MODEL PARAMETER ESTIMATION FOR MIXTURE DENSITY POLYNOMIAL SEGMENT MODELS

Toshiaki Fukada *Yoshinori Sagisaka* *Kuldip K. Paliwal* *

ATR Interpreting Telecommunications Research Laboratories
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
fukada, sagisaka, kkp@itl.atr.co.jp

ABSTRACT

In this paper, we propose parameter estimation techniques for mixture density polynomial segment models (henceforth MDPSM) 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, or (3) EM training of mean and variance trajectories. These parameter estimation methods were evaluated in TIMIT vowel classification experiments. The experimental results showed that modeling both the mean and variance trajectories are consistently superior to modeling only the mean trajectory. We also found that modeling 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 to address some of the shortcomings of HMMs. Broadly speaking, there are two HMM limitations that various models have tried to address: (1) weak duration modeling 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[1][2][3][4][5][6][7][8]. 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 segment modeling proposed by Gish and Ng [4] is one such technique for relaxing the independence assumption. This modeling technique, however, has a serious shortcoming; i.e., 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.

This paper presents model parameter estimation method for mixture density polynomial segment models (MDPSM) with time variant variances[9]. The model parameters of the MDPSM are the mean trajectory coefficients, the variance coefficients and the mixture weights. In our segmental modeling approach, higher order regression models are used not only for mean trajectory modeling but also for time-varying variance modeling. [4] can be viewed as a special case (i.e. 0-th order regression model) of our method. Recently, similar approach was also proposed by Gish and Ng [10]. However, they restricted the time-variation of the covariance to be limited to having three different covariance matrices existing over a segment, while there is no restriction in our modeling except time-varying variance can be modeled with higher order regression models.

The paper is organized as follows. Section 2 starts with an overview of single Gaussian segment modeling, describes two ways of model parameter estimation for MDPSM with time-invariant covariance, and finally provides model parameter estimation formulae for the time-variant covariance case. 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. Single Gaussian Segment Model

Consider an L (in frames) length sequence of observation vectors $y_1^L = [y_1, \ldots, y_L]$ generated by label a, where y_t is a D-dimensional observation (e.g. cepstrum) vector at time t. This sequence defines a segment corresponding to the label a. In the polynomial trajectory model, this segment is represented as follows:

$$y_t = \mu_{a_t} + e_t, \qquad 1 \le t \le L, \tag{1}$$

where μ_{a_t} and e_t are the D-dimensional mean vector and residual error vector, respectively, at time t. The mean vector μ_{a_t} is represented as an R-th order polynomial $\mu_{a_t} = z_L^t B_a$, where $B_a = [b_{a0}, b_{a1}, \ldots, b_{aR}]^T$, and z_L^t is an $(R+1)$-dimensional vector:

$$z_L^t = \begin{cases} [1, 0, 0, \ldots, 0], t = 1 \\ \left[1, \frac{t-1}{L-1}, (\frac{t-1}{L-1})^2, \ldots, (\frac{t-1}{L-1})^R\right], 1 < t \le L. \end{cases} \tag{2}$$

*On leave from Griffith University, Brisbane, Australia.

Note that the polynomial is defined on normalized time.

When the error vectors e_t are considered independent and identically distributed as a Gaussian with zero mean and an invariant covariance matrix Σ_a, the likelihood of the segment can be expressed as,

$$\mathrm{P}(y_1,\ldots,y_L|a) = \prod_{t=1}^{L} f(y_t)$$

$$= \prod_{t=1}^{L} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_a|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_t-\mu_{a_t})^T \Sigma_a^{-1}(y_t-\mu_{a_t})}. \quad (3)$$

The model parameters, contained in B_a and Σ_a can be estimated using a maximum likelihood criterion as described in [4]. From now on, we omit a for simplification.

