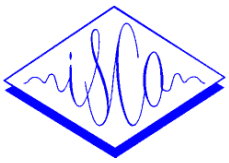

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

BAYESIAN LEARNING OF THE PARAMETERS OF DISCRETE AND TIED MIXTURE HMMs FOR SPEECH RECOGNITION

Qiang Huo†, *Chorkin Chan*† *and Chin-Hui Lee*‡

†Department of Computer Science
The University of Hong Kong, Hong Kong

‡Speech Research Department
AT&T Bell Laboratories, Murray Hill, NJ 07974

ABSTRACT

In this paper a theoretical framework for Bayesian adaptive learning of discrete HMM parameters is presented. Formulations of MAP and segmental MAP estimation of DHMM parameters are developed. An empirical Bayes method to estimate the hyperparameters of prior density based on the moment estimate is proposed. We applied the proposed method to speaker adaptation problems using a 26-word English alphabet vocabulary. Speaker-adaptive training algorithm is shown to be effective in improving the performance of both speaker-dependent and speaker-independent speech recognition problems. The method proposed in this paper will also be applicable to other problems in HMM training for speech recognition such as sequential or batch training, context adaptation, parameter smoothing, and so on.

Keywords: Bayesian learning; hidden Markov model; speaker adaptation.

INTRODUCTION

Recently, Bayesian adaptive learning of Hidden Markov Model (HMM) parameters has been proposed and adopted in a number of speech recognition applications. By assuming the HMM parameters to be random variables, the joint distribution of the observation vectors and the HMM parameters can be constructed. Given the joint distribution which prescribes the correlation between the observation vectors and the HMM parameters, statistical inference about the HMM parameters can be made based on a small amount of observation vectors. Therefore the Bayesian learning framework is an efficient way to handle sparse training data problem which exists in most statistical pattern recognition applications including speech recognition applications such as sequential training, speaker adaptation and parameter smoothing.

A theoretical framework of Bayesian learning was first proposed by Lee *et al* [7] for estimating the mean and covariance matrix parameters of a continuous density HMM (CDHMM) with a multivariate Gaussian state observation density. It was then extended to handle all the parameters of a CDHMM with mixture Gaussian state observation densities [3, 4, 5, 8]. Two algorithms for performing Bayesian adaptive learning, namely the forward-backward MAP algorithm [4] and the segmental MAP algorithm [7, 4, 8], have been developed and shown to be effective for many speech recognition applications. By using the same Bayesian learning framework, we have also extended this formulation to estimate parameters of discrete HMMs (DHMM) and semi-continuous HMMs (SCHMMs, also called tied-mixture HMMs) [6]. In addition to the two above MAP estimation algorithms, a computationally efficient segmental quasi-Bayes estimation algorithm for SCHMM parameters is developed. A new method for estimating the prior density parameters based on the moment estimate is also proposed.

In this paper, We study practical issues related to the use of Bayesian adaptive learning algorithms in estimating DHMM parameters for speaker adaptation (SA) applications. Other results related to SCHMMs will be reported in a separate paper due to the limitation on paper size in this conference. The rest of the paper is organized as follows: After a brief introduction of the concept of the Bayesian point estimation in Section 2, the formulation of MAP estimates for DHMM is presented in Section 3. In Section 4, an empirical Bayes method to estimate the hyperparameters of prior density based on the moment estimate is proposed. The experimental results are presented in Section 5. Finally, concluding remarks are presented in Section 6.

BAYESIAN POINT ESTIMATION

In the Bayesian approach, if θ is the unknown parameter vector to be estimated from a sequence of n observations $x_1, x_2, \cdots, x_n$, it is assumed that an investigator's prior knowledge about θ can be summarized in a prior probability density function (PDF) $p(\theta)$, with $\theta \in \Omega$, where Ω denotes an admissible region of the parameter space. By the use of Bayes' theorem, this information can be combined with the sample density function $p(x_1, x_2, \cdots, x_n|\theta)$ (which is the likelihood function if viewed as a function of θ) to yield a posterior PDF $p(\theta|x_1, x_2, \cdots, x_n)$. Such a PDF can be used to make inferences about the parameter θ:

$$p(\theta|x_1, x_2, \cdots, x_n) = \frac{p(x_1, x_2, \cdots, x_n|\theta)p(\theta)}{\int_\Omega p(x_1, x_2, \cdots, x_n|\theta)p(\theta)d\theta}. \quad (1)$$

