# 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, USA*

Received 26 September 1991
Revised 23 January 1992

**Abstract.** An investigation into the use of Bayesian learning of the parameters of a multivariate Gaussian mixture density has been carried out. In a framework of continuous density hidden Markov model (CDHMM), 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 parameter smoothing, speaker adaptation, speaker clustering and corrective training are given.

**Zusammenfassung.** Wir berichten über eine Untersuchung zum Einsatz der Bayes'schen Lerntheorie zur Schätzung der Parameter von multi-variaten Gauss'schen Verteilungsdichten. Im Rahmen eines "Hidden Markov Modells" mit kontinuierlicher Dichteverteilungen (CDHMM) stellt die Bayes'sche Theorie einen einheitlichen Ansatz dar zur Parameterglättung, Sprecheradaption, Sprecherklusterung und zum korrigierenden Training. Das Ziel ist, die Modellrobustheit eines auf CDHMM basierenden Spracherkennungssystems in Hinblick auf die Ergebnisse zu verbessern. Unser Ansatz ist, Bayes'sches Lernen zu benutzen, um Vorwissen in Form von initialen Dichten der HMM-Parameter in den Trainingsprozess einzubringen. Wir stellen die theoretische Basis für dieses Verfahren dar und wenden es zur Glättung von Parametern, Sprecheradaption, Sprecherklusterung und im korrigierenden Training an.

**Résumé.** Une étude sur l'utilisation de l'apprentissage bayésien des paramètres de densités multigaussiennes a été effectuée. Dans le cadre des modèles markoviens cachés à densitiés d'observations continues (CDHMM), l'apprentissage bayésien est un outil très général applicable au lissage des paramètres, à l'adaptation au locuteur, à l'estimation de modèles par groupe de locuteurs et à l'apprentissage correctif. Le but est d'augmenter la robustesse des modèles d'un système de reconnaissance afin d'en améliorer les performances. Notre approche consiste a utiliser l'apprentissage bayésien pour incorporer une connaissance a priori dans le processus d'apprentissage sous forme de densités de probabilités des paramètres des modèles markoviens. La base théorique de cette procédure est présenteée, ansi que les résultats obtenus pour le lissage des paramètres, l'adaptation au locuteur, l'estimation de modèles propres à chaque sexe et l'apprentissage correctif.

**Keywords.** Bayesian learning; Hidden Markov models; parameter smoothing; speaker adaptation; speaker clustering; corrective training.

## 1. 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 cannot 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.

[1] 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.

Techniques used to alleviate the problem of insufficient training data 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 (Jelinek and Mercer, 1980) 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 (Gauvain and Lee, 1991). The prior knowledge consists of prior densities of the HMM parameters. Such an approach was shown to be effective for speaker adaptation in isolated word recognition where the parameters of multivariate Gaussian state observation densities of whole-word HMMs were adapted (Lee et al., 1990a). 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 is shown to serve as a unified approach for parameter smoothing, speaker adaptation, speaker clustering and corrective training.

## 2. 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 $\theta$ is the parameter vector to be estimated from a sequence of $n$ observations $x_1, \ldots, x_n$, given a prior density $\mathrm{P}(\theta)$, then one estimate for $\theta$ is the maximum a posteriori (MAP) estimate which corresponds to the mode of the posterior density,

$$\theta_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\ \mathrm{P}(x_1, \ldots, x_n \mid \theta)\mathrm{P}(\theta). \tag{1}$$

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

Given the MAP formulation in (1) two problems remain: the choice of the prior distribution family and the 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 requires the existence of a sufficient statistic of a fixed dimension (DeGroot, 1970). If the observation density possesses such a statistic $s$ and if $g(\theta \mid s, n)$ is the associated kernel density, MAP estimation is reduced to the evaluation of the mode of the product $g(\theta \mid s, n)\mathrm{P}(\theta)$. In addition, if the prior density is chosen from the same family as the kernel density, $\mathrm{P}(\theta) = g(\theta \mid t, m)$, the previous product is simply equal to $g(\theta \mid u, m+n)$ 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, $g(\theta \mid u, m+n)$ 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 components in the joint prior density. To simplify solving (1), we restrict our choice to a product of a Dirichlet density and kernel densities of the mixture exponential density, $\mathrm{P}(\theta) \propto \prod_{k=1}^{K} \omega_k^{m_k} g(\theta_k \mid t_k, m_k)$, where $K$ is the number of mixture components and the $\omega_k$'s are the mixture weights.

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.

### 2.1. Normal density case

Bayesian learning of a normal density is well known (DeGroot, 1970). If $x_1, \ldots, x_n$ is a random sample from $\mathrm{N}(x|m, r)$, where $m$ and $r$ are the mean and the precision (reciprocal of the variance), respectively, and if $\mathrm{P}(m, r)$ is a normal-gamma prior density, $\mathrm{P}(m, r) \propto r^{1/2} \exp(-\frac{1}{2}\tau r(m-\mu)^2) r^{\alpha-1} \exp(-\beta r)$, 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 $\hat{\mu}$ and $(\hat{\alpha}-0.5)/\hat{\beta}$, respectively.

