IMPROVED ACOUSTIC MODELING WITH BAYESIAN LEARNING

Jean-Luc Gauvain† and Chin-Hui Lee

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

ABSTRACT

We study the use of *Bayesian learning* for the estimation of the parameters of a multivariate mixture Gaussian density. For speech recognition algorithms based on the *continuous density hidden Markov model* (CDHMM) framework, Bayesian learning serves as a unified approach for the following four applications, namely *parameter smoothing*, *speaker adaptation* (SA), *speaker group modeling*, and *corrective training* (CT). In our approach, we use Bayesian learning techniques to incorporate prior knowledge into the CDHMM training process in the form of prior densities of the HMM parameters. The theoretical basis for this procedure is presented. All the four applications have been evaluated. Experimental results on the TI connected digit task and the Naval Resource Management (RM) task are provided to show the effectiveness of Bayesian adaptation of CDHMM.

INTRODUCTION

When training sub-word units for continuous speech recognition using probabilistic modeling techniques, we are faced with the general problem of sparse training data. This limits the effectiveness of conventional *maximum likelihood* (ML) approaches. The sparse training data problem can not always be solved by the acquisition of more training data. For example, in the case of rapid adaptation to new speakers or environments, the amount of data available for adaptation is usually much less than what is needed to achieve good performance for speaker-dependent applications.

Our solution to the problem is to use Bayesian learning to incorporate prior knowledge into the HMM training process. This information is characterized in terms of prior densities of the HMM parameters. Such an approach was shown to be effective for speaker adaptation in isolated word recognition of a 39-word, English alpha-digit vocabulary where adaptation involved only the parameters of a diagonal covariance Gaussian state observation density of whole-word HMM's [1]. In this paper, Bayesian adaptation is extended to handle parameters of mixture Gaussian densities for sub-word HMM's and applied to various recognition problems.

† Jean-Luc Gauvain is with the Speech Communication Group at LIMSI/CNRS, Orsay, France. This study was completed while he was visiting Bell Labs. in 1990-1991.

MAP ESTIMATE OF CDHMM

The difference between maximum likelihood estimation (MLE) and Bayesian learning lies in the assumption of an appropriate prior distribution for the parameters to be estimated. Given a sequence of n observations $\mathbf{X} = \{x_1, \cdots, x_n\}$ with a density $P(\mathbf{X} \mid \theta)$, we are interested in estimating the parameter vector θ. Let $P(\theta)$ be the prior density, one way to estimate θ is to assume that θ is a random vector and obtain the *maximum a posteriori* (MAP) estimate which corresponds to the mode of the posterior density $P(\theta \mid \mathbf{X})$, i.e.

$$\theta_{MAP} = \arg\max_{\theta} P(\theta \mid \mathbf{X}) = \arg\max_{\theta} P(\mathbf{X} \mid \theta)\, P(\theta) \quad . \quad (1)$$

On the other hand, if θ is assumed to be a fixed but unknown parameter vector, then there is no prior knowledge about θ. This is equivalent to assuming a non-informative prior, i.e. $P(\theta)$ = constant. Eq. (1) is now the familiar ML formulation.

MAP of Gaussian mixture densities

MAP formulation for parameters of multivariate Gaussian densities is well-known [2]. MAP estimation of the parameters of the Gaussian mixture density has also been formulated [3-4]. To simplify the notation, we assume here a mixture of univariate normal densities:

$$P(x \mid \theta) = \sum_{k=1}^{K} \omega_k\, N(x \mid m_k, r_k) \quad (2)$$

where $\theta = (\omega_k, m_k, r_k)_{k=1}^{K}$. There exists no sufficient statistic of fixed dimension for θ and hence no conjugate density. We propose to use a prior joint density which is the product of a Dirichlet density and normal-gamma densities:

$$P(\theta) \propto \prod_{k=1}^{K} \omega_k^{\lambda_k} r_k^{1/2} \exp\left(-\frac{\tau_k r_k}{2}(m_k-\mu_k)^2\right) r_k^{\alpha_k - 1} \exp(-\beta_k r_k) \quad . \quad (3)$$

