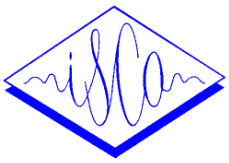

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities

Jean-Luc Gauvain[1] *and Chin-Hui Lee*

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

ABSTRACT

An investigation into the use of *Bayesian learning* of the parameters of a multivariate Gaussian mixture density has been carried out. In a *continuous density hidden Markov model* (CDHMM) framework, Bayesian learning serves as a unified approach for parameter smoothing, speaker adaptation, speaker clustering and corrective training. The goal is to enhance model robustness in a CDHMM-based speech recognition system so as to improve performance. Our approach is to use Bayesian learning to incorporate prior knowledge into the training process in the form of prior densities of the HMM parameters. The theoretical basis for this procedure is presented and results applying it to HMM parameter smoothing, speaker adaptation, speaker clustering, and corrective training are given.

The following word error reductions were observed on the DARPA RM task: 10% with HMM parameter smoothing, 31% for speaker adaptation with 2 minutes of speaker specific training data, and 15% with sex-dependent modeling.

INTRODUCTION

When training sub-word units for continuous speech recognition using probabilistic methods, we are faced with the general problem of sparse training data. This limits the effectiveness of the conventional *maximum likelihood* approach. 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.

Techniques used to alleviate the insufficient training data problem include probability density function (pdf) smoothing, model interpolation, corrective training, and parameter sharing. The first three techniques have been developed for HMM with discrete pdfs and cannot be directly extended to the general case of *continuous density hidden Markov model* (CDHMM). For example, the classical scheme of model interpolation [3] can be applied to CDHMM only if tied mixture HMMs or an increased number of mixture components are used.

Our solution to the problem is to use Bayesian learning to incorporate prior knowledge into the CDHMM training process [10]. The prior information consists of prior densities of the HMM parameters. Such an approach was shown to be effective for speaker adaptation in isolated word recognition where adaptation involved only the parameters of a multivariate Gaussian state observation density of whole-word HMMs [9]. In this paper, Bayesian adaptation is extended to handle parameters of mixtures of Gaussian densities. The theoretical basis for Bayesian learning of parameters of a multivariate Gaussian mixture density for HMM is developed. In a CDHMM framework, Bayesian learning serves as a unified approach for parameter smoothing, speaker adaptation, speaker clustering, and corrective training.

[1]Jean-Luc Gauvain is on leave from the Speech Communication Group at LIMSI/CNRS, Orsay, France.

MAP ESTIMATE OF CDHMM

The difference between maximum likelihood (ML) estimation and Bayesian learning lies in the assumption of an appropriate prior distribution of the parameters to be estimated. If θ is the parameter vector to be estimated from a sequence of n observations $x_1, ..., x_n$, given a prior density $P(\theta)$, then one estimate for θ is the maximum a posteriori (MAP) estimate which corresponds to the mode of the posterior density,

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(x_1, ..., x_n|\theta)P(\theta) \qquad (1)$$

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

Given the MAP formulation in Equation 1 two problems remain: the choice of the prior distribution family and the effective evaluation of the maximum a posteriori. In fact these two problems are closely related, since the choice of an appropriate prior distribution can greatly simplify the estimation of the maximum a posteriori. The most practical choice is to use conjugate densities which are related to the existence of a sufficient statistic of a fixed dimension [1]. If the observation density possesses such a statistic s and if $g(\theta|s,n)$ is the associated kernel density, MAP estimation is reduced to the evaluation of the mode of the product $g(\theta|s,n)P(\theta)$. In addition, if the prior density is chosen in the same family as the kernel density, $P(\theta) = g(\theta|t,m)$, the previous product is simply equal to $g(\theta|u, m+n)$ since the kernel density family is closed under multiplication. In this case, the MAP estimation problem is closely related to the MLE problem - finding the mode of the kernel density. In fact, $g(\theta|u, m+n)$ can be seen as the kernel of the likelihood of a sequence of $m+n$ observations.