2.2. Mixture Density Polynomial Segment Model

The discussion in the previous section was concerned with single Gaussian segment modeling. We extend here the previous modeling to a mixture density case. In this case, Eq. (3) is represented by an M-mixture Gaussian:

$$f(y_t) = \sum_{k=1}^{M} w_k \; f_k(y_t) = \sum_{k=1}^{M} w_k \; \mathcal{N}(y_t, B_k, \Sigma_k) \quad (4)$$

where

$$\mathcal{N}(y_t, B_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_t - z_L^t B_k)^T \Sigma_k^{-1}(y_t - z_L^t B_k)} \quad (5)$$

and w_k is the weight of the k-th mixture component. The mixture components satisfy the relation $\sum_{k=1}^{M} w_k = 1$. The model parameters B_k, Σ_k, and w_k in Eq. (4) can be estimated by segment clustering or by EM training. These methods are described in detail in **2.2.1.** and **2.2.2.**, respectively.

2.2.1. Clustering Method

One simple way of generating MDPSM is based on segment clustering. That is, the training segments for a label (e.g. phone) are partitioned into M regions using the K-means clustering algorithm. The distance measure used during the clustering is a "multivariate Gaussian distance measure". B_k and Σ_k can be obtained in the same way as in the single mixture case using the segments assigned to cluster k. w_k is calculated as the relative frequency of the segments.

2.2.2. EM Method

To obtain model parameters that maximize $\mathrm{P}(y_1^L|a)$, we derive here re-estimation formulas based on the EM algorithm. The EM re-estimation formulas can be derived by maximizing the following auxiliary function Q with the mixture components being the hidden variables:

$$Q(\bar{\Phi}|\Phi) = E[\log \mathrm{P}(y_1^L, k|\bar{\Phi})|y_1^L, \Phi]$$

$$= \sum_{k=1}^{M} \frac{\mathrm{P}(k, y_1^L|\Phi)}{\mathrm{P}(y_1^L|\Phi)} \log \mathrm{P}(y_1^L, k|\bar{\Phi}) \quad (6)$$

where Φ and $\bar{\Phi}$ are the sets of the current model parameters and the re-estimated model parameters, respectively. k denotes the index of a mixture component. Maximizing Eq. (6) is equivalent to individually maximizing the following two functions:

$$Q_1(\bar{\Phi}|\Phi) = \sum_{k=1}^{M} \frac{\mathrm{P}(k, y_1^L|\Phi)}{\mathrm{P}(y_1^L|\Phi)} \log \mathrm{P}(y_1^L|\bar{\Phi}) \quad (7)$$

with respect to B_k and Σ_k, and

$$Q_2(\bar{\Phi}|\Phi) = \sum_{k=1}^{M} \frac{\mathrm{P}(k, y_1^L|\Phi)}{\mathrm{P}(y_1^L|\Phi)} \log \mathrm{P}(k|\bar{\Phi}) \quad (8)$$

with respect to w_k.

Let the probability $\mathrm{P}(k, y_1^L|\Phi)/\mathrm{P}(y_1^L|\Phi)$ in Eq. (7) and Eq. (8) be denoted as $\gamma_{k,t}$ using the current model parameters Φ. Then, we can estimate $\gamma_{k,t}$ efficiently as,

$$\gamma_{k,t} = \begin{cases} \dfrac{\alpha_t \beta_{t+1} w_k f_k(y_{t+1})}{\mathrm{P}(y_1^L|\Phi)}, & t = 1,\ldots,L-1 \\ \dfrac{\alpha_T}{\mathrm{P}(y_1^L|\Phi)}, & t = L \end{cases} \quad (9)$$

where α_t and β_t are obtained recursively:

$$\alpha_t = \begin{cases} f(y_1), & t = 1 \\ \alpha_{t-1} f(y_t), & t = 2,\ldots,L \end{cases} \quad (10)$$

$$\beta_t = \begin{cases} 1, & t = L \\ \beta_{t+1} f(y_{t+1}), & t = L-1,\ldots,1. \end{cases} \quad (11)$$

First, we consider obtaining the mean trajectory parameters of the m-th component, $\bar{B}_m = [\bar{b}_{m0}, \bar{b}_{m1}, \ldots, \bar{b}_{mR}]^T$. These parameters can be obtained through differentiation of Eq. (7) with respect to $\bar{b}_{mr}$ and solving the equation $\partial Q_1/\partial \bar{b}_{mr} = 0$.

