acceptable performance.

In the case of SPC-based prior estimation, a similar condition exists which can also limit its use. Nevertheless, better system performance can be expected from both approaches.

Future work may concentrate on adaptation of other system parameters using SPC-based MAP estimation and application of this approcach to mixture-Gaussian triphones.

References

[1] J-L. Gauvain and C-H Lee. "Bayesian Learning for Hidden Markov Model with Gaussian Mixture Observation Densities". *Speech communication,* 11: 205 - 213, 1992.

[2] C-H. Lee and J-L Gauvain. "A Study on Speaker Adaptation for Continuous Speech Recognition" *In Proc. ARPA Cont. Speech Rec. Workshop*, p.p. 59-64, Stanford, September 1992.

[3] A. P. Dempster, N.M. Laird, and D.B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society,* 39, Ser. B:1-38, 1977.

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

[5] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 2nd Edition, 1985.

[6] C-H. Lee, C-H Lin, and B-H. Juang. "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models". *IEEE Trans. Sig. Proc.,* SP-39(4): 806 - 814, April 1991.

[7] H. Robbins. "The Empirical Bayes Approach to Statistical Decision Problems". *Annals Math. Stat.,* 35 (1): 1-20, March 1964.

[8] J.S. Maritz and T. Lwin. *Empirical Bayes Methods.* Chapman and Hall, 2nd Edition, 1989.

[9] Q. *Huo* and C. Chan. "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition", Technical report, Department of Computer Science, University of Hong Kong, September 1992.

[10] S.M. Ahadi. "Improved Prior Estimation for Bayesian Adaptation". In *Proc. ICEE-97,* Sharif University, Tehran.

[11] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. "Large Vocabulary Continuous Speech Recognition Using HTK". In *Proc. ICASSP-94,* Adelaide.

