SPEAKER ADAPTATION EXPERIMENTS USING NONSTATIONARY-STATE HIDDEN MARKOV MODELS: A MAP APPROACH

C. Rathinavelu[1] *Li Deng*

Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L3G1, Canada
[1]Currently at: Speech Processing Group, Bell Labs, Lucent Technologies, Naperville, IL 60566, USA

ABSTRACT

In this paper, we report our recent work on applications of the MAP approach to estimating the time-varying polynomial Gaussian mean functions in the nonstationary-state or trended HMM. Assuming uncorrelatedness among the polynomial coefficients in the trended HMM, we have obtained analytical results for the MAP estimates of the time-varying mean and precision parameters. We have implemented a speech recognizer based on these results in speaker adaptation experiments using TI46 corpora. Experimental results show that the trended HMM always outperforms the standard, stationary-state HMM and that adaptation of polynomial coefficients only is better than adapting both polynomial coefficients and precision matrices when fewer than four adaptation tokens are used.

1. INTRODUCTION

Bayesian learning has been widely used for obtaining maximum a posteriori (MAP) estimates of the hidden Markov model (HMM) parameters (e.g. [6, 4]). The MAP estimation framework provides a way of incorporating prior information in the training process. This is particularly useful for dealing with problems arising from sparse training data, out of which the classical maximum likelihood (ML) approach gives poor estimates of model parameters. This MAP approach has been shown to be effective for speaker adaptation of alpha-digit recognition and a number of other tasks where the time-invariant (given HMM states) Gaussian densities are adapted to sparse training data obtained from new speakers [5].

The formulation of the trended HMM, also called the parametric nonstationary-state HMM or parametric trajectory model, has been proposed as a superior model for speech acoustics than the conventional HMM, and has been successfully used in speech recognition applications [1, 2, 7]. The parameters of the trended HMM, especially the state-dependent time-varying Gaussian means, used in the past were trained by a modified Viterbi algorithm based on the joint-state ML principle [2]. In our previous study, we extended the ML training algorithm to the minimum classification error (MCE) training algorithm for discriminatively estimating the state-dependent polynomial coefficients in the trended HMM [7]. Just as an extension of the ML trained unimodal Gaussian trended HMM to the MCE-trained trended HMM is a step towards superior discrimination of speech classes, we expect that the same superiority can be achieved in our trended-HMM framework (due to its superior modeling capabilities) by extending the ML-trained HMM to MAP-trained trended HMM for speaker adaptation applications.

In this study, we investigate the problem of the MAP approach to estimating the time-varying polynomial Gaussian mean functions in the trended HMM. Assuming uncorrelatedness among the polynomial coefficients in the trended HMM, we obtain analytical results for the MAP estimates of the time-varying mean and precision parameters. According to these results, the MAP estimates can be viewed as a weighted average of the estimate that the classical ML method would give and an estimate based on prior information. To examine the performance of the extended techniques, the MAP framework is applied to speaker adaptation experiments using TI46 corpora. The properties of the MAP formulation for training the trended HMM is analyzed by examining goodness-of-fit of the raw speech data to the polynomial trajectories in the model, and comparative experimental results on alphabet classification are reported which demonstrated the effectiveness of the MAP algorithm for the trended HMM.

2. MAP ESTIMATES FOR TRENDED HMMS

Consider the trended HMM given by [2]:

$$\mathcal{O}_t \quad = \quad \mathcal{X}_t^{Tr}\mathcal{B}_i + \mathcal{R}_t(\sigma_i^2), \tag{1}$$

where $\mathcal{R}_t \sim i.i.d.\mathcal{N}(0,\sigma_i^2)$, $\mathcal{B}_i = [\mathcal{B}_i(0)\ \mathcal{B}_i(1)\ \cdots\ \mathcal{B}_i(P)]^{Tr}$ is a $(P+1)\times 1$ vector of state-dependent polynomial regression coefficients, $\mathcal{X}_t = [(t-\tau_i)^0\ (t-\tau_i)^1\ \cdots\ (t-\tau_i)^P]^{Tr}$ is a $(P+1)\times 1$ vector of exogenous explanatory variables with $(t-\tau_i)$ representing the sojourn time in state i. To simplify the presentation of our approach, the data feature vectors $\mathcal{O}_t$, $t = 1, 2, \cdots, T$ are assumed to be scalar-valued observation data sequence of length T. The MAP formulation requires a joint prior distribution for both $\mathcal{B}_i$ and σ_i^2 (which are treated as random variables in the Bayesian analysis [4]).