Furthermore, if an investigator has a loss function which reflects the cost of an incorrect estimation, it is generally possible to obtain an estimate, say $\hat{\theta}$, which minimizes the posterior expected loss. Under a wide range of conditions, $\hat{\theta}$ will also be a function of the sample observations which minimizes the average risk. In this latter case, $\hat{\theta}$ is formally termed the Bayesian point estimator relative to the given loss function and prior PDF employed. It is well known that the mean of the posterior PDF is the Bayesian point estimator given that the loss function is quadratic while the mode of the posterior PDF is the one usually called modal or MAP (maximum *a posteriori*) estimator corresponding to the special zero-one loss function structure. Both of them are reasonable candidate of the point estimate of θ [11]. In particular, when the prior PDF $p(\theta)$ is constant over the parameter space Ω, the MAP estimator is the same as the classical maximum likelihood (ML) estimator.

MAP ESTIMATES FOR DHMM

Consider an N-state DHMM with parameter vector $\lambda = (\pi, A, B)$, where $\pi^t = [\pi_1, \pi_2, \cdots, \pi_N]$ is the initial state probability vector, $A = [a_{ij}], i, j = 1, 2, \cdots, N$, is the transition probability matrix, and $B = [b_{jk}], j = 1, \cdots, N, k = 1, \cdots, K$, b_{jk} is the probability of observing symbol v_k in state j. The observation symbol set is denoted as $V = \{v_1, v_2, \cdots, v_K\}$.

For simplicity, prior independence of π, A and B is assumed. The prior density for λ is then:

$$g(\lambda) = g(\pi) \cdot g(A) \cdot g(B). \tag{2}$$

Such an independence assumption may not be unduly restrictive. If the rows of π, A and B are assumed independently distributed *a priori*, and their densities assume the form of Dirichlet distributions (sometimes called multivariate beta PDF), then $g(\lambda)$ becomes a special case of the matrix beta PDF [9]:

$$g(\lambda) = K_c \cdot \prod_{i=1}^{N} \{\pi_i^{\eta_i - 1} \cdot (\prod_{j=1}^{N} a_{ij}^{\eta_{ij} - 1}) \cdot (\prod_{k=1}^{K} b_{ik}^{\nu_{ik} - 1})\} \tag{3}$$

where K_c is a normalizing factor. $\{\eta_i\}, \{\eta_{ij}\}, \{\nu_{ik}\}$ are sets of positive parameters for the prior PDF of π, A, B assigned by the investigator to represent his or her prior knowledge of the parameters.

For an observation sequence $\mathbf{x} = (x_1, x_2, \cdots, x_T)$, let $\mathbf{s} = (s_1, s_2, \cdots, s_T)$ be the unobserved state sequence. Given the observation sequence $\mathbf{x}$ and the prior density $g(\lambda)$, the MAP estimate of λ can be obtained by

$$\lambda_{MAP} = \arg\max_{\lambda} P(\mathbf{x}|\lambda) g(\lambda). \tag{4}$$

By viewing it as a missing data problem, as noted by Dempster *et al* [2], the EM (expectation-maximization) algorithm can be easily modified to produce this MAP estimate. The EM reestimation formulation for the three parameter sets π, A, B is as follows [6]:

$$\hat{\pi}_i = \frac{e_i + \eta_i - 1}{\sum_{i=1}^{N} e_i + \sum_{i=1}^{N} \eta_i - N} \qquad i = 1, 2, \cdots, N \tag{5}$$

$$\hat{a}_{ij} = \frac{c_{ij} + \eta_{ij} - 1}{\sum_{j=1}^{N} c_{ij} + \sum_{j=1}^{N} \eta_{ij} - N} \qquad i, j = 1, 2, \cdots, N \tag{6}$$

$$\hat{b}_{jk} = \frac{d_{jk} + \nu_{jk} - 1}{\sum_{k=1}^{K} d_{jk} + \sum_{k=1}^{K} \nu_{jk} - K} \qquad k = 1, 2, \cdots, K \tag{7}$$