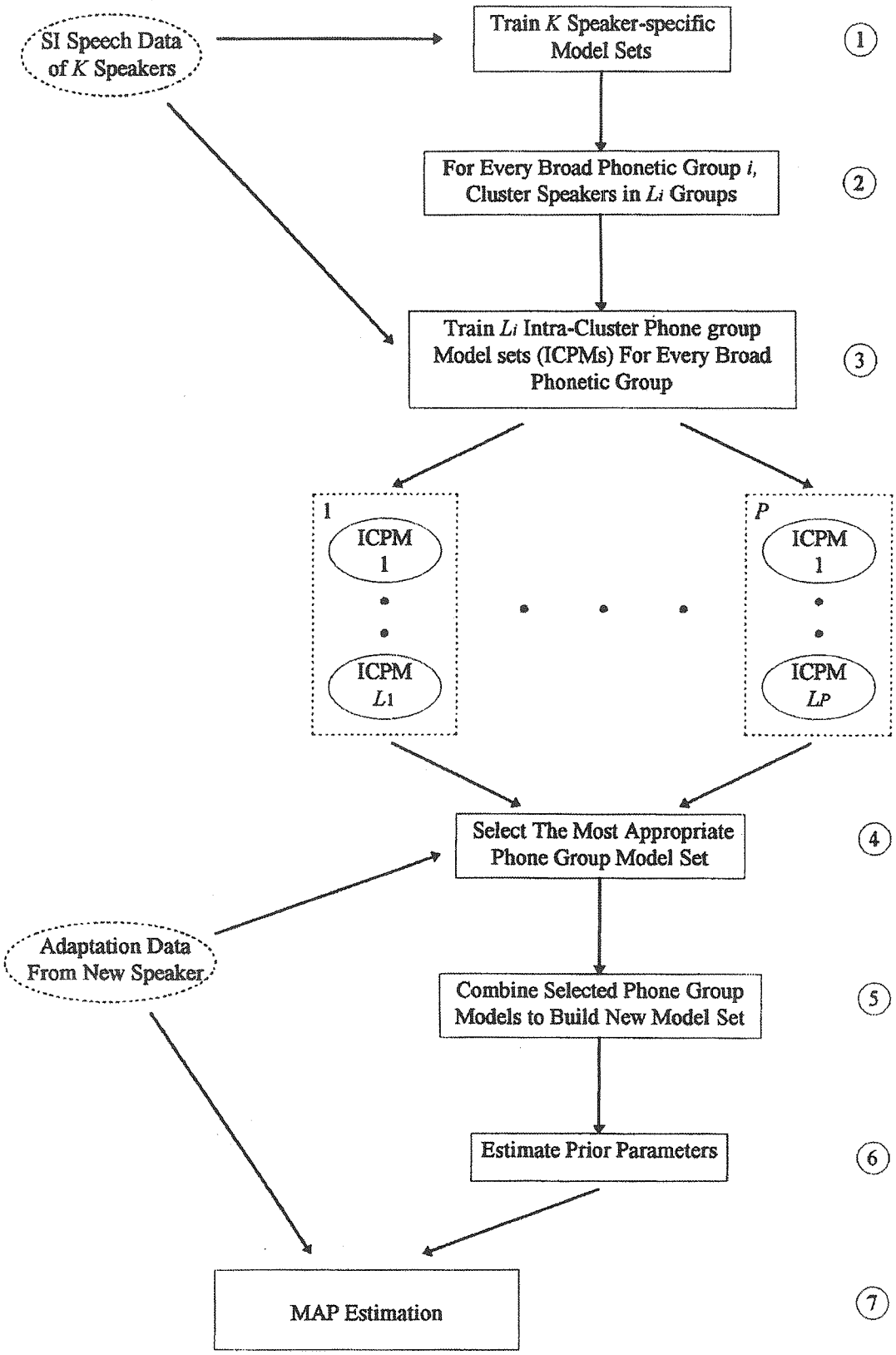

Figure 91) Block diagram for Speaker Phone Clustering Algorithm

groups, 3 clusters per phonetic group were obtained as shown in Table 3.

Note that due to unevenness of the distribution of speakers within the clusters, in few cases, some with very small number of speakers were resulted which were omitted from the list of clusters in order to prevent clusters with small amounts of training data to form. In one case (stops), this resulted in only 2 remaining clusters within a phone group.

Table (3) Number of speakers per phone group cluster.

Phone Group	Vowels	Stops	Nasals	Fricatives	Glides
Cluster 1	30	30	23	30	30
Cluster 2	52	77	56	13	48
Cluster 3	27	-	14	63	17

Table 4 displays the results of SPC-based prior estimated MAP approach to speaker adaptation. All the experiments in this case, i.e. both adaptation and clustering, were performed on single-Gaussian monophones for easier implementation. The table compares the amounts of improvement over the baseline SI system, obtained due to the application of SI-based and SPC-based prior estimated MAP adaptations (called MAP and SPC-MAP respectively). Results show that even without any MAP estimation, the resultant system of SPC shows an improvement of 16.0 percent in word error rate over an SI system. In other words, the application of SPC can provide a better baseline system which naturally leads to a better MAP estimation. SPC-MAP shows a consistent improvement in its results and always outperforms MAP with a noticeable margin. The baseline SI system in these experiments had a word error rate of 22.98%.

Table (4) Percent improvement in recognition word error rate over SI system obtained by speaker adaptation using MAP and SPC-MAP approaches.

Adaptation Sentences	0	10	40	100	600
MAP	0.0	6.2	22.3	31.6	38.9
SPC-MAP	16.0	15.6	30.8	37.8	48.2

5 - Conclusions

Two approaches to the estimation of prior parameters for Bayesian adaptation of CDHMM parameters were introduced and evaluated. The results show considerable improvements in MAP estimation due to the use of any of these two approaches, in comparison to the *ad hoc* prior estimated MAP adaptation. However, the computation or memory overheads in these approaches, in some situations, might even give priority to *ad hoc* approach.

In the use of moments method for prior estimation, one method of implementation, which was also our choice in our experiments, was the use of several speaker dependent systems for calculating sample moments. In comparison to the *ad hoc* approach, this is costly both in the amount of computations needed for training SD systems and in the amount of memory (disk space) needed for saving such trained models. These factors may in some cases lead to the choice of *ad hoc* method due to its simplicity in prior calculations and reasonably

adapting (or training) means, mixture weights and variances of the model output distributions. The averaged word error rate for the baseline SI system in this case was 5. 9%.