This approach has been widely used for sequential learning of the mean vectors of feature- and template-based recognizers, see for example (Zelinski and Class, 1983; Stern and Lasry, 1987). Ferretti and Scarci (1989) 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. (1983) used Bayesian estimation for speaker adaptation of CDHMM parameters in a connected digit recognizer. More recently Lee et al. (1990a) 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.

### 2.2. Mixture of normal densities

For this study we use CDHMM where the state observation densities are mixtures of multivariate normal densities (Lee et al., 1990b, 1990c). However, to simplify the presentation of our approach, we assume here a mixture of univariate normal densities,

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

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

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

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 $\mathrm{P}(\theta|x_1, \ldots, x_n)$ 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 EM algorithm (Dempster et al., 1977) to find the maximum a posteriori. The following recursive formulas estimate the MAP of the 3 parameter sets:

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

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

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

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

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 EM 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 (6) and (7) are applied to each vector component.

It should be noted that the convergence of this algorithm can be proved even when the above regularity conditions are not satisfied (Gauvain and Lee, 1992).

### 2.3. Segmental MAP algorithm

The above procedure to evaluate the MAP of a mixture of Gaussians can be used to estimate the observation density parameters of an HMM state given a set of observations $\mathscr{X}$ assumed to be independently drawn from the state distribution. Following the scheme of the segmental $k$-means algorithm (Rabiner et al., 1986), we obtain a segmental MAP algorithm (Lee et al., 1990a; Gauvain and Lee, 1991). 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 $\mathscr{X}$ 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 (Lee et al., 1990b), we used the segmental MAP algorithm to evaluate the HMM parameters. However, if it is desired to maximize $P(\mathscr{X} | \theta)P(\theta)$ over the HMM and not only state by state along the best state sequence, a Bayesian version of the Baum-Welch algorithm can also be designed (Gauvain and Lee, 1992).

As in the case of MLE, one simply replaces $c_{ik}$ by $c_{ijk}$ in the re-estimation formulas and applies the summations over all the observations for each state $s_j$:

$$c_{ijk} \triangleq \gamma_{ij} \frac{\omega_k N(x_i | m_{jk}, r_{jk})}{P(x_i | \theta_j)}, \tag{8}$$

where $\gamma_{ij}$ is the probability of being in the state $s_j$ at time $i$, given that the model generates $\mathscr{X}$. For the segmental MAP approach $\gamma_{ij}$ is equal to 0 or 1.

### 2.4. 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 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 obtained 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 label 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 (Gauvain and Lee, 1991).

## 3. Experiments

In the following subsections we discuss experiments on parameter smoothing, speaker adaptation, speaker clustering and corrective training. In these experiments three sets of phone models were used: 1769 CD phone models and 47 and 21 CI 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 (Lee et al., 1990c) 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 and from the TI/NIST connected digits corpus as provided by NIST. For telephone bandwidth compatibility, the original RM speech signal was filtered from 100 Hz to 3.8 kHz and down-sampled at 8 kHz, and the TI/NIST digits were low-pass filtered at 3.2 kHz and down-sampled at 6.67 kHz. For RM, 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.

### 3.1. 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-occurrence 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 either as a way to add extra constraints to the model parameters so as to reduce the effect of insufficient training data, or 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. Here the reduced set is obtained by removing the context dependency, but other prior parameter tyings can be investigated.

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 (Lee et al., 1990b), was used for the two approaches. Experiments were carried out using mixtures of 16 Gaussian components on the FEB89, OCT89, JUN90 and FEB91 DARPA tests containing 1380 sentences (11 843 words). An average word error reduction of 10% (from 6.0 to 5.5) was obtained using parameter smoothing. Although this improvement is small (we suspect because the 1769 phone model set had originally been designed (Lee et al., 1990b) to be trained with an MLE approach on the SI-109 training data), it validates the approach.

### 3.2. 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. The prior densities are simultaneously estimated during the speaker-independent training process along with the estimation of the parameters for the SI CD models. The speaker-specific models are built from the adaptation data using the segmental MAP algorithm.

The iterative adaptation process is initialized with the SI models. After segmenting 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 (2 minutes of speech) 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 45% of the CD phones appear (28% for 20 sentences), whereas typically all the CI phones appear. Table 1 summarizes the test results 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, reductions in word error of 16% and 31% were obtained as compared to the SI results. On this test speaker adaptation appears to be effective only for the female speakers for whom the SI results were lower than for the male speakers.