where

$$e_i = Pr(s_1 = i|\mathbf{x}, \lambda) \tag{8}$$

$$c_{ij} = \sum_{t=1}^{T-1} Pr(s_t = i, s_{t+1} = j|\mathbf{x}, \lambda) \tag{9}$$

$$d_{jk} = \sum_{t: x_t \sim v_k} Pr(s_t = j, x_t \sim v_k|\mathbf{x}, \lambda) \tag{10}$$

and "$x_t \sim v_k$" denotes that the observation x_k is encoded as the symbol v_k. These terms can be efficiently computed by using the forward-backward algorithm [1]. Strictly speaking, to derive above reestimation formulas, three conditions must be obeyed: (1) $e_i + \eta_i > 1$, (2) $c_{ij} + \eta_{ij} > 1$ and (3) $d_{jk} + \nu_{jk} > 1$. This is usually the case in practice; otherwise, these simple formulas cannot be derived. Extension to the case of multiple independent observation sequences is straightforward and the readers are referred to [6].

If instead of maximizing $g(\lambda|\mathbf{x})$, the joint posterior density of parameters λ and state sequence $\mathbf{s}$, $g(\lambda, \mathbf{s}|\mathbf{x})$, is maximized. The so called segmental MAP estimate [7, 4] of λ can be derived [6]. By applying the Viterbi algorithm to the training data, apart from the most likely state sequences, the sets of observations associated with each HMM state are also available. Let $n_i^{(1)}$ denote the numbers of observations in state i at time $t = 1$, and n_{ij} be the transition count from state i to state j in the most likely state sequences. Furthermore, let f_{jk} denote the count of observing symbol v_k in state j. It is straight forward to show that the segmental MAP reestimation formulas are the same as equation (5) to (7) by replacing the e_i by $n_i^{(1)}$, c_{ij} by n_{ij} and d_{jk} by f_{jk}.

It can be seen that the above formulation computes each MAP estimate as a weighted sum of two terms, namely the corresponding prior parameter and the observed data. The weights are also recomputed iteratively and depend on the parameters and the data in a nonlinear fashion. The weight contribution of prior knowledge and new observed data are adjusted according to the desired applications.

PRIOR DENSITY ESTIMATION

Prior density estimation and choice of density parameters depend on the particular application of interest. In speaker adaptation application presented in this paper, prior density $g(\lambda|\varphi)$ represents the information of the variability of a certain model among the different speakers. Taking the empirical Bayes approach [10], the training data set $\mathbf{X}$ for estimating hyperparameters φ can be divided into different sets $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_Q$ correspond to Q different speakers or speaker groups. One may use these observation sets to estimate the corresponding HMMs $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_Q$ with the classical Baum-Welch algorithm,

and then pretend to view $\{\hat{\lambda}_i\}$ as the random observations with density $g(\lambda)$. In the case of DHMM where $g(\lambda)$ is assumed to have the form of equation (3), i.e. a matrix beta PDF, with the properties of the moments for matrix beta PDF, we have [6]

$$\eta_i = E(\pi_i)\{\frac{E(\pi_i)[1-E(\pi_i)]}{Var(\pi_i)} - 1\} \quad (11)$$

$$\eta_{ij} = E(a_{ij})\{\frac{E(a_{ij})[1-E(a_{ij})]}{Var(a_{ij})} - 1\} \quad (12)$$

$$\nu_{ik} = E(b_{ik})\{\frac{E(b_{ik})[1-E(b_{ik})]}{Var(b_{ik})} - 1\}. \quad (13)$$

Replacing $E(\pi_i)$, $Var(\pi_i)$, $E(a_{ij})$, $Var(a_{ij})$, $E(b_{ik})$, $Var(b_{ik})$ by their corresponding sample moments of random observations $\{\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_Q\}$, the moment estimates of η_i, η_{ij}, ν_{ik} are thus obtained.

When enough training data are available, the above method of moment will lead to a reasonable estimate of hyperparameters φ. This estimate may be improved by the following iterative scheme [4]: starting with an initial estimate $\varphi^{(m)}$, get the MAP estimates $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_Q$ from $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_Q$ with methods presented in the previous section; and then an improved $\varphi^{(m+1)}$ can be obtained by using the above method of moment.