Table (1) Percent improvement in recognition word error rate over SI system obtained by speaker adaptation using MAP and MM-MAP approaches, compared to SD system.

Adaptation Sentences	0	10	40	100	600
MAP	0.0	4.9	16.0	38.5	59.8
MM-MAP	0.0	11.3	22.1	40.5	61.3
SD	60.1				

These results indicate a considerable improvement in the results of MAP due to the application of the moments method in prior estimation in comparison to *ad hoc* prior estimation, especially for smaller numbers of adaptation sentences where the effect of prior parameters is dominant, while in larger numbers of adaptation sentences the effect of adaptation data is dominant. However, it can be seen that even with larger amounts of adaptation data, MM-MAP always outperforms (although slightly) the *ad hoc* method.

4 - 2 - Evaluation of Prior Estimation by Speaker Phone clustering

In clustering experiments, speaker-specific model sets were single-Gaussian monophones. This was chosen because of the simplicity and small number of parameters of such models, which can therefore lead to better training of models in very limited training data situation encountered during clustering process. A second benefit of using such models is getting higher speeds in the model training for clustering.

Moreover in clustering experiments reported here, only mean parameters which are believed to have larger influence on the results of adaptation, are updated. The possibility of providing prior parameter estimates for updating other parameters such as covariance matrices or mixture weights, in a Bayesian framework, using the speaker clustering technique has been left for future work.

For speaker phone clustering (SPC) experiments, the 47 phone models of the system were divided into 5 broad phonetic groups, i.e. vowels, stops, nasals, fricatives and glides. These groups and their members are shown in Table 2.

Table (2) The five broad phonetic groups used in speaker phone clustering.

Vowels	aa ae ah ao aw ax ay eh en er ey ih ix iy ow oy uh uw
Stops	b d dd dh dx g k kd p pd t td th ts
Nasals	m n ng
Fricatives	ch f jh s sh v z
Glides	hh l r w y

The clustering process was carried out within these 5 phonetic groups. As mentioned earlier, the clustering thresholds could be chosen to be different within the broad phonetic group. In practice, these thresholds were chosen so that the desired number of clusters within each group were obtained. Eventually, using different values of distance thresholds for different phone

In *ad hoc* prior estimation, as performed by Gauvain and Lee [2] in a so-called segmental MAP estimation framework, some *ad hoc* constraints were applied to prior parameter estimation procedure and the prior parameters were calculated as follows:

$$\psi_{ik} = \hat{\omega}_{ik} \sum_{k=1}^{K} \tau_{ik} \quad (20)$$

$$\mu_{ik} = \hat{m}_{ik} \quad (21)$$

$$\alpha_{ik} = 1/2\,(\tau_{ik} + 1) \quad (22)$$

$$\beta_{ik} = 1/2\tau_{ik}\,\hat{r}_{ik}^{-1} \quad (23)$$

where $\hat{\omega}_{ik}$, $\hat{m}_{ik}$ and $\hat{r}_{ik}^{-1}$ are the mixture weights, means and precision parameters extracted from the SI system. Note that most of the calculated prior parameters depend on the value of τ_{ik}, set for each mixture component. However, in the experiments reported in [2], the value of τ_{ik} was set to a fixed value for all the mixture components. Hence, apart from the mean parameter values, the rest of prior parameters were estimated in an *ad hoc* fashion. Also note that in this approach the adaptation of transition probability parameters was not carried out as their influence on the system performance is negligible.

A similar implementation of the MAP estimation of HMM parameters was performed in order to enable us compare our results with the results of *ad hoc* prior estimation approach. However, in our case, a forward-backward MAP estimation procedure was followed (in place of a segmental MAP estimation one). A value of 10 was used for τ_{ik} in these experiments.

4 - 1 - Evaluation of Prior Estimation by Moments Method

As a systematic approach to prior parameter estimation, the method of moments is expected to show a better performance in comparison to *ad hoc* method since the subjective constraints applied in that approach are allowed to be relaxed. This can be accomplished by the application of equations (13) through (19) to estimate the prior parameters in a CDHMM system. The approach followed to realise the calculations in the above - mentioned equations was to use the parameters from several different speaker dependent systems to calculate the required sample moments.