When there is no sufficient statistic of a fixed dimension, MAP estimation, like ML estimation, has no analytical solution. However, the problems are still very similar. For the general case of mixture densities of the exponential family, we propose to use a product of kernel densities of the exponential family assuming independence between the parameters of the mixture com-

ponents in the joint prior density. To simplify solving Equation 1, we can restrict our choice to a product of a Dirichlet density and kernel densities of the mixture exponential density, $P(\theta) \propto \prod_{k=1}^{K} \omega_k^{m_k} g(\theta_k|t_k, m_k)$, where K is the number of mixture components and ω_k's are the mixture weights. However, this choice may be too restrictive to adequately represent the real prior information and in practice it may be of interest to choose a slightly larger family.

In the following subsections, we focus our attention on the cases of normal density and mixture of normal densities for two reasons: solutions for the MLE problem are well known and we are using CDHMM based on mixtures of normal densities.

Normal density case

Bayesian learning of a normal density is well known [1]. If $x_1, ..., x_n$ is a random sample from $\mathcal{N}(x|m, r)$, where m and r are respectively the mean and the precision (reciprocal of the variance), and if $P(m, r)$ is a normal-gamma prior density, $P(m, r) \propto r^{1/2} \exp(-\frac{\tau r}{2}(m-\mu)^2) r^{\alpha-1} \exp(-\beta r)$, the joint posterior density is also a normal-gamma density whose parameters $\hat{\mu}$, $\hat{\beta}$, $\hat{\alpha}$ and $\hat{\tau}$ may be directly obtained from the prior parameters and the sample mean and variance. The MAP estimates of m and r are respectively $\hat{\mu}$ and $\frac{\hat{\alpha}-0.5}{\hat{\beta}}$.

This approach has been widely used for sequential learning of the mean vectors of feature- and template-based recognizers, see for example [4, 7]. Ferretti and Scarci [8] used Bayesian estimation of mean vectors to build speaker-specific codebooks in an HMM framework. In all these cases, the precision parameter was assumed to be known and the prior density limited to a Gaussian.

Brown *et al.* [5] used Bayesian estimation for speaker adaptation of CDHMM parameters in a connected digit recognizer. More recently Lee *et al.* [9] investigated various training schemes of Gaussian mean and variance parameters using normal-gamma prior densities for speaker adaptation. They showed that on the alpha-digit vocabulary, with a small amount of speaker specific data (1 to 3 utterances of each word), the MAP estimates gave better results than the ML estimates.

Mixture of normal densities

For this study we used CDHMM where the state observation densities are mixtures of multivariate normal densities [11, 12]. However, to simplify the presentation of our approach, we assume here a mixture of univariate normal densities:

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

where $\theta = (\omega_1, ..., \omega_K, m_1, ..., m_K, r_1, ..., r_K)$. For such a density there exists no sufficient statistic of fixed dimension for θ and therefore no conjugate distribution.

We propose to use a joint prior density which is the product of a Dirichlet density and gamma-normal densities:

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

The choice of such a prior density can be justified by the fact that the Dirichlet density is the conjugate distribution of the multinomial distribution (for the mixture weights) and the gamma-normal density is the conjugate density of the normal distribution (for the mean and the precision parameters). The problem now is to find the mode of the joint posterior density.

If we assume the following regularity conditions, 1) $\lambda_k = \tau_k$ and 2) $\alpha_k = (\tau_k + 1)/2$, then the posterior density $P(\theta|x_1, ..., x_n)$ can be seen as the likelihood of a stochastically independent union of a set of $\sum_{k=1}^{K} \tau_k$ categorized observations and a set of n uncategorized observations. (A mixture of K densities can be interpreted as the density of a mixture of K populations, and an observation is said to be categorized if its population of origin is known with probability 1.) This suggests the use of the E.M. algorithm [2] to find the maximum a posteriori. The following recursive formulas estimate the MAP of the 3 parameter sets.

$$c_{ik} \triangleq \frac{\omega_k \mathcal{N}(x_i|m_k, r_k)}{P(x_i|\theta)} \tag{4}$$

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

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

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

By using a non-informative prior density (i.e. an improper prior with $\lambda_k = 0$, $\tau_k = 0$, $\alpha_k = 1/2$, and $\beta_k = 0$) the classical E.M. reestimation formulas to compute the maximum likelihood estimates of the mixture parameters can be recognized.

Generalization to a mixture of multivariate normal densities is relatively straightforward. For the general case where the covariance matrices are not diagonal, the joint prior density is the product of a Dirichlet density and multivariate normal-Wishart densities. In the case of diagonal covariance matrices, the problem for each component reduces to the 1-dimensional case, and formulas 6 and 7 are applied to each vector component.