The choice of such a prior density can be justified by the fact that the Dirichlet density is the conjugate density of the multinomial density (for the mixture weights) and the normal-gamma density is the conjugate density of the normal distribution (for the mean and the precision parameters). It can be shown [3] that the classical EM algorithm [5] can be used to find the MAP estimate. Define the normalized weight

$$c_{ik} \equiv \frac{\omega_k N(x_i \mid m_k, r_k)}{P(x_i \mid \theta)} \quad , \quad (4)$$

the MAP estimate for θ is solved by the following equations:

0-7803-0532-9/92 $3.00 © 1992 IEEE

$$\omega_k' = \frac{\lambda_k + \sum_{i=1}^{n} c_{ik}}{n + \sum_{k=1}^{K} \lambda_k} \quad (5)$$

$$m_k' = \frac{\tau_k \mu_k + \sum_{i=1}^{n} c_{ik} x_i}{\tau_k + \sum_{i=1}^{n} c_{ik}} \quad (6)$$

$$r_k' = \frac{2\alpha_k - 1 + \sum_{i=1}^{n} c_{ik}}{2\beta_k + \sum_{i=1}^{n} c_{ik}(x_i - m_k')^2 + \tau_k(\mu_k - m_k')^2} . \quad (7)$$

The above iterative re-estimation equation can be shown to converge to a solution that locally maximizes $P(\theta \mid X)$ [3]. It can also be shown that when more adaptation data are used ($n \to \infty$), the MAP estimate converges to the MLE asymptotically [3]. By using a non-informative prior density (i.e. an improper distribution with $\lambda_k=0$, $\tau_k=0$, $\alpha_k=1/2$, and $\beta_k=0$) the EM re-estimation formulas to compute the MLE of the mixture parameters can be recognized.

MAP estimates for HMM

The above procedure can be applied to estimate the parameters of an HMM state given a set of n observations $X = \{x_1, \ldots, x_n\}$ assumed to be independently drawn from the mixture Gaussian state distribution. Following the scheme of the segmental k-means algorithm to estimate the parameters of an HMM, first the Viterbi algorithm is used to segment the training data X into sets of observations associated with each HMM state and then the MAP estimate procedure is applied to each state. This *segmental MAP algorithm* was originally proposed in [1] and was recently extended [3-4] to handle mixture Gaussian state densities.

A Bayesian version of the Baum-Welch algorithm can also be designed [3]. As in the case of MLE, one simply replaces c_{ik} in eq. (4) by c_{ijk} in the re-estimation formulas and applies the summations over all the observations for each state s_j:

$$c_{ijk} \equiv \gamma_{ij} \frac{\omega_{jk} N(x_i \mid m_{jk}, r_{jk})}{P(x_i \mid \theta_j)} \quad (8)$$

where $\theta_j = (\omega_{jk}, m_{jk}, r_{jk})_{k=1}^{K}$ is the mixture density parameter vector for state s_j, γ_{ij} is the probability of being in state s_j at time i, given that the model generates X. For the segmental MAP approach γ_{ij} is simply equal to 0 or 1 [3].

Prior parameter estimation

In all applications of Bayesian learning, the prior density parameters were estimated along with the estimation of the SI model parameters using the segmental k-means algorithm. In our formulation, more parameters are needed for the prior density than for the mixture density itself. It is therefore of interest to use *tied parameters* for the prior densities. Information about the variability to be modeled with the prior densities was associated with each frame of the SI training data. This information was simply represented by a class number which can be the speaker ID, the speaker sex or the phonetic context. The HMM parameters for each class C_l given the mixture component were then computed. For the experiments reported in this paper, the prior density parameters were estimated as follows:

$$\alpha_{jk} = (\tau_{jk}+1)/2 \quad (9)$$

$$\beta_{jk} = \tau_{jk}/2r_{jk} \quad (10)$$

$$\mu_{jk} = m_{jk} \quad (11)$$

$$\lambda_{jk} = \omega_{jk} \sum_{k=1}^{K} \tau_{jk} \quad (12)$$

$$\tau_{jk} = \frac{p + \sum_l c_{jkl}}{\sum_l [c_{jkl}(y_{jkl} - m_{jk})'(\sum_k \omega_{jk} r_{jk}^{-1})^{-1}(y_{jkl} - m_{jk})]} \quad (13)$$