The results of the application of moments method to prior parameter estimation for Bayesian adaptation (MM-MAP), together with the results of an *ad hoc* MAP implementation (MAP) and SD results are reported in Table 1 as improvements in word error rate over the baseline SI system. The baseline system used in these experiments was a 6-component mixture triphone system discussed earlier. The results of adaptation reported here and in the rest of experiments were obtained by applying the adaptation algorithm with specified number of adaptation sentences to all available speakers of SD database individually and averaging the results. The SD systems indicated in the results were trained using all available 600 sentences per speaker, utilising the normal Baum-Welch training algorithm in similar conditions and the reported results were also obtained by averaging over all 12 available speakers. All results reported in Table 1 were obtained after

MAP estimation is used to overcome, as far as possible, the problem of sparseness of training data. This is not in contradiction with the whole process which is supposed to provide better prior paramerters for MAP estimation process for the purpose of speaker adaptation. However, in this case, the intermediate MAP estimation process is implemented with *ad hoc* prior parameters. These parameters are extracted from a trained SI HMM system which naturally provides better initial models for any new speaker.

In step 2, using the model sets already generated, and dividing the phone models into a number of groups, the speakers are clustered within each group individually according to the similarity of their models in that group, using the same clustering algorithm described in [10]. This results in L_i clusters in any broad phonetic group i. Note that the number of clusters per phonetic group and also the thresholds used in the clustering process can be different for different groups. The *Intra-Cluster Phone Model* sets (ICPMs) are then trained in step 3 using the data from speakers in each cluster.

Hence, L_i ICPMs are trained for any broad phonetic group i. Again, usual MAP estimation technique was used for this purpose, using all data available for that cluster and the resultant models for that phonetic group are extracted from the set of trained models and added to the rest of the models from the baseline SI system to form the ICPM.

In Step 4, the ICPM with the highest average log likelihood per frame for the utterances of the new speaker, i.e. the one to the speech of whom the adaptation process should be applied, is found for any phonetic group i. This specifies the most appropriate cluster in each phonetic group for this speaker. Assuming the number of broad phonetic groups to be P, P such clusters are found and phone group models are extracted once again from each, and then combined in step 5 to build new set of models. In step 6, in place of the baseline SI system, the newly formed model set is used in prior parameter estimation for Bayesian adaptation.

4 - Experimental Evaluation

The evaluation of the adaptation techniques discussed so far was carried out on a 1000 word continuous speech database of English language. The training and adaptation data consisted of two sections: one with more than 100 speakers with about 1.5 to 2 minutes speech of each, while the other one consisted of 12 speakers with about 30 minutes speech of each. The first section described above was used for training SI baseline recognition system, as well as clustering purposes, while the second section was used as SD data for adaptation purposes . For testing purposes, about 5 minutes of speech per SD speaker was available and was used in all experiments.

The parameterisation of the whole database was carried out using 12 mel frequency cepstral coefficients (MFCC), normalised log energy and the first and second differentials of these parameters. Two baseline systems were used in these experiments: a monophone and a word-internal triphone system, both trained with all the available SI training data. The triphone system was developed using the state-clustering algorithm [11].

are the parameters of normal-gamma distribution. The values of E (r_{ikv}), Var (r_{jkv}), E (m_{ikv}) and Var (m_{ikv}) can then be replaced by their corresponding sample moments in order to find the estimates of the above parameters. Similar equaions can also be derived for full covariance matrix case using the normal-Wishart distribution properties.

3 - 2 - Speaker Clustering For Prior Estimation

The method of *Speaker Clustering* (SC) was introduced in [10] . The basic concept in this approach was to provide a better baseline system to be used for each group of speakers in place of the SI system in the *ad hoc* method. In this approach, in place of using a speaker independent speech database to train a set of SI models to be used as the base for prior parameter estimation, it was divided into separate databases for each of the K individual speakers. K different speaker-specific HMM sets were then trained using these databases. The speakers of the SI speech database were then divided into a few clusters, say L, according to the similarity of their speaker-specific model parameters.