The variance of the m-th component, $\bar{\Sigma}_m$, can also be obtained through differentiation of Eq. (7) with respect to the i-th column and j-th row element $\bar{\sigma}_{mij}$ and solving the equation $\partial Q_1/\partial \bar{\sigma}_{mij} = 0$. It can be shown that differentiation with respect to $\bar{b}_{mu}$ results in

$$\sum_{u=1}^{R} C(u+r)\bar{b}_{mu} = V(r) \quad (12)$$

where

$$C(l) = \sum_{t=1}^{L} g_{ml,t}, \qquad l = 0,\ldots,2R \quad (13)$$

$$V(r) = \sum_{t=1}^{L} g_{mr,t} \; y_t, \qquad r = 0,\ldots,R \quad (14)$$

$$g_{ml,t} = \gamma_{m,t}\{(t-1)/(L-1)\}^l. \quad (15)$$

As for the variance, $\bar{\Sigma}_m$ is

$$\bar{\Sigma}_m = \sum_{t=1}^{L} \gamma_{m,t} \; (y_t - \bar{\mu}_{m,t})(y_t - \bar{\mu}_{m,t})^T \Big/ \sum_{t=1}^{L} \gamma_{m,t}. \quad (16)$$

The weighting coefficient $\bar{w}_m$ can be obtained from Eq. (8) by application of a Lagrange optimization using Lagrange multipliers:

$$\bar{w}_m = \sum_{t=1}^{L} \gamma_{m,t} / \sum_{t=1}^{L} \sum_{k=1}^{M} \gamma_{k,t}. \quad (17)$$

The results of the model parameters obtained from clustering method are used as initial model parameters for the EM algorithm.

2.2.3. Variance Trajectory Model

In the previous modeling, the variance is time invariant throughout a segment. Here, we try to extend the segment model with time variant variances for more precise modeling. This model can be realized by representing a change in the variance as a trajectory.

In the variance trajectory model, variance Σ_k in Eq. (5) is represented as a W-th order polynomial $\Sigma_{k,t} = z_L^t G_k$, where

$$G_k = [g_{k0}, g_{k1}, \ldots, g_{kW}]^T. \quad (18)$$

z_L^t defined by Eq. (2) is given as a $(W+1)$-dimensional vector. In this model, estimates of the mean trajectory and weight parameters can be obtained in a similar way to that described in **2.2.2.**. However, the computation of variance differs as follows.

The variance of the m-th component $\bar{\Sigma}_{m,t}$ can be obtained through differentiation of Eq. (7) with respect to the i-th column and j-th row element $\bar{\sigma}_{m,t,ij} = [\bar{s}_{m,ij0}, \bar{s}_{m,ij1}, \ldots, \bar{s}_{m,ijW}]^T$ and solving the equation $\partial Q_1 / \partial \bar{s}_{m,ijr} = 0$. From this equation, we obtain

$$\sum_{t=1}^{L} g_{mr,t} \left\{ \bar{\sigma}_{m,t,ij} - (y_t - \bar{\mu}_{m,t})_i (y_t - \bar{\mu}_{m,t})_j \right\} / \bar{\sigma}^2_{m,t,ij} = 0. \quad (19)$$

However, as $\bar{\sigma}_{m,t,ij}$ in the denominator is a time dependent value, Eq. (19) can not be solved as a linear equation. Therefore, we use an approximation assuming $\bar{\sigma}_{m,t,ij}$ in the denominator is replaced by a value calculated from the current variance trajectory $\sigma_{m,t,ij}$. Equation (19) then becomes

$$\sum_{t=1}^{L} g_{mr,t} \left\{ \bar{\sigma}_{m,t,ij} - (y_t - \bar{\mu}_{m,t})_i (y_t - \bar{\mu}_{m,t})_j \right\} / \sigma^2_{m,t,ij} = 0. \quad (20)$$