where ω_{jk}, m_{jk} and r_{jk} are the SI HMM parameters for each state s_j and each mixture component k in state s_j, and m_{jk} and r_{jk} are vectors of dimension p. The class mean vector y_{jkl} is equal to $\sum_i c_{ijkl} x_i / c_{jkl}$, where c_{ijkl} is defined as $c_{ijkl}=c_{ijk}$ if $x_i \in C_l$ and $c_{ijkl}=0$ otherwise, and $c_{jkl}=\sum_i c_{ijkl}$.

EXPERIMENTAL SETUP

All the experiments presented in this paper used various sets of context-independent (CI) and context-dependent (CD) phone models. Each model is a left-to-right HMM with Gaussian mixture state observation densities. Diagonal covariance matrices are used and it is assumed that the transition probabilities are fixed and known [6]. As described in Lee *et al* [7], a 38-dimensional feature vector composed of 12 LPC-derived cepstrum coefficients, 12 delta cepstrum coefficients, the delta log energy, 12 delta-delta cepstrum coefficients, and the delta-delta log energy is used.

For RM evaluation, results are reported using the standard word-pair grammar with a perplexity of about 60. For digit recognition, we use the TI/NIST connect digit database. Both databases were down-sampled to telephone bandwidth. All four recognition applications based on Bayesian learning have been tested and are discussed in details in the following.

MODEL SMOOTHING AND ADAPTATION

CD model smoothing was evaluated in [4] and it was found that the word error rates was reduced by 10% compared with the results obtained without model smoothing. Speaker adaptation was tested on the JUN90 data with 1 minute and 2 minutes of speaker-specific adaptation data. A 16% and 31% reduction in word error were obtained compared to the SI results [4]. Sex-dependent modeling can be also achieved similar to the speaker adaptation case by replacing the speaker labels with the gender labels for each training sentence. On the FEB91 test, using Bayesian learning for HMM parameter smoothing and a combined male/female (M/F) modeling, a 21% word error reduction compared to the baseline system results was obtained [4].

To further investigate the correlation observed between speaker sex and performance, the experiment has been carried out on the FEB91-SD test material which includes data from 7 male and 5 female speakers. For studying issues related to both SD and SA training, a set of 47 CI phone models was used. Two, five and thirty minutes of the SD training data were used for training and adaptation. The SI, SA, SD and M/F results (in word error rate) are summarized in Table 1.

Training data	0 min	2 min	5 min	30 min
SD	—	31.5	12.1	3.5
SA (SI)	13.9	8.7	6.9	3.4
SA (M/F)	11.5	7.5	6.0	3.5

Table 1: Summary of SI, SA, SD and M/F results (FEB91-SD)

It is noted that the SD word error rates was 31.5% for 2 minutes of training data. The word error rates for the SI model was 13.9% which is comparable with 5 minutes of SD training. The performance of SA model in general is better than the SD model when an equal amount of training (or adaptation) data were used. When all the training data are used (600 sentences or roughly 30 minutes), the SA and SD results were comparable which is consistent with the Bayesian formulation in equations (5)-(7) that the MAP estimate converges to the MLE asymptotically. Compared to the SI results, the word error reduction was 37% with two minutes of adaptation data, comparable to the improvement observed on the JUN90 data with CD unit models. Similar to the findings in our previous experiments [4], the improvement is larger for the female speakers (51%) than for the male speakers (22%).

Speaker adaptation can also be performed on sex-dependent models. Starting from the combined M/F model (0 minute training), the performance improved as more speaker-specific data were used for adaptation. When compared with the SA(SI) results shown in Table 1, it is noted that speaker adaptation is always more effective when initializing with sex-dependent seed models. The best result on the FEB91-SD test, using 47 CI unit models, was 96.6% word accuracy when all 600 utterances were used for adaptation.