Suppose the prior information about $\mathcal{B}_i$ conditioned on the value for σ_i^{-2} is represented by a Gaussian random variable $\mathcal{N}(\mathcal{B}_i;\mu_i,\sigma_i^2\mathcal{M}_i)$. Its probability density function is

$$f(\mathcal{B}_i|\sigma_i^{-2}) \quad = \quad (2\pi\sigma_i^2)^{-\frac{1}{2}}|\mathcal{M}_i|^{-\frac{1}{2}}\exp\left[-0.5\sigma_i^{-2}(\mathcal{B}_i-\mu_i)^{Tr}\right.$$
$$\left.\mathcal{M}_i^{-1}(\mathcal{B}_i-\mu_i)\right]. \tag{2}$$

Thus, prior to the observation of data samples, the best guess to the value of $\mathcal{B}_i$ is represented by the $(P+1)\times 1$ vector μ_i, and the confidence in this guess is summarized by the $(P+1)\times(P+1)$ matrix $\sigma_i^2\mathcal{M}_i$; a lower degree of the confidence is represented by a larger diagonal element of $\mathcal{M}_i$. Knowledge about the exogenous variable $\mathcal{X}_t$ is presumed to have no effects on the prior distribution; hence Eqn.(2) also describes the density $f(\mathcal{B}_i|\sigma_i^{-2},\mathcal{X}_t)$. Following [6], it is convenient to describe the prior distribution not in terms of the variance σ_i^2 but rather in terms of the reciprocal of the variance, σ_i^{-2}, which is known as the precision. The prior distribution for precision σ_i^{-2} is provided by the gamma distribution [5]:

$$f(\sigma_i^{-2}) = \frac{q_i^{p_i}}{\Gamma(p_i)}\sigma_i^{-2(p_i-1)}\exp\left(-q_i\sigma_i^{-2}\right), \quad (3)$$

where $p_i>0$ and $q_i>0$ are parameters that describe the prior information. Thus, $f(\mathcal{B}_i,\sigma_i^{-2})$, the joint prior density for $\mathcal{B}_i$ and σ_i^{-2}, is given by the product of Eqn.(2) and Eqn.(3), or a normal-gamma distribution. The choice of such a prior density is made because the normal-gamma density is the conjugate density of the normal distribution, a fact that is essential for the analytical derivation of the MAP estimates.

The MAP estimates are obtained according to

$$\hat{\theta}_{MAP} = argmax_\theta\left[\mathcal{Q}(\theta|\theta_0)+\log f(\mathcal{B}_i,\sigma_i^{-2})\right], \quad (4)$$

where the simplified log likelihood function is given by

$$\begin{aligned}\mathcal{Q}(\theta|\theta_0) &= \sum_{t=1}^{T}\sum_{i=1}^{N}\gamma_t(i)\left[0.5\log(\sigma_i^{-2})-0.5\sigma_i^{-2}\right.\\ &\quad\left.(\mathcal{O}_t-\mathcal{X}_t^{Tr}\mathcal{B}_i)^2\right]. \quad (5)\end{aligned}$$

The quantity $\gamma_t(i)$ is either to be one if the model generates $\mathcal{O}_t$ in state i at time t or to be zero otherwise. The maximization of Eqn.(4) is solved by using the expectation-maximization (EM) algorithm. Due to space limitations, we only summarize the final results for the MAP estimates of $\mathcal{B}_i$, σ_i^2, and of their prior parameters here:

$$\hat{\mathcal{M}}_i = \left[\sum_{t=1}^{T}\gamma_t(i)\mathcal{X}_t\mathcal{X}_t^{Tr}+\mathcal{M}_i^{-1}\right]^{-1}, \quad (6)$$

$$\hat{\mu}_i = \hat{\mathcal{M}}_i\sum_{t=1}^{T}\gamma_t(i)\mathcal{X}_t\mathcal{O}_t+\mu_i, \quad (7)$$

$$\begin{aligned}\hat{q}_i &= 2q_i+\sum_{t=1}^{T}\gamma_t(i)(\mathcal{O}_t-\mathcal{X}_t^{Tr}\hat{\mathcal{B}}_i)^2\\ &\quad+(\hat{\mathcal{B}}_i-\mu_i)^{Tr}\mathcal{M}^{-1}(\hat{\mathcal{B}}_i-\mu_i), \quad (8)\end{aligned}$$