We can now solve this linear equation and obtain the following formula:

$$\sum_{u=1}^{W} H(u+r) \bar{s}_{m,iju} = Z(r) \quad (21)$$

where

$$H(l) = \sum_{t=1}^{L} g_{ml,t} / P_{W,m,ijt}, l = 0, \ldots, 2W \quad (22)$$

$$Z(r) = \sum_{t=1}^{L} g_{mr,t} (y_t - \bar{\mu}_{m,t})_i (y_t - \bar{\mu}_{m,t})_j / P_{W,m,ijt}, \quad r = 0, \ldots, W \quad (23)$$

$$P_{W,m,ijt} = \{ \sum_{n=1}^{W} s_{m,ijn} (\frac{t-1}{L-1})^n \}^2. \quad (24)$$

Note that both $H(l)$ and $Z(r)$ are dimension dependent vector.

3. EXPERIMENTS

3.1. Conditions

To investigate the relative effectiveness of the three kinds of MDPSM, we performed experiments on a speaker-independent 16-vowel classification task using the TIMIT corpus. 462 speakers (41,014 tokens) were employed for context-independent MDPSM training and 168 speakers (14,981 tokens) were employed for testing. The regression order of the mean trajectories and the time varying variance trajectories were set to 2. We generated MDPSM with diagonal covariance matrices from 10-dimensional MFCCs and their derivatives with a 5 ms frame rate. As for the initial variances for variance trajectory model, the results obtained from the clustering method were used. That is, g_{k1} and g_{k2} in Eq.(18) were set to zero for the initial values. The duration probabilities, which were computed from a histogram of the training segment durations, were used in the classification. Segement y_1^L can be classified as phoneme $\hat{m}$ by

$$\hat{m} = \underset{m}{\operatorname{argmax}} \left\{ \log \mathrm{P}(y_1^L | m) + L \log \mathrm{P}(L|m) \right\}. \quad (25)$$

In order to match the dynamic ranges of $\log \mathrm{P}(y_1^L|m)$ and $\log \mathrm{P}(L|m)$, $L \log \mathrm{P}(L|m)$ was used instead of $\log \mathrm{P}(L|m)$ in Eq.(25) . We found this operation gaving consistently higher classification performance than that of Eq.(25) .

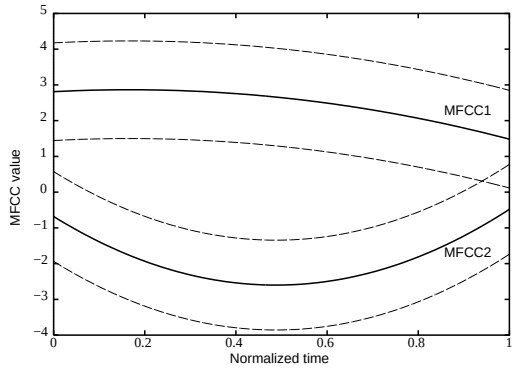
3.2. Effectiveness of Variance Trajectory Modeling

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

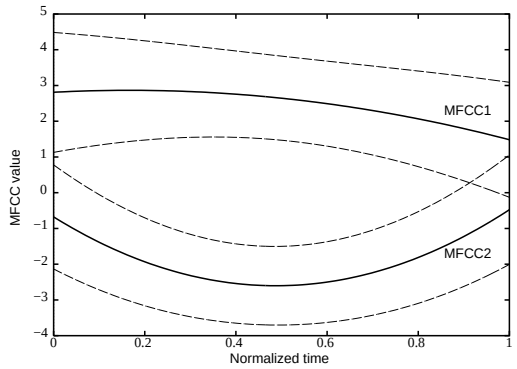
$$\bar{\mu}_t = \mu_t \pm \sigma_t. \quad (26)$$

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 figure that variance trajectory model can capture these phenomena.

