Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models

Conference Paper · January 1991

DOI: 10.3115/112405.112457 · Source: DBLP

Chin-Hui Lee
Georgia Institute of Technology

Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models

Jean-Luc Gauvain[†] *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 of this study 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 CDHMM training process in the form of prior densities of the HMM parameters. The theoretical basis for this procedure is presented and preliminary results applying to HMM parameter smoothing, speaker adaptation, and speaker clustering are given.

Performance improvements were observed on tests using the DARPA RM task. For speaker adaptation, under a supervised learning mode with 2 minutes of speaker-specific training data, a 31% reduction in word error rate was obtained compared to speaker-independent results. Using Baysesian learning for HMM parameter smoothing and sex-dependent modeling, a 21% error reduction was observed on the FEB91 test.

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 conventional *maximum likelihood* 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.

Techniques used to alleviate the insufficient training data problem include probability density fuction (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 [4] [9] 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. 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 of a 39-word, English alpha-digit vocabulary where adaptation involved only the parameters of a multivariate Gaussian state observation density of whole-word HMM's [12]. 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.

In the case of speaker adaptation, Bayesian learning may be viewed as a process for adjusting speaker-independent (SI) models to form speaker-specific ones based on the available prior information and a small amount of speaker-specific adaptation data. The prior densities are simultaneously estimated during the SI training process along with the estimation of the SI model parameters. The joint prior density for the parameters in a state is assumed to be a product of normal-gamma densities for the mean and variance parameters of the mixture Gaussian components and a Dirichlet density for the mixture gain parameters. The SI models are used to initialize the iterative adaptation process. The speaker-specific models are derived from the adaptation data using a *segmental MAP algorithm* which uses the Viterbi algorithm to segment the data and an EM algorithm to estimate the mode of the posterior density.

In the next section the principle of Bayesian learning for CDHMM is presented. The remaining sections report preliminary results obtained for model smoothing, speaker adaptation and sex-dependent modeling.

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 way to estimate θ is to use the maximum a posteriori (MAP) estimate which corresponds to the mode of the posterior density $P(\theta|x_1, ..., x_n)$, i.e.

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

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

On the other hand, if θ is assumed to be 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] [2]. 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 conjugate family, i.e. in same family of 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. The MAP estimate is then

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, g(\theta|u, m+n) \tag{2}$$

In this case, the MAP estimation problem is closely related to the MLE problem which consists of 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, the MAP estimation, like ML estimation, has no analytical solution, but 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 the problem of finding the solution to equation 1, we restrict our choice to a product of a Dirichlet density and kernel densities of the mixture exponential density, i.e.

$$P(\theta) \propto \prod_{k=1}^{K} \omega_k^{m_k} g(\theta_k|t_k, m_k) \tag{3}$$

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 with parameters $\hat{\mu}$, $\hat{\beta}$, $\hat{\alpha}$ and $\hat{\tau}$ such that:

$$\hat{\mu} = \frac{\tau}{\tau+n}\mu + \frac{n}{\tau+n}\bar{x} \tag{4}$$

$$\hat{\beta} = \beta + \frac{n}{2}S_x + \frac{\tau n(\bar{x}-\mu)^2}{2(\tau+n)} \tag{5}$$

$$\hat{\alpha} = \alpha + n/2 \tag{6}$$

$$\hat{\tau} = \tau + n \tag{7}$$

where S_x is the variance of the random sample. The MAP estimates of μ 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-based or template-based speech recognizers, see for example [5] and [8]. Ferretti and Scarci [11] 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 was limited to a Gaussian.

Brown *et al.* [6] have used Bayesian estimation for speaker adaptation of CDHMM parameters in a connected digit recognizer. More recently Lee *et al.* [12] investigated various training schemes of the 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 ML estimates.