CORRECTIVE TRAINING

Bayesian learning provides a scheme for model adaptation which can also be used for corrective training. In order to do so, the state segmentation step of the segmental MAP algorithm was modified to obtain not only the frame/state association for the *sentence model* states but also for the states corresponding to the model of all the possible sentences (*general model*). In the re-estimation formulas, the values c_{ijk} for each state s_j are evaluated using equation (8), such that γ_{ij} is equal to 1 in the sentence model and to -1 in the general model. While convergence is not guaranteed, in practice it was found that by using large values for $\tau_{jk}(\approx 200)$, the number of training sentence errors decreased after each iteration until convergence. It should be noted that if the Viterbi alignment is replaced by the Baum-Welch algorithm we obtain a corrective training algorithm for CDHMM's very similar to the recently proposed *corrective MMIE training* algorithm [8].

Corrective training was evaluated on both the TI/NIST SI connected digit task and the RM task. Only the Gaussian mean vectors and the mixture weights were corrected. For the connected digit task a set of 21 phonetic HMMs were trained on the 8565 digit strings. Results on the 8578 test strings are given in Tables 2 and 3. Both string accuracy and string error count are listed for comparison. The CT-16 results in Table 2 were obtained with 8 iterations of corrective training while the CT-32 results in Table 3 were based on only 3 iterations of adaptation. Iteration 0 represents the results obtained with MLE without using corrective training. Here one full iteration of corrective training is implemented as one recognition run which produces a set of *"new"* training strings (i.e. errors and/or barely correct strings) followed by ten iterations of Bayesian adaptation using the data of these strings. String error rates of 1.5% and 1.3% were obtained with 16 and 32 mixture components per state respectively. These represent 25% and 15% word error reductions compared to the cases without corrective training (MLE-16 and MLE-32).

Iteration Number	Training Accuracy (String Error Count)	Testing Accuracy (String Error Count)
0	98.4 (134)	98.0 (168)
1	99.0 (82)	98.4 (139)
2	99.3 (60)	98.4 (138)
4	99.5 (14)	98.5 (129)
8	99.8 (18)	98.6 (122)

Table 2: Corrective training results on the TI-digit task (21 CI models with 16 mixture components per state)

Iteration Number	Training Accuracy (String Error Count)	Testing Accuracy (String Error Count)
0	99.2 (67)	98.5 (126)
1	99.5 (44)	98.6 (117)
2	99.6 (32)	98.7 (110)
3	99.7 (29)	98.7 (111)

Table 3: Corrective training results on the TI-digit task (21 CI models with 32 mixture components per state)

The corrective training procedure is also effective for continuous sentence recognition. Table 4 gives results for the RM task, using 47 models with 32 mixture components. The CT-32 corrective training assumes a fixed beam width. Since the number of string errors was small in the training set, the effective amount of data for corrective training is rather limited. To increase the amount, we propose the use of a smaller beam width in recognizing strings in the training set. It was observed that this *improved corrective training* (ICT-32) procedure not only reduces the error rate in training but also increases the separation between the correct string and the other competing strings. The number of training errors also increased as predicted. The regular and the improved corrective training gave an average word error rate reduction of 15% and 20% respectively on the test data.

Test Set	MLE-32	CT-32	ICT-32
TRAIN	7.7	1.8	3.1
FEB89	11.9	10.2	8.9
OCT89	11.5	9.8	8.9
JUN90	10.2	8.8	8.1
FEB91	11.4	10.3	10.2
FEB91-SD	13.9	11.3	11.0
Overall Test	11.8	10.1	9.4

Table 4: Corrective training results on the RM task (47 CI models with 32 mixture components per state)

PDF SMOOTHING

We have shown that Bayesian learning can be used for context-dependent model smoothing. This approach can be seen either as a way to add extra constraints to the model parameters so as to reduce the effect of insufficient training data, or it can be seen as an "interpolation" between two sets of parameter estimates: one corresponding to the desired model and the other to a smaller model which can be trained using MLE on the same data. Instead of defining a reduced parameter set by removing the context dependency, we can alternatively reduce the mixture size of the observation densities and use a single Gaussian per state in the smaller model. Cast in the Bayesian learning framework, this implies that the same marginal prior density is used for all the components of a given mixture. Variance clipping can also be viewed as a MAP estimation technique with a uniform prior density constrained by a maximum (positive) value for the precision parameters [1]. However, this does not have the appealing interpolation capability of the conjugate priors.