When the above regularity conditions on the prior joint density are not satisfied we have no proof of convergence of this algorithm. However, in practice we have not encountered any problems when these conditions were only approximately satisfied.

Segmental MAP algorithm

The above procedure to evaluate the MAP of a mixture of Gaussians can be applied to estimate the observation density parameters of an HMM state given a set of observations $\mathcal{X}$ assumed to be independently drawn from the state distribution. Following the scheme of the segmental k-means algorithm [6], we obtain a segmental MAP algorithm [9, 10]. First, the HMM parameters are initialized with values corresponding to the mode of the prior density. Second, the Viterbi algorithm is used to segment the training data $\mathcal{X}$ into sets of observations associated with each HMM state, and third, the MAP estimate procedure is applied to each state. The second and third steps are iterated until convergence.

In order to compare our results to results previously obtained with the k-means segmental algorithm [11] we used the segmental MAP algorithm to evaluate the HMM parameters. However, if it is desired to maximize $P(\mathcal{X}|\theta)P(\theta)$ over the HMM and not only state by state along the best state sequence, a Bayesian version of the Baum-Welch algorithm can easily be designed [10].

Prior density estimation

The method of estimating the prior parameters depends on the desired goals. We envisage the following three types of applications for Bayesian learning.

Sequential training: The goal is to update existing models with new observations without reusing the original data in order

to save time and memory. After each new data set has been processed, the prior densities must be replaced by an estimate of the posterior densities. In order to approach the HMM MLE estimators the size of each observation must be as large as possible. The process is initialized with non-informative prior densities.

Model adaptation: For model adaptation most of the prior density parameters are derived from parameters of an existing HMM. (This justifies the term "model adaptation" even if the only sources of information for Bayesian learning are the prior densities and the new data.) To estimate parameters not directly obtained from the existing model, training data is needed in which the "missing" prior information can be found. This can be the data already used to build the existing models or a larger set containing the variability we want to model with the prior densities.

Parameter smoothing: Since the goal of parameter smoothing is to obtain robust HMM parameters, shared prior parameters must be used. These parameters are estimated on the same training data used to estimate the HMM parameters via Bayesian learning. For example, with this approach context-dependent (CD) models can be built from context-independent (CI) ones.

In this study we were mainly interested in the problems of speaker-independent training and speaker adaptation. Therefore parameter smoothing and model adaptation in which the prior density parameters must be evaluated from SI or SD models and from SI training data were investigated. The prior density parameters were estimated along with the estimation of the SI model parameters using the segmental k-means algorithm. Information about the variability to be modeled by the prior densities was associated with each frame of the SI training data. This information was represented by a class number corresponding to the speaker number, sex, or phonetic context. The prior density parameters were estimated from the class mean vectors and the SI HMM parameters [10].

EXPERIMENTS

The 3 first experiments used a set of 1769 CD phone models. Each model is a 3 state left-to-right HMM with Gaussian mixture state observation densities (except for silence which is a one-state model). Diagonal covariance matrices are used and the transition probabilities are assumed to be fixed and known. A 38-dimensional feature vector [12] composed of 12 cepstrum coefficients, 12 delta cepstrum coefficients, the delta log energy, 12 delta-delta cepstrum coefficients, and the delta-delta log energy is used. The training and testing materials were taken from the DARPA Naval Resource Management task as provided by NIST. For telephone bandwidth compatibility, the original speech signal was filtered from 100 Hz to 3.8 kHz and down-sampled at 8 kHz. Results are reported using the standard word-pair grammar with a perplexity of about 60. The SI training data consisted of 3969 sentences from 109 speakers (78 males and 31 females), subsequently referred to as the SI-109 training data.

CD model smoothing

It is well known that HMM training requires smoothing, particularly if a large number of CD phone models are used with limited training data. While several solutions have been investigated to smooth discrete HMMs, such as model interpolation, co-occurence smoothing, and fuzzy VQ, only variance smoothing has been proposed for continuous density HMMs. We investigated the use of Bayesian learning to train CD phone models with prior densities obtained from CI phone training. This approach can be seen as model interpolation between CI and CD models for the case of continuous density HMMs.