$$\hat{p}_i = \sum_{t=1}^{T}\gamma_t(i)+(P+1)+2(p_i-1), \quad (9)$$

$$\hat{\mathcal{B}}_i = \hat{\mu}_i, \quad (10)$$

$$\hat{\sigma}_i^2 = \frac{\hat{q}_i}{\hat{p}_i} \quad (11)$$

According to the above formula, the MAP estimates can be interpreted as a weighted average of the corresponding prior information and of the sample data. The weights are computed iteratively based on a combination of the prior speaker-independent model parameters and of the new-speaker data in a non-linear fashion. The difference between the ML estimation procedure and the MAP procedure lies in the assumption of an appropriate prior distribution of the parameters to be estimated. By using a diffuse prior information, represented as $p_i=0$, $q_i=0$ and $\mathcal{M}_i^{-1}=0$, the MAP estimates for $\mathcal{B}_i$ and σ_i^2 obtained above would become identical to the ML estimates derived in [2].

3. SPEAKER ADAPTATION EXPERIMENTS

The experiments conducted to evaluate the MAP approach are aimed at recognizing the 26 letters in the English alphabet, contained in the TI46 speaker dependent isolated word corpus. It is produced by 16 speakers, eight males and eight females. The speaker-independent (SI) training set consists of 26 tokens per word from each of six male and six female speakers. For the remaining four speakers, up to ten tokens of each word are used as adaptation training data, and the remaining 16 tokens used as speaker dependent test data.

The preprocessor produces a vector of 26 elements consisting of 13 Mel-frequency cepstral coefficients (MFCCs) and 13 delta MFCCs for every 10 msec of speech. In computing MFCCs, 25 triangular band pass filters are simulated, spaced linearly from 0 to 1 kHz and exponentially from 1 kHz to 8.86 kHz, with the adjacent filters overlapped in the frequency range by 50%. The FFT power spectral points are combined using a weighted sum to simulate the output of the triangular filter. The MFCCs are then computed according to

$$C_p = \sum_{r=1}^{25}S_r\cos\left(p\times[r-0.5]\times\frac{\pi}{25}\right),\ 0\le p\le 12$$

where S_r is the log-energy output of the rth mel-filter [9]. The delta MFCCs are constructed by taking the difference between two frame forward and two frame backward of the MFCCs. This window length of 50ms is found to be optimal in capturing the slope of the spectral envelope, i.e. the transitional information [8]. The augmented MFCCs and delta MFCCs are provided as the data input for every frame of speech into the modeling stage.

Each word is represented by a single left-to-right, three-state HMM (no skips). The speaker-dependent (SD) models are trained from adaptation data using five-iterations of the modified Viterbi algorithm with single mixture for each state in the HMMs [2]. To set up a baseline speaker-independent (SI) performance on the test data set, we created the SI models with a single mixture distribution for each state in the HMMs, by combining the parameters in the mixture components which had been well trained using the SI training set. The combination formulas are

$$\mathcal{B}_i(p) = \sum_{m=1}^{M}W_m\mathcal{B}_{im}(p),\quad p=0,1,\cdots,P$$

Number of Adaptation Tokens	Polynomial Order P=0 (SI=69.95%)		
	SD	SA1	SA2
1	58.35%	78.97%	74.39%
2	71.15%	82.33 %	80.41%
3	77.70%	83.77%	82.99%
4	82.69%	84.80%	84.08%
5	85.40%	84.86%	85.82%
6	86.66%	86.60%	86.24%
7	87.56%	87.50%	87.14%
8	87.98%	88.46%	87.56%
9	87.86%	88.58%	88.88%
10	88.28%	88.65%	89.66%

Table 1. Summary of speaker adaptation results for constant-trended HMM (benchmark, P=0).

Number of Adaptation Tokens	Polynomial Order P=1 (SI=75.48%)		
	SD	SA1	SA2
1	46.82%	84.14%	75.78%
2	74.58%	86.84%	84.08%
3	82.51%	88.11%	87.02%
4	85.64%	88.58%	89.00%
5	86.48%	88.76%	90.08%
6	88.52%	89.66%	91.05%
7	88.58%	90.08%	91.89%
8	89.54%	90.14%	90.93%
9	90.20%	90.39%	91.47%
10	91.05%	91.65%	92.07%

Table 2. Summary of speaker adaptation results for linear-trended HMM (P=1).

and the variance to be the variance of the corresponding Gaussian mixture distribution:

$$\sigma_i^2 = \sum_{m=1}^{M} W_m \sigma_{im}^2,$$

where M, the actual number of mixture components used in each state, is set to five, W_m is the mixture weight, $\mathcal{B}_{im}(p)$ is the time-varying polynomial mean coefficients, and σ_{im}^2 is the variance of the mth mixture component residing in the ith state.