Mixture of normal densities

In the current implementation of the recognizer used in this study [13] [14] the state observation density is a mixture of multivariate normal densities. However, to simplify the presentation of our approach, we assume here a mixture of univariate normal densities:

$$P(x|\theta) = \sum_{k=1}^{K} \omega_k \mathcal{N}(x|m_k, r_k) \tag{8}$$

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

$$P(\theta) \propto \prod_{k=1}^{K} \omega_k^{\lambda_k} r_k^{1/2} \exp(-\frac{\tau_k r_k}{2}(m_k-\mu_k)^2) r_k^{\alpha_k-1}\exp(-\beta_k r_k) \tag{9}$$

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 is now to find the mode of the posterior joint 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 fact suggests the use of the E.M. algorithm [3] to find the maximum a posteriori. The following recursive formulas estimate the MAP of the 3 parameter sets.

$$c_{ik} \triangleq \frac{\omega_k \mathcal{N}(x_i|m_k, r_k)}{P(x_i|\theta)} \tag{10}$$

$$\omega'_k = \frac{\lambda_k + \sum_{i=1}^{n} c_{ik}}{n + \sum_{k=1}^{K} \lambda_k} \tag{11}$$

$$m'_k = \frac{\tau_k \mu_k + \sum_{i=1}^{n} c_{ik} x_i}{\tau_k + \sum_{i=1}^{n} c_{ik}} \tag{12}$$

$$r'_k = \frac{2\alpha_k - 1 + \sum_{i=1}^{n} c_{ik}}{2\beta_k + \sum_{i=1}^{n} c_{ik}(x_i - m'_k)^2 + \tau_k(\mu_k - m'_k)^2} \tag{13}$$

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 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 prior joint 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 (12) and (13) 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 n observations $x_1, ..., x_n$ assumed to be independently drawn from the state distribution. Following the scheme of the segmental k-means algorithm [7] to estimate the parameters of an HMM, first the Viterbi algorithm is used to segment the training data $\mathcal{X}$ into sets of observations associated with each HMM state and then the MAP estimate procedure is applied to each state. The following segmental MAP algorithm originally proposed in [12] is obtained:

1. Set $\hat{\theta} = \text{argmax}_\theta P(\theta)$
2. Obtain the optimal state sequence $\hat{S}$, i.e.
$$\hat{S} = \underset{S}{\text{argmax}}\, P(\mathcal{X}|S, \hat{\theta})P(\hat{\theta})$$
3. Given the state sequence $\hat{S}$, use the E.M. algorithm to find $\hat{\theta}$ such that
$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, P(\mathcal{X}|\hat{S}, \theta)P(\theta)$$
4. Iterate 2 and 3 , until convergence.

In order to compare our results to results previously obtained with the k-means segmental algorithm [13] 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 also be designed. As in the case of maximum likelihood estimation, simply replace c_{ik} by c_{ijk} in the reestimation formulas and apply the summations over all the observations for each state j:

$$c_{ijk} \triangleq \gamma_{ij} \frac{\omega_k \mathcal{N}(x_i|m_{jk}, r_{jk})}{P(x_i|\theta_j)} \tag{14}$$

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

Prior density estimation

If the prior density defined by equation (9) for a mixture of univariate Gaussians is used, more parameters need to be evaluated for the prior density than for the mixture density itself. As in the case for the HMM parameters, it is therefore of interest to use tied parameters for the prior densities in order to obtain more robust estimators or to simply reduce the memory requirements.

The method of estimating these 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 use of 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 data 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) models.

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. This approach was used to smooth the parameters of CD models, for speaker adaptation, and to build sex-dependent models.