Models were built with MLE and MAP approaches using the SI-109 training data. For the MAP estimation, the prior densities were based on a 47 CI model set. Covariance clipping, as reported in [11], has been used for the two approaches. Experiments were carried out using mixtures of 16 Gaussian components on the FEB89, OCT89, JUN90 and FEB91 DARPA tests including 1380 sentences (11843 words). An average word error reduction of 10% (from 6.0 to 5.5) was obtained using parameter smoothing. This improvement is small since the 1769 phone model set had originally been designed to be trained with a MLE approach on the SI-109 training data [11] but it validates the approach.

Speaker adaptation

In the framework of Bayesian learning, speaker adaptation may be viewed as adjusting speaker-independent models to form speaker-specific ones, using the available prior information and a small amount of speaker-specific adaptation data. Along with the estimation of the parameters for the SI CD models, the prior densities are simultaneously estimated during the speaker-independent training process. The speaker-specific models are built from the adaptation data using the *segmental MAP* algorithm. The SI models are used to initialize the iterative adaptation process. After segmenting all of the training sentences with the models generated in the previous iteration, the speaker-specific training data is used to adapt the CD phone models both with and without reference to the segmental labels. Three types of adaptation were investigated: adapting all CD phones with the exact triphone label (type 1), those with the same CI phone label (type 2), and all models without regard to the label (type 3). Each frame of the sentence is distributed over the models based on the observation densities of the preceding iteration. When the model labels are not used, this method can be viewed as probabilistic spectral mapping constrained by the prior densities. It was found that a combination of adaptation types 1 and 2 was the most effective for fast speaker adaptation. While a maximum of 8 mixture components per density was allowed, the actual average number of components was 7. This represents a total of 3 million parameters to be estimated and adapted.

Experiments were conducted using approximately 1 and 2 minutes of adaptation data to build the speaker-specific models. In 40 utterances, roughly 2 minutes of speech, only about 45% of the CD phones appear (28% for 20 sentences), whereas typically all the CI phones appear. Table 1 summarizes the test results[2] on the JUN90 data for the last 80 utterances of each speaker, where the first 20 (or 40) utterances were used for supervised adaptation of types 1 and 2. Speaker-independent recognition results are shown for comparison. With 1 minute and 2 minutes of speaker-specific training data, a 16% and 31% reduction in word error were obtained compared to the SI results. On this test speaker adaptation appears to be effective only for the female speakers for whom SI results were lower than for the male speakers.

Experiments have also been carried out using unsupervised speaker adaptation, which is more applicable to on-line situations. Starting with the SI models, adaptation of SI phone models is performed every 40 utterances using type 2 adaptation. The results on the JUN90 test are shown in Table 2 for the last 80 sentences of each speaker. There is an overall error reduction of 16%.

[2]Results reported in this subsection were obtained with a recognizer using a guided search strategy [14] which has been found to give slightly biased and better performance than a regular beam search strategy.

Speaker	SI	SA (1 min)	SA (2 min)	Err. Red. (2 min)
BJW(F)	4.7	3.4	2.2	53%
JLS(M)	3.6	3.0	3.4	5%
JRM(F)	9.2	7.0	5.3	42%
LPN(M)	3.2	4.7	3.2	0%
Overall	5.1	4.3	3.5	31%

Table 1: Speaker adaptation results on the JUN90 test data.

Speaker	SI	SA (2 x 2 min)
BJW(F)	4.7	3.4
JLS(M)	3.6	3.5
JRM(F)	9.2	6.6
LPN(M)	3.2	3.7
Overall	5.1	4.3

Table 2: Unsupervised speaker adaptation results on the JUN90 test.

Sex-dependent modeling

It has recently been reported that the use of different models for male and female speakers reduced recognizer errors on the RM task using a word-pair grammar with models trained on the SI-109 data set (e.g. [13]). We investigated the same idea within the framework of Bayesian learning. Two sets of 1769 CD phone models were generated using data from the male speakers for one set and from the female speakers for the other set. For both sets the same prior density parameters, which had been estimated along with SI training on all 109 speakers, were used. Recognition was performed by computing the likelihoods of the sentence for the two sets of models and by selecting the solution with to the highest likelihood. In order to avoid problems due to likelihood disparities caused by implementation details, all HMM parameters other than the Gaussian mean vectors were assumed to be known and set to the parameter values of the SI models.