Table 1
Speaker adaptation results given as word error rate (%) and word error reduction on the JUN90 test data

| Speaker | SI | SA<br>(1 min) | SA<br>(2 min) | Err. red.<br>(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 2
Unsupervised speaker adaptation results given as word error rate (%) and word error reduction on the JUN90 test data

| Speaker | SI | SA (2 × 2 min) | Err. red |
|---|---|---|---|
| BJW(F) | 4.7 | 3.4 | 28% |
| JLS(M) | 3.6 | 3.5 | 3% |
| JRM(F) | 9.2 | 6.6 | 28% |
| LPN(M) | 3.2 | 3.7 | −16% |
| Overall | 5.1 | 4.3 | 16% |

Experiments have also been carried out using unsupervised speaker adaptation, which is more applicable to on-line situations. Adaptation of the SI phone models is performed every 40 utterances using type 2 adaptation with the recognized labels. The adapted models are then used to recognize the next 40 utterances. 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%.

In order to compare speaker adaptation to ML training of SD models, an experiment has been carried out on the FEB91-SD test material which includes data from 12 speakers (7 male and 5 female), using a set of 47 CI phone models. Two, five and thirty minutes of the SD training data were used for training and adaptation. The SD, SA (SI) word error rates are given in the two first rows of Table 3. The SD word error rate for 2 minutes of training data was 31.5%. The SI word error rate (0 minutes of adaptation data) for the SI-109 model was 13.9% which is comparable to the SD results with 5 minutes of SD training data. The SA models

Table 3
Summary of SD, SA (SI) and SA (M/F) results on FEB91-SD test. Results are given as a word error rate (%)

| Training | 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 |

were shown to perform better than SD models when relatively small amounts of data were used for training or adaptation. When all the available training data (about 30 minutes) was used, the SA and SD results were comparable, consistent with the Bayesian formulation that the MAP estimate asymptotically converges to the MLE. Relative to the SI results, the word error reduction was 37% with 2 minutes of adaptation data, an improvement comparable to that observed on the JUN90 test data with CD models. As in the previous experiment, a larger improvement was observed for the female speakers (51%) than for the male speakers (22%).

### 3.3. *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 (Huang et al., 1990). 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 during 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 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-SI test data (5m/5f speakers) gave a 4.6% word error rate with both sets of models, compared to 5.4% with the SI model set. This error rate is 33% less than that obtained using sex-dependent models trained with MLE. These results reinforce the interest of the speaker clustering and validate Bayesian learning as a way to generate sex-dependent models.

Given that we have demonstrated the effectiveness of sex-dependent SI models, we pose the question of whether or not we can obtain additional improvement with speaker adaptation starting from the combined M/F models. To address this question we trained male and female models for the 47 CI units and evaluated speaker adaptation with the FEB91-SD test data. The results are given in the third row of Table 3 for 2, 5 and 30 minutes of adaptation data. The word error rate for the sex-dependent models with no speaker adaptation is 11.5%. The error rates are reduced to 7.5% with 2 minutes and 6.0% with 5 minutes, of adaptation data. Comparing the last 2 rows of the table it can be seen that speaker adaptation is more effective when sex-dependent seed models are used. The error reduction with 2 minutes of training data is 35% compared to the sex-dependent model results and 46% compared to the SI model results.

### 3.4. *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 that will also improve performance on the test data. One simple way to do corrective training is to use the training sentences which were incorrectly recognized as new data.

In order to do so, 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 re-estimation formulas, the values $c_{ijk}$ for each state $s_j$ are replaced by $\gamma_{ij}\omega_{jk}\mathrm{N}(x_i|m_{jk}, r_{jk})/\mathrm{P}(x_i|\theta_j)$, where $\gamma_{ij}$ 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 $\tau_{jk}$ ($\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 training algorithm for CDHMM very similar to the corrective MMIE training proposed by Normandin and Morgera (1991).

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 are 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

Table 4
Corrective training results in string and word error rates (%) on the TI-digits for 21 CI models with 16 and 32 mixture components per state; string error counts are given in parentheses

| Training conditions | Training |  | Test |  |
|---|---|---|---|---|
|  | String | Word | String | Word |
| MLE-16 | 1.6 (134) | 0.5 | 2.0 (168) | 0.7 |
| CT-16 | 0.2 (18) | 0.1 | 1.4 (122) | 0.5 |
| MLE-32 | 0.8 (67) | 0.2 | 1.5 (126) | 0.5 |
| CT-32 | 0.3 (29) | 0.1 | 1.3 (111) | 0.4 |

are given in Table 4 using 16 and 32 mixture components for the observation pdfs. String and word error rates are given with and without corrective training for both test and training data. The CT-16 results were obtained with 8 iterations of corrective training while the CT-32 results were based on only 3 iterations. 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, compared to 2.0% and 1.5% without corrective training. These represent string error reductions of 27% and 12%. We note that corrective training helps more with smaller models, as the ratio of adaptation data to the number of parameters is larger.