In these three cases, 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 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 number, the speaker sex, or the phonetic context. The HMM parameters for each class $\mathcal{C}_l$ given the mixture component were then computed. For the experiments reported in this paper, the prior density parameters were estimated as follows:

$$\alpha_{jk} = \frac{\tau_{jk} + 1}{2} \tag{15}$$

$$\beta_{jk} = \frac{\tau_{jk}}{2 r_{jk}} \tag{16}$$

$$\mu_{jk} = m_{jk} \tag{17}$$

$$\lambda_{jk} = \omega_{jk} \sum_{k=1}^{K} \tau_{jk} \tag{18}$$

$$\tau_{jk} = \frac{p \sum_l c_{jkl}}{\sum_l c_{jkl} (y_{jkl} - m_{jk})^t (\sum_k \omega_k r_{jk}^{-1})^{-1} (y_{jkl} - m_{jk})} \tag{19}$$

where ω_{jk}, m_{jk}, and r_{jk} are the SI HMM parameters for each state j and each mixture component k (m_{jk} and r_{jk} are vectors of p components). 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 $x_i \in \mathcal{C}_l$ and $c_{ijkl} = 0$ if $x_i \notin \mathcal{C}_l$, and $c_{jkl} = \sum_i c_{ijkl}$. It can seen that when the τ_{jk}'s are known all the other prior parameters are directly estimated from the SI HMM parameters. The prior density parameters τ_{jk} can be regarded as a weight associated with the k^{th} Gaussian of state s_j. When this weight is large, the prior density is sharply peaked around the values of the SI HMM parameters and these values will be modified only slightly by the adaptation process. Conversely, if τ_{jk} is small the adaptation will be very fast. By choosing these estimators for the prior parameters the ability of the prior density to accurately model the inter-class variability is reduced but more robust estimators are obtained. Additionally, to further increase the robustness, the τ_{jk} values can be constrained to be identical for all Gaussians of a given state, or for all states of an HMM, or even for all the HMMs. For the experiments reported in this paper a common value for all the HMMs was estimated. This is clearly too strong a constraint and we plan to relax it in future experiments.

The state log-energy density parameters can be adapted using the same Bayesian learning principle. In the current models, a discrete pdf is used to model the state log-energy. Like for the mixture parameters, these pdfs were estimated using Bayesian learning. The prior density, a Dirichlet distribution, was estimated in the same way as the mixture weights. Bayesian learning of the log-energy pdf was not used for fast speaker adaptation since we could only adapt the parameters corresponding to a few observed log-energy values. In fact, here the more general problem is Bayesian learning of discrete HMMs based on multinomial distributions, for which only the statistics of the observed symbols can be adapted. One solution to this problem is to view, only for training purposes, the multinomial distribution as a mixture of Gaussians with a common covariance matrix.

CD MODEL SMOOTHING

It is well known that HMM training requires smoothing, particularly if a large number of context dependent (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.

All the experiments presented in this paper use 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 the silence model which has only one state. Diagonal covariance matrices are used and it is assumed that the transition probabilities are fixed and known. As described in [14], a 38-dimensional feature vector 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.

For the parameter smoothing experiments, the training data consisted of 3969 sentences from 109 speakers (78 males and 31 females). This data set will be subsequently referred to as 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 [13], has been used for the two approaches. Results are reported with a mixture of 16 Gaussian components for each state. Table 1 shows word error rates obtained for the FEB89, OCT89, JUN90, and FEB91 test sets using models estimated with the MLE and MAP methods.

An average error rate reduction of about 10% was ob-

Model type	FEB89	OCT89	JUN90	FEB91
MLE	6.2	6.0	6.3	5.8
MAP47	5.3	6.0	5.3	5.4

Table 1: Parameter smoothing with Bayesian learning.

served using parameter smoothing with prior densities estimated on a set of 47 units. This improvement is limited since the 1769 phone model set was originally designed to be trainable with a MLE approach on the SI-109 training data [13]. We intend to run some other experiments with a larger number of CD units to futher explore this approach.

SPEAKER ADAPTATION

Previous works on speaker-adaptation within the framework of the DARPA RM task have been reported for fast-adaptation (using less than 2 min of speech). Model interpolation has been proposed to adapt SI models [9] and probabilistic spectral mapping has been proposed to adapt SD models [10] and multi-speaker models [15]. 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. For fast speaker adaptation, it was found that a combination of adaptation types 1 and 2 was the most effective. The same set of 1769 CD phone units, where the observation densities are mixtures of 38-element multivariate Gaussian distributions was used for evaluation. 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 2 summarizes the test results[1] 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 also 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 speaker-independent results. On this test speaker adaptation appears to be effective only for the female speakers for whom SI results were lower than the male speakers.