Recognition of the FEB91 test data (5m/5f speakers) gives a 4.6% word error rate with both sets of models as compared to 5.4% with the SI model set. This result confirms the interest of the speaker clustering and validates Bayesian learning as a way to generate sex-dependent models.

Corrective training

Bayesian learning provides a scheme for model adaptation which can also be used for corrective training. Corrective training maximizes the recognition rate on the training data hoping that will also improve performance on the test data. One simple way is to use the training sentences which were incorrectly recognized as new data.

In order to do that, the second 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 reestimation formulas, the values c_{ik} for each state j are replaced by $\gamma_{ij}\omega_k\mathcal{N}(x_i|m_{jk},\tau_{jk})/P(x_i|\theta_j)$ where γ_{ij} is equal to 1 in the sentence model and to -1 in the general model. The convergence is not guaranteed but in practice by using large values for τ_k ($\simeq 200$) 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 algorithm for CDHMMs very similar to the corrective MMIE training proposed in [15].

Preliminary experiments have been carried out on the TI/NIST connected digits database using a set of 21 phonetic HMMs trained on the 8565 digit strings. Only the Gaussian mean vectors and the mixture weights were corrected. On the 8578 test strings, string error rates of 1.5% and 1.3% were obtained with 16 and 32 mixture components per state respectively, compared to 2.0% and 1.5% without corrective training.

SUMMARY

An investigation into the use of Bayesian learning of CDHMM parameters has been carried out. The theorical framework for training HMMs with Gaussian mixture densities was presented. It was shown that Bayesian learning can serve as a unified approach for parameter smoothing, speaker adaptation, speaker clustering and corrective training. Encouraging results have been obtained for these applications. On the DARPA RM task we observed an 10% word error reduction with HMM parameter smoothing, 31% for speaker adaptation with 2 minutes of speaker specific training data, and 15% with sex-dependent modeling. On the TI connected digit recognition task, 15% to 25% string error reduction was achieved with corrective training.

REFERENCES

[1] M. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

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

[3] F. Jelinek, R. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data", *Pattern Recognition in Practice*, pp. 381-397, North-Holland Publishing Company, 1980.

[4] R. Zelinski, F. Class, "A learning procedure for speaker-dependent word recognition systems based on sequential processing of input tokens," *Proc. ICASSP83*, pp. 1053-1056, Boston, May 1983.

[5] P. Brown, C.-H. Lee, J. Spohrer, "Bayesian Adaptation in Speech Recognition," *Proc. ICASSP83*, pp. 761-764, Boston, May 1983.

[6] L. Rabiner, J. Wilpon, B.-H. Juang, "A Segmental *k*-Means Training Procedure for Connected Word Recognition", *AT&T Tech. Journal.*, Vol. 65, No. 3, pp. 21-32, May-June 1986.

[7] R. Stern, M. Lasry, "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," *IEEE Trans. on ASSP*, Vol. ASSP-35, No. 6, June 1987.

[8] M. Ferretti, S. Scarci, "Large-Vocabulary Speech Recognition with Speaker-Adapted Codebook and HMM Parameters", *Proc. Eurospeech89*, pp. 154-156, Paris, Sept. 1989.

[9] C.-H. Lee, C.-H. Lin, B.-H. Juang, "A Study on Speaker Adaptation of Continuous Density HMM Parameters", *Proc. ICASSP90*, pp. 145-148, Albuquerque, April 1990.

[10] J.-L. Gauvain, C.-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models", *Proc. DARPA Speech and Natural language Workshop*, Pacific Grove, February 1991.

[11] 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.

[12] C.-H. Lee, E. Giachin, L. Rabiner, R. Pieraccini, A. Rosenberg, "Improved Acoustic Modeling for Continuous Speech Recognition", *Proc. DARPA Speech and Natural language Workshop*, Hidden Valley, June 1990.

[13] X. Huang, F. Alleva, S. Hayamizu, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, "Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition," *Proc. DARPA Speech and Natural language Workshop*, Hidden Valley, June 1990.

[14] R. Pieraccini, C.-H. Lee, E. Giachin, L. Rabiner, "Implementation Aspects of Large Vocabulary Recognition Based on Intraword and Interword Phonetic Units," *Proc. DARPA Speech and Natural language Workshop*, Hidden Valley, June 1990.

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