	63·6	63·8
VTM (with duration)	58·7	63·4	65·0	66·0	66·2
HMM	54·1	58·7	60·9	62·0	62·4

16 vowels are trained using the EM algorithm with 20 iterations. No model parameter is tied. No state skip is allowed in the transition, that is, three frames are required for the minimum duration. The total number of the free parameters for each phoneme HMM is $S(MD+MD+M+1)$(SMD for means, SMD for variances, SM for mixture weights, and S for transition probabilities), where S is the number of HMM states. As for VTM, the total number of free parameters for each phoneme VTM is $(R+1)(MD+MD)+M+68$ ($(R+1)MD$ for means, $(R+1)MD$ for variances, M for mixture weights, and 68 for duration probabilities (In this paper, the minimum and maximum duration lengths are set to 3 and 70, respectively.) When $M=5$, the total number of the free parameters for HMM, VTM without duration probability, and VTM with duration probability become 618, 605 and 673, respectively. Classification results for the VTM and HMM systems are listed in Table IV. From this table, it can be seen that VTMs provide about 4% improvement for each mixture against the three-state HMM whose number of free parameters is equal to that of the VTMs'.

4. Conclusions

In this paper, we have proposed mixture density polynomial segment models (MDPSM), where higher order regression models are used not only for mean trajectory modelling but also for time-varying variance modelling. We have developed a theoretical formulation for estimating the model parameters using the EM algorithm. We have conducted speaker-independent vowel classification experiments using the TIMIT database and reported the classification results. These results indicate that the proposed model gives a consistently better performance than the MDPSM proposed by Gish and Ng (1993). In addition, the proposed model shows significant improvement in classification performance over the conventional HMM.

As the proposed modelling requires the explicit evaluation of different segmentations, the computational requirements in PSM or VTM decoding are generally higher than those in HMM decoding. To reduce the cost of segment evaluations, several techniques have been studied (Zue, Glass, Phillips & Seneff, 1989; Digalakis, Ostendorf & Rohlicek, 1992; Fukada, Aveline, Schuster & Sagisaka, 1997). These researches are important to apply segment models to continuous speech recognition systems.

References

Deng, L. (1992). A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing* **27**, 65–78.

Digalakis, V., Ostendorf, M. & Rohlicek, J. R. (1992). Fast search algorithms for phone classification and recognition using segment-based models. *IEEE Transactions on Signal Processing Parameter estimation for segment model* **40**, 2885–2896.

Fukada, T., Aveline S., Schuster, M. & Sagisaka, Y. (1997). Segment boundary estimation using recurrent neural networks. *Proceedings of EUROSPEECH'97*, Rhodes, Greece, pp. 2839–2842.

Gales, M. & Young, S. J. (1993). Segmental hidden Markov models. *Proceedings of EUROSPEECH'93*, Berlin, Germany, pp. 1579–1582.

Ghitza, O. & Sondhi, M. M. (1993). Hidden Markov models with templates as non-stationary states: an application to speech recognition. *Computer Speech and Language* **2**, 101–119.

Gish, H. & Ng, K. (1993). A segmental speech model with applications to word spotting. *Proceedings of ICASSP'93*, Minneapolis, U.S.A., pp. II-447–II-450.

Gish, H. & Ng, K. (1996). Parametric trajectory models for speech recognition. *Proceedings of ICSLP'96*, Philadelphia, U.S.A., pp. 466–469.

Goldenthal, W. D. & Glass, J. R. (1994). Statistical trajectory models for phonetic recognition. *Proceedings of ICSLP'94*, Yokohama, Japan, pp. 1871–1873.

Gong, Y. & Haton, J. P. (1994). Stochastic trajectory modeling for speech recognition. *Proceedings of ICASSP'94*, Adelaide, Australia, pp. I-57–I-60.

Holmes, W. J. & Russell, M. J. (1995). Speech recognition using a linear dynamic segmental HMM. *Proceedings of EUROSPEECH'95*, Madrid, Spain, pp. 1611–1614.

Ostendorf, M. & Roukos, S. (1989). A stochastic segment model for phoneme-based continuous speech recognition. *IEEE Transactions on Acoustic, Speech and Signal Processing* **37**, 1857–1869.

Paliwal, K. K. (1993). Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer. *Proceedings of ICASSP'93*, Minneapolis, U.S.A., pp. II-215–II-218.

Robinson, T., Hochberg, M. & Renals, S. (1994). IPA: improved phone modelling with recurrent neural networks. *Proceedings of ICASSP'94*, Adelaide, Australia, pp. I-37–I-40.

Zue, V., Glass, J., Philips, M. & Seneff, S. (1989). Acoustic segmentation and phonetic classification in the SUMMIT system. *Proceedings of ICASSP'89*, Glasgow, Scotland, pp. 389–392.

(*Received 11 February 1997 and accepted for publication 11 May 1998*)