We experimented with this pdf smoothing approach on the TI digit and RM databases. A set of 213 CD phone models with 32 mixture components (213 CD-32) for the TI digits and a set of 2421 CD phone models with 16 mixture components (2421 CD-16) for RM were used for evaluation. Results are listed in Tables 5 and 6 for MLE training, MLE with variance clipping (MLE+VC), and MAP estimation with pdf smoothing. In Table 5, string accuracy (SACC) and word accuracy (WACC) are given. When compared with the variance clipping scheme, it can be seen that the MAP estimate reduces the number of string errors from 101 to 76, which represents a 25% string error reduction. Using pdf smoothing on the 213 CD models, the string accuracy was 99.1% which is the *best* result reported on this task.

	SACC (Strings Correct)	WACC
MLE	98.7 (8464)	99.6
MLE+VC	98.8 (8477)	99.6
MAP	99.1 (8502)	99.7

Table 5: TI test results for pdf smoothing (using 213 inter-word CD-32 models)

	FEB89	OCT89	JUN90	FEB91
MLE	93.3	92.5	92.1	92.9
MLE+VC	95.0	95.0	94.8	95.9
MAP(SI)	95.0	95.5	95.0	96.2
MAP(M/F)	95.2	96.2	95.2	96.7

Table 6: RM test results for pdf smoothing (using 2421 inter-word CD-16 models)

As for the RM testing shown in Table 6, we also observed a consistent improvement over the variance clipping scheme (MLE+VC) when pdf smoothing is applied. When combined with sex-dependent modeling, the MAP(M/F) scheme shown in Table 6 gives a slight improvement over the MAP(SI) results. The 96.7% word accuracy on FEB91 test data represents the best results we have achieved on this set.

SUMMARY

A study on the use of Bayesian learning of CDHMM parameters has been carried out. The theoretical framework for training HMM's with mixture Gaussian state observation densities was presented. It was shown that Bayesian learning serves as a unified approach for parameter smoothing, speaker adaptation, speaker group modeling and corrective training.

Tested on the RM task, encouraging results have been obtained for all four applications. For speaker adaptation, a 37% word error reduction over the best SI results was obtained on the FEB91-SD test with 2 minutes of speaker-specific training data. It was also found that speaker adaptation based on sex-dependent models gave a better result than that obtained with speaker-independent seed. Compared with speaker-dependent training, speaker adaptation achieved a better performance based on the same amount of training/adaptation data. Corrective training was also found effective in reducing word errors by 15-25%. It is noted that corrective training helps more with models with a smaller number of parameters. The best results on SI RM testing was obtained with pdf smoothing and sex-dependent modeling. For the FEB91 SI test, we achieved a word accuracy of 96.7% using the word-pair grammar. The overall average for the four SI sets was 95.8% word accuracy.

Only pdf smoothing and corrective training were applied to the TI/NIST connected digit task. It was found that corrective training is effective for improving both CI and CD models. When more models were estimated (213 CD models), pdf smoothing provided a robust model that gave a 99.1% string accuracy on the testing data. This represents the best performance reported on this database.

REFERENCE

[1] C.-H. 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 ASSP*,April 1991.

[2] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, New York, 1970.

[3] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *to be submitted.*

[4] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Proc. EuroSpeech 91*, Genova, 1991.

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

[6] C.-H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, "Acoustic modeling for large vocabulary speech recognition", *Computer Speech and Language*, 4, pp. 127-165, 1990.

[7] C.-H. Lee, *et al*, "Improved Acoustic Modeling for Continuous Speech Recognition", *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990.

[8] Y. Normandin and D. Morgera, "An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary, Continuous Speech Recognition", *Proc. ICASSP91*, pp. 537-540, May 1991.