The initial prior density parameters are estimated first from those of the SI mixture HMMs according to

$$p_i = \frac{1}{\sum_{m=1}^{M} W_m \sigma_{im}^2}, \quad (12)$$

$$q_i = 1.0, \quad (13)$$

$$\mu_i(p) = \sum_{m=1}^{M} W_m \mathcal{B}_{im}(p), \quad (14)$$

$$\mathcal{M}_i(p) = \frac{1}{p_i \sum_{m=1}^{M} W_m (\mathcal{B}_{im}(p) - \mu_i(p))}, \quad (15)$$

These prior parameters are then updated over iterations of the batch MAP algorithm according to Eqns.(6)-(9). Note that $M_i(p)$ above denotes the pth element of the diagonal correlation matrix M_i. In the MAP batch estimation, the parameters are updated after processing all tokens for each iteration, in contrast with sequential adaptation where the parameters are adjusted at the end of processing each token. We will not address the sequential adaptation procedure in this study. In all of our experiments, a total of five batch adaptation iterations are performed.

The speech recognition rates, averaged over two males and two females, are summarized in Table 1 and Table 2, for conventional, stationary-state HMM (benchmark) and for the trended HMM, respectively. Four experimental setups have been used: 1) speaker-independent (SI); 2) speaker-dependent (SD) ; 3) speaker-adaptation and adapting only polynomial coefficients for the time-varying means (SA1); and 4) speaker-adaptation and adapting both polynomial coefficients and precision matrices (SA2). The results in Table 1 and Table 2 are shown as a function of the number of word tokens used in training from a new speaker. Comparing results in Table 1 and Table 2, the effectiveness of the MAP training on the trended HMM is clearly demonstrated. For example, in the SA1 experiments, the error rate reduction of 26.8% is obtained when moving from $P = 0$ (83.77%) model to $P = 1$ (88.11%) model with three adaptation takens. The best recognition rate of 92.1% is achieved when both polynomial coefficients and precision matrices are adapted using all ten tokens of adaptation data. The rate drops gradually with fewer adaptation tokens for both SA1 and SA2 experiments, with somewhat faster drop for SA2 than for SA1. In contrast, for the SD experiments, the recognition rates drop rapidly when the training tokens reduce from ten to one.

The results in Table 1 and Table 2 also show that the MAP estimates (SA1 and SA2) become approaching the ML estimates (SD) in performance when the number of training token increases from one to ten. This is reassuring because under the asymptotic condition, the posterior density would be dominated by the sample data likelihood function as demonstrated in Eqn.(10) and Eqn.(11) with $T \to \infty$.

4. DATA FITTING RESULTS

To analyze the mechanisms underlying the superiority of the MAP training on the trended HMM, we performed data fitting experiments. Once the structure of the trended HMM is determined, the MAP algorithm discussed in Section 2 is used to reestimate the ML-trained trended HMM parameters using a fixed set of adaptation data. The MAP models are constructed using the SA1 experimental setup with one adaptation token. Fig. 1 shows the results of fitting a test utterance (letter *a* from a first female speaker in the TI46 speech corpus) using the benchmark (P=0) and trended (P=1) HMMs. (Use of first-order MFCC, C_1, as speech data here, shown in solid lines in Fig. 1, is for illustration purposes only. Similar results are available for higher order cepstral coefficients.) The top two subplots of Fig.1 show the data-fitting results (dashed lines) for SI benchmark HMM (left) and trended HMM (right) when both models are trained by the ML method. The bottom