We also tried an *ad hoc* method to estimate the hyperparameters of prior density. Let $\hat{e}_i$, $\hat{c}_{ij}$ and $\hat{d}_{jk}$ be the respective estimated counts of related events at the last iteration of SI training. These counts are divided by the number of training tokens for each word and then plus one. The hyperparameters are then set to these values.

ADAPTATION EXPERIMENTS

In this paper, we choose the 26-word English alphabet as the vocabulary for all experiments. Two databases are used for evaluating the adaptation algorithms, *viz.*, the OGI ISOLET and the TI46 corpora. These two databases were recorded at two separate sites with a time gap of 10 years. The sampling rates and quantization precisions are 16 KHz with 16-bit quantization and 12.5 KHz with 12-bit quantization respectively. They have therefore very different acoustic characteristics. The speech data in the two corpora were lowpass-filtered at 3.3KHz and down-sampled to 8 KHz so that hopefully, they will become more compatible to each other. To simplify filter design problem, we adopt a multistage implementation to convert the sampling rate from 12.5 KHz to 8 KHz by first up-sampling to 50 KHz, then to 200 KHz, and then down-sampling to 40 KHz and finally to 8 KHz. The feature vectors used in this study consist of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30ms frame length and a 10ms frame shift. For speaker independent (SI) training and the prior density estimation, The OGI ISOLET database consisting of 150 talkers (75 females and 75 males) was used. Each talker uttered each of the letters twice. Since there are not enough data to estimate a model for each letter for each speaker, speaker clustering based on vector quantization was performed to obtain 16 speaker clusters from which 16 sets of models needed to obtain the moment estimates of the prior parameters were derived. For speaker dependent (SD) or adaptive training and testing, the English alphabet subset of the TI46 isolated word corpus was used. It was produced by 16 talkers, 8 females and 8 males. Each person uttered each of the letters 26 times, 10 of them were used for SD/SA training and the remaining 16 tokens for testing. Only the experimental results using the test data from one female ("f1") and one male ("m2") (they are selected in random) are reported here.

The speaker clustering process began from two natural male/female groups. The clustering algorithm is as follows:

1. View all male speakers as one group and all female speakers as another group. Generate respectively two codebooks with size 256.
2. Do "speaker classification" respect to each codebook of each speaker group with VQ method.
3. Reformulate the codebook for each speaker group with the speaker classification result in step 2.
4. If speaker classification is stable or a predefined maximum number of iterations reached or the variation of the total quantization error (for all speakers) is less than a predifined threshold, then go to step 5; else go to step 2.
5. If predefined number of speaker groups reached, stop; else go to step 6.
6. Split the codebook by a simple perturbation method, go to step 2.

The criterion used here is the "total quantization error", so the speakers in each group will not be even. We used all the training utterances of each speaker for "speaker classification". Speaker clustering results show that most clusters are dominated by male or female and this seems a very positive sign for our clustering method.

Throughout the experiments, each of the 26 letters in the vocabulary was modeled by a single left-to-right 5-state DHMM with arbitrary state skipping. A 256-vector codebook was generated from ISOLET corpus by using LBG algorithm with Euclidean distortion measure and was used in all subsequent experiments. The SI/SD word models were trained by using the standard Baum-Welch algorithm [1] and the SA ones were obtained by using the MAP estimates presented in Section 3. In recognition, the decision rule assigns the unknown word to the vocabulary word whose model has the highest forward-backward probability.

The recognition results for the female and male talkers based on SA models are listed in Tables 1 and 2 respectively. "SA1" corresponds to the experiments with the *ad hoc* prior parameters and "SA2" refers to the ones with prior parameters estimated by the method of moment. For comparison purposes, the word recognition rates for the SD models are also reported.