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 2: Speaker adaptation results on the JUN90 test data.

Speaker	SI	SA (2 × 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 3: Unsupervised speaker adaptation results on the JUN90 test data.

Preliminary 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 3 for the last 80 sentences of each speaker. There is an overall error reduction of 16%.

SEX-DEPENDENT MODELING

It has recently been reported that the use of different models for male and female speakers reduced recognizer errors by 6% on the FEB89 and OCT89 tests using a word-pair grammar with models trained on the SI-109 data set [16]. 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 is performed by computing the likelihoods of the sentence for the two sets of models and by selecting the solution corresponding to the highest likelihood. In order to avoid problems due to likelihood disparities caused by implementation details, all the HMM parameters with the exception of the Gaussian mean vectors were assumed to be known and set to the parameter values of the SI models trained on the 109 speakers.

Table 4 shows the results obtained on the FEB91 test using the speaker independent set (SI), the male set (MA),

[1] Results reported in this section were obtained with a recognizer using a guided search strategy [17] which has been found to give slightly biased and better performance than a regular beam search strategy.

Speaker	SI	MA	FE	MA+FE
ALK(F)	9.3	11.5	8.6	8.6
CAL(F)	3.8	5.1	3.8	3.8
CAU(F)	3.3	3.7	3.7	3.7
EAC(F)	7.2	8.9	6.4	7.2
JLS(M)	1.6	2.0	2.0	2.0
JWG(M)	7.9	6.6	12.9	6.6
MEB(M)	4.1	3.3	6.5	3.3
SAS(M)	1.9	2.2	3.7	2.2
STK(M)	5.0	3.3	5.0	3.3
TRB(F)	10.9	18.3	5.7	5.7
overall	5.4	6.4	5.8	4.6

Table 4: Results on FEB91 test using separate male/female models.

the female set (FE), and the male and female sets together (MA+FE). Looking at the results speaker by speaker it can be seen that sex models do the job for which they have been designed; The best result for each speaker is obtained with the models of his/her sex. For the FEB91 test, the male models gave the higher likelihood for 153 sentences and the female models for 147 sentences. The overall improvement obtained using separate models for male and female speakers is a reduction in error rate of about 16%. This improvement is observed for both male and female speakers.

On the FEB91 test, using Baysesian learning for HMM parameter smoothing and sex-dependent modeling, a 21% error reduction compared to the baseline system results is obtained (5.8% to 4.6%).

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, and speaker clustering. Encouraging results have been obtained for these three applications.

Bayesian learning applied to HMM parameter smoothing had an overall 10% reduction on the word errors compared to results obtained using conventional segmental k-means training. Using Bayesian learning for sex-dependent modeling, an additional 15% error reduction was obtained. For speaker adaptation, a 31% error reduction was obtained on the JUN90 test with 2 minutes of speaker-specific training data. Since the extent of these tests is relatively limited, other experiments should be carried out to obtain more statistically significant results in order to fully validate this approach.

REFERENCES

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

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

[3] 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.

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

[5] R. Zelinski and 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.

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

[7] L. R. Rabiner, J. G. Wilpon, and 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.

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

[9] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system", *Ph.D. Thesis*, Carnegie-Mellon University, 1988.

[10] M.-W. Feng, F. Kubala, R. Schwartz, and J. Makhoul, "Improved Speaker Adaptation Using Text Dependent Spectral Mappings", *Proc. ICASSP88*, pp. 131-134, New York, April 1988.

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

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

[13] 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.

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

[15] F. Kubala and R. Schwartz, "A New Paradigm for Speaker-Independent Training and Speaker Adaptation", *Proc. DARPA Speech and Natural language Workshop*, Hidden Valley, June 1990.

[16] 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.

[17] R. Pieraccini, C.-H. Lee, E. Giachin, L. R. 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.