The corrective training procedure is also effective for continuous sentence recognition of the RM task. Table 5 gives results for this task, using 47 SI-CI models with 32 mixture components. Corrective training gives an average word error rate reduction of 20% on the test data. For this experiment we used a small beam search width to recognize the training data so as to increase the amount of corrective training data and also to speed up the training process. It was observed that this procedure not only reduces the error rate in training but also increases the separation between the correct string and the other competing strings, resulting in better performance on the test data. (The error rate reduction is only 15% with a standard beam width.)

From these results we can conclude that this approach works in that the performance on the training data was dramatically improved and that increasing the performance on the training data gave improved recognition of the test data.

## 4. Summary

An investigation into the use of the Bayesian learning of CDHMM parameters has been carried out. The theoretical 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. Performance improvements have been observed for these four applications. On the DARPA RM task we observed a word error reduction of 10% with HMM parameter smoothing, 31% to 46% for speaker adaptation with 2 minutes of speaker specific training data, and 15% to 17% with sex-dependent modeling. It was also found that speaker adaptation based on sex-dependent models gave a better result than that obtained with a speaker-independent seed. Compared to speaker-dependent training, speaker adaptation achieved a better performance with the same amount of training/adaptation data. Corrective training was also found to be effective on the RM and TI/NIST connected digit tasks. We observed a word error reduction of 20% on the RM task and a string error reduction of 12–27% on the TI/NIST task. It was observed that corrective training helps more with models having a smaller number of parameters.

Table 5
Corrective training results on the RM task for 47 CI models with 32 mixture components per state; results are given as word error rate (%)

| Test | Training | FEB89 | OCT89 | JUN90 | FEB91 | FEB91-SD |
|---|---|---|---|---|---|---|
| MLE-32 | 7.7 | 11.9 | 11.5 | 10.2 | 11.4 | 13.9 |
| CT-32 | 3.1 | 8.9 | 8.9 | 8.1 | 10.2 | 11.0 |

## References

P. Brown, C.-H. Lee and J. Spohrer (1983), "Bayesian adaptation in speech recognition," *Proc. Internat. Conf. Acoust. Speech Signal Process. 83*, Boston, pp. 761-764.

M. DeGroot (1970), *Optimal Statistical Decisions* (McGraw-Hill, New York).

A. Dempster, N. Laird and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Statist. Soc. Ser. B*, Vol. 39, pp. 1-38.

M. Ferretti and S. Scarci (1989), "Large-vocabulary speech recognition with speaker-adapted codebook and HMM parameters", *Proc. Eurospeech 89*, Paris, pp. 154-156.

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, February 1991.

J.-L. Gauvain and C.-H. Lee (1992), "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", to be submitted.

X. Huang, F. Alleva, S. Hayamizu, H.-W. Hon, M.-Y. Hwang and K.-F. Lee (1990), "Improved hidden Markov modeling for speaker-independent continuous speech recognition," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990.

F. Jelinek and R. Mercer (1980), "Interpolated estimation of Markov source parameters from sparse data", *Pattern Recognition in Practice*, ed. by E.S. Gelsema and L.N. Kanal (North-Holland, Amsterdam), pp. 381-397.

C.-H. Lee, C.-H. Lin and B.-H. Juang (1990a), "A study on speaker adaptation of continuous density HMM parameters", *Proc. Internat. Conf. Acoust. Speech Signal Process. 90*, Albuquerque, NM, pp. 145-148.

C.-H. Lee, L.R. Rabiner, R. Pieraccini and J. G. Wilpon (1990b), "Acoustic modeling for large vocabulary speech recognition", *Comput. Speech Language*, Vol. 4, pp. 127-165.

C.-H. Lee, E. Giachin, L. Rabiner, R. Pieraccini and A. Rosenberg (1990c), "Improved acoustic modeling for continuous speech recognition", *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, June 1990.

Y. Normandin and D. Morgera (1991), "An improved MMIE training algorithm for speaker-independent small vocabulary, continuous speech recognition", *Proc. Internat. Conf. Acoust. Speech Signal Process. 91*, pp. 537-540.

L. Rabiner, J. Wilpon and B.-H. Jang (1986), "A segmental *k*-means training procedure for connected word recognition", *AT&T Tech. J.*, Vol. 65, No. 3, May-June 1986, pp. 21-32.

R. Stern and M. Lasry (1987), "Dynamic speaker adaptation for feature-based isolated word recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, No. 6, June 1987.

R. Zelinski and F. Class (1983), "A learning procedure for speaker-dependent word recognition systems based on sequential processing of input tokens," *Proc. Internat. Conf. Acoust. Speech Signal Process. 83*, Boston, pp. 1053-1056.