Tables 1 and 2 clearly show that the regular MLE training procedure ("SD") is inadequate when the amount of

Table 1: Summary of results for female speaker f1
(SI recognition rate: 53.37)

tokens	SD	SA1	SA2
1	50.48	61.54	58.65
2	54.57	63.70	62.50
3	63.46	67.79	67.55
4	64.18	68.75	66.35
5	67.07	68.51	67.07
6	65.38	67.79	65.87
7	65.38	66.11	67.07
8	67.07	65.38	68.75
9	68.27	67.07	69.95
10	70.91	68.75	70.43

Table 2: Summary of results for male speaker m2
(SI recognition rate: 58.65)

tokens	SD	SA1	SA2
1	53.37	68.51	65.38
2	63.94	74.52	72.60
3	69.71	76.68	75.24
4	74.28	80.05	78.12
5	74.52	80.29	79.09
6	74.52	81.49	79.81
7	80.77	81.49	81.73
8	79.81	83.89	82.21
9	79.09	84.62	83.17
10	82.21	86.06	83.89

available training data is insufficient. The performance for "SD" improves as the number of speaker specific training tokens increases. The SD performance is inferior to the SI performance when only one token per letter is available for training. But the relative low SI performance also shows the serious mismatch between SI training set and SD testing set. The results here show that speaker adaptation can be used to reduce this mismatch. SA is not only doing speaker adaptation but also in some sense doing signal matching. When one additional training token is available for speaker adaptation, the SA models always outperform the SI models. Much better performance can also be achieved when more adaptation training tokens are incorporated. It is noted that when using the same amount of training data, SA training outperforms SD training in most of the cases tested. This implies that SA training utilizes training data more effectively than SD training, especially in cases of insufficient training data. As expected, the SA performance quickly becomes equivalent to the SD performance when the number of adaptive training tokens increases. It is also noted that the performance of "SA1" is in most cases better than that of "SA2". The hyperparameters of prior distribution estimated with *ad hoc* method seem more robust in our experiments here than the ones estimated with the method of moment which may suffer more from the sparse training data problem. The low SI recognition performance may be improved by using more SI training data. This in turn will produce better prior density estimation. Thus a better SA performance is expected.

SUMMARY

In this paper a theoretical framework for Bayesian adaptive learning of discrete HMM parameters is presented. Formulations of MAP and segmental MAP estimation of DHMM parameters are developed. An empirical Bayes method to estimate the hyperparameters of prior density based on the moment estimate is proposed. We applied the proposed method to speaker adaptation problems using a 26-word English alphabet vocabulary. When compared with recognition results obtained using SI models, the adaptive training procedure achieved better performance. Compared with SD training, speaker adaptation achieved an equal or better performance with the same amount of training/adaptation data in most cases. The method proposed in this paper will also be applicable to other HMM training problems for speech recognition such as sequential or batch training, context adaptation, parameter smoothing, and so on.

REFERENCES

[1] L.E. Baum (1972), "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," *Inequalities*, Vol. 3, pp. 1-8.

[2] A.P. Dempster, N.M. Laird and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, Vol. 39, pp. 1-38.

[3] J.-L. Gauvain and C.-H. Lee (1991), "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, pp. 272-277, Feb. 1991.

[4] J.-L. Gauvain and C.-H. Lee (1992), "MAP Estimation of Continuous Density HMM: Theory and Applications," *Proc. DARPA Speech and Natural Language Workshop*, Arden House, pp. 185-190.

[5] J.-L. Gauvain and C.-H. Lee (1992), "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, Vol. 11, Nos. 2-3, pp. 205-213.

[6] Q. Huo and C. Chan (1992), "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition," *Technical Report*, TR-92-08, Department of Computer Science, University of Hong Kong.

[7] C.-H. Lee, C.-H. Lin and B.-H. Juang (1991), "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on ASSP*, Vol. ASSP-39, No. 4, pp.806-814.

[8] C.-H. Lee and J.-L. Gauvain (1993), "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proc. ICASSP-93*, Minneapolis, pp. II-588-591.

[9] J. J. Martin (1967), *Bayesian Decision Problems and Markov Chains*, Wiley, New York.

[10] H. Robbins (1964), "The Empirical Bayes Approach to Statistical Decision Problems," *Ann. Math. Statist.*, Vol. 35, pp. 1-20.

[11] R. L. Winkler (1972), *Introduction to Bayesian Inference and Decision*, Holt, Rinehart and Winston, Inc.