Figure 2 shows log likelihood as function of iterations on the /aa/ vowel segments (3,054 segments in total). Solid line shows the log likelihood for three component MDPSMs with constant variance described in **2.2.2.**. Dotted line is for three component MDPSMs with variance trajectory model (VTM). We can see from this figure that VTM gives higher log likelihood than the constant variance model at more than five iterations.



(a) Conventional PSM (variance is constant)



(b) Variance trajectory model

Figure 1. Comparison between constant variance model and variance trajectory model.

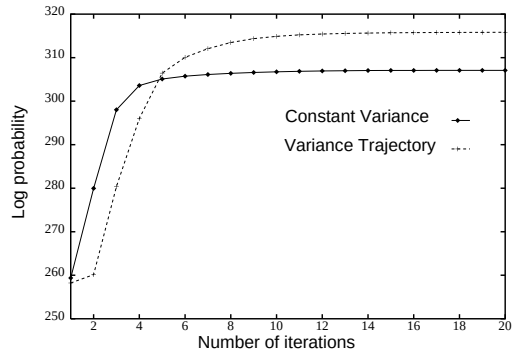


Figure 2. Log likelihood as function of iterations on training data (vowel /aa/).

3.3. Classification Results

The classification results are shown in Table 1. For comparison, results obtained using a three-state HMM are also listed in the table. From this table, it can be seen that (1) variance trajectory models (VTM) give consistently higher classification rates compared to constant variance models (EM) and (2) 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'.

Table 1. Classification rate (%).

method	mixtures				
	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
HMM	54.1	58.7	60.9	62.0	62.4

4. CONCLUSIONS

We proposed parameter estimation techniques for mixture density polynomial segment models (MDPSM) with time variant variances. In this method, higher order regression models are used not only for mean trajectory modeling but also for time-varying variance modeling. The classification results showed that the proposed method gave consistently better performance than the MDPSM proposed by Gish and Ng[4]. In addition, the proposed method achieved significant improvement over the conventional HMM.

REFERENCES

[1] M. Ostendorf and S. Roukos: "A stochastic segment model for phoneme-based continuous speech recognition," IEEE Trans. on Acoust., Speech and Signal Proc., vol 37, 12, pp. 1857–1869, 1989.

[2] L. Deng: "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol27, pp. 65-78, 1992.

[3] K.K. Paliwal: "Use of Temporal Correlation Between Successive Frames in a Hidden Markov Model Based Speech Recognizer," *Proc. ICASSP-93*, pp. II-215–II-218, 1993.

[4] H. Gish and K. Ng: "A Segmental Speech Model with Applications to Word Spotting," *Proc. ICASSP-93*, pp. II-447–II-450, 1993.

[5] M. Gales and S. Young: "Segmental hidden Markov models," *Proc. EUROSPEECH-93*, pp. 1579–1582, 1993.

[6] O. Ghitza and M. Sondhi: "Hidden Markov Models with Templates as Non-stationary States: An Application to Speech Recognition," *Computer Speech and Language*, 2, pp. 101-119, 1993.

[7] T. Robinson, M. Hochberg and S. Renals: "IPA: Improved Phone Modelling with Recurrent Neural Networks," *Proc. ICASSP-94*, pp. I-37–I-40, 1994.

[8] W. Goldenthal and J. Glass: "Statistical Trajectory Models for Phonetic Recognition," *Proc. ICSLP-94*, pp. 1871-1873, 1996.

[9] T. Fukada, Y. Taniguchi and Y. Sagisaka: "Model Parameter Estimation for Mixture Density Stochastic Segment Models," Tech. Rep. SP 96-24, IEICE, June 1996 (*in Japanese*).

[10] H. Gish and K. Ng: "Parametric trajectory models for speech recognition," *Proc. ICSLP-96*, pp. 466–469, 1996.