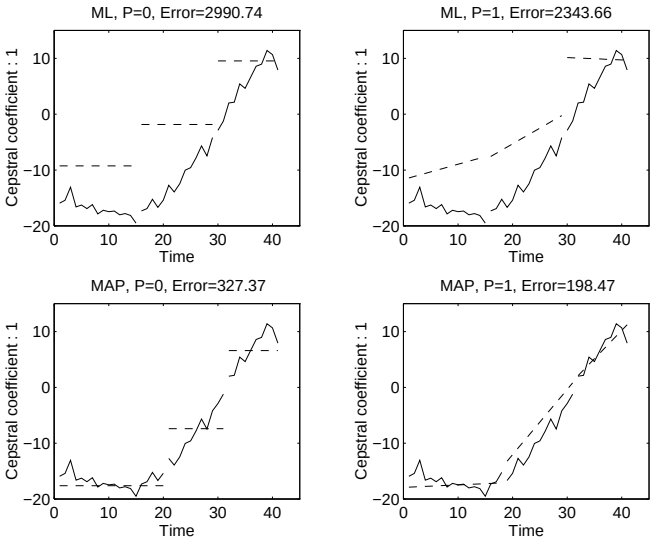

Figure 1. Fitting three-state *a*/ey/ models (dashed-lines) to a speech data sequence (solid lines)

two subplots show the corresponding results (dashed lines) using the MAP-trained HMMs (SA1). In all the plots, the solid lines are the real speech data, $\mathcal{O}_t$, of the C_1 sequence from a test token not used in adapting the HMMs. The vertical axis represents the magnitude of C_2 and the horizontal time axis is expressed in terms of the frame number. For each sub-plot of Fig. 1, the two break-points in the otherwise continuous solid lines correspond to the frames at which the optimal state transitions occur from state one to state two, and from state two to state three, respectively. The dashed lines in all sub-plots of Fig. 1 are the four different trend functions, varying in the polynomial order ($P = 0$ or $P = 1$) and in the training procedure (ML or MAP). These labels are shown at the head of each sub-plot, together with the data-fitting error computed by a linear summation of the residual squares over the states and over the state-bound time frames.

It is observed that the MAP-trained trended HMM fits the test token better than any other alternatives. For the benchmark HMM, error reduction in data fitting by incorporating the MAP training goes from 2990 to 327. The MAP method for the trended HMM plays a more significant role of reducing the data-fitting error (a measure of better modeling capability) from 2343 to 198. This suggests that the time-varying mean parameters in the trended HMM represent essential characteristics of a particular speaker and they can be effectively estimated with a very small amount of training data using the MAP training procedure.

5. SUMMARY AND CONCLUSIONS

In this study, the Bayesian adaptation technique using the MAP approach is derived, implemented and evaluated for optimally estimating the time-varying polynomial Gaussian mean functions in the trended HMM. The main conclusions can be summaried as follows. First, compared with speaker-independent models, the MAP adaptive training procedure achieves consistently better performance even with a single token in the adaptation data. Second, the trended HMM always outperforms the benchmark HMM (with only one exception where one training token is used in the speaker-dependent mode). When ten training tokens are used to obtain adaptive estimates for both the polynomial coefficients and the precisions, the recognizer achieves the best recognition rate of 92.1% (averaged over four speakers). Third, adaptation of polynomial coefficients only is shown to be better than adapting both polynomial coefficients and precision matrices when fewer than four adaptation tokens are used, while the opposite is true for more adaptation tokens. Comparisons of the alphabet classification performance and of data-fitting results demonstrate the effectiveness of the MAP-trained trended HMMs. A more detailed experiments with use of higher order polynomial functions (P greater than one) using the MAP approach is currently under way and will be reported in the near future.

REFERENCES

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

[2] L. Deng, M. Aksmanovic, X. Sun and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states", *IEEE Trans. on Speech and Audio Processing*, Vol.2, No. 4, pp. 507-520, October 1994.

[3] J. L. Gauvain and C. H. Lee, "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities", *Speech Communication*, No. 11, pp. 205-213, 1992.

[4] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No.2, pp. 291-298, April 1994.

[5] Q. Huo, C. Chan, and C.H. Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition", *IEEE Trans. on Speech and Audio Processing*, Vol.3, No. 5, pp. 334-345, Sept. 1995.

[6] C. Lee, C. H. Lin and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models", *IEEE Trans. on Signal Processing*, Vol. 39, No. 4, pp. 806-814, April 1991.

[7] C. Rathinavelu and L. Deng, "Speech trajectory discrimination using the minimum classification error learning", *IEEE Trans. on Speech and Audio Processing*, submitted for publication.

[8] C. Rathinavelu and L. Deng, "Use of generalized dynamic feature parameters for speech recognition: maximum likelihood and minimum classification error approaches", *IEEE Proc. ICASSP*, 1995, pp. 373-376.

[9] C. Rathinavelu and L. Deng, "HMM-based speech recognition using state-dependent, linear transforms on Mel-warped DFT features", *IEEE Proc. ICASSP*, 1996, pp. 9-12.