For each of these speaker-specific clusters, an intra-cluster HMM system, known *as Intra-Cluster Model set* (ICM), was trained pooling all the speech data available from the speakers of that cluster. This led to L intra-cluster model sets. Once the speech data from a new speaker, to whom HMMs were to be adapted, was available, these utterances (i.e. adaptation data) were used to select the closest ICM to the new speaker. Therefore , a system different from the usual pooled SI system, which was believed to be closer to the new speaker's actual acoustic models, was selected and used in prior parameter estimation process.

A further boost can be given to the results of this approach using the concept of *Speaker Phone Clustering* (SPC). In this case, the same concept of clustering is used in a somewhat wider sense. The basic idea in using the clustering approach in providing better prior parameters for Bayesian adaptation was that some speech features might have been shared among certain speakers. Hence, dividing them to a few groups according to their speech similarities might have been useful. Here, the idea is that among different speakers there may be only some common acoustic features while their other features may not even be close to each other. Hence, it might be useful to group speakers according to their acoustic similarities in phone groups.

This is carried out by dividing the phones into a number of broad phonetic groups hoping that better clusters of speakers with acoustic similarities could be found within these phonetic groups. Examples of these broad phonetic groups are vowels, stop consonants, fricatives, etc.

The block diagram for the implementation of this algorithm is given in Figure 1. In this approach, similar to SC, firstly, a number of speaker - specific model sets are trained. This is carried out in the same way, by using the speech data from K speakers of an SI speech database. As these model sets will be used in clustering speakers, a larger number of such model sets is desirable. This is the main reason for using SI speech data. However, the problem of small amount of training data per speaker, available in such a database, can lead to undertraining. Hence, similar to SC approach,

prior distribution, i.e. if a Bayes decision rule involves a parameter ϕ of a prior distribution, then if ϕ is replaced by any estimate derived from observed data, the resulting rule may be referred to as an empirical Bayes rule [7] [8]. In this case, it is assumed that a current observation **o** is to be used for estimating λ. (**O**, Λ) denotes a sequence of independent sets of past observations and their associated unknown HMM parameters when current observation is made. The actual values of $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ (realisations of Λ) are assumed not to be ever known, but there exists a common prior p.d.f. for them all. The empirical p.d.f. of **O**, f_n (**O**), is an estimate of the marginal p.d.f. of **O**, f (**O** | Λ),, so that as $n \rightarrow \infty$ for every **O**, f_n (**O**) $\rightarrow$ f (**O** |Λ) . Then, it might be possible to find a p.d.f. G (Λ) such that in

$$f(\mathbf{O} \mid \phi) = \int f(\mathbf{O} \mid \Lambda)\, \hat{G}(\Lambda \mid \phi)\, d\Lambda \qquad (10)$$

$\hat{G}(\Lambda \mid \phi) \rightarrow G(\lambda \mid \phi)$ as $n \rightarrow \infty$ [8]. However, due to the difficulty of finding maximum likelihood estimates based on f (**O**|ϕ) in such cases, a simpler approach should be followed.

A few methods have been introduced in literature to overcome this problem [5] [8]. One useful method is the method of moments. In this method, the first few sample moments are equated to the corresponding population moments to obtain as many equations as needed . For example, in this case, the observation sets $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n$ can be used for the estimation of corresponding parameters $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n$, using ordinary parameter reestimation procedures such as Baum-Welch algorithm. These estimated parameters can then be assumed to be observations with density G (λ) as defined in (4).

Then, the Dirichlet distribution properties can be used to find the population moments for initial state probabilities as [4] [9]

$$E(\pi_i) = \frac{\eta_i}{\Sigma_{i=1}^{N} \eta_i} \qquad (11)$$

$$Var(\pi_i) = \frac{\eta_i(\Sigma_{i=1}^{N} \eta_i - \eta_i)}{(\Sigma_{i=1}^{N} \eta_i)^2 (\Sigma_{i=1}^{N} \eta_i + 1)} \qquad (12)$$

Thus,

$$\eta_i = \left\{ \frac{E(\pi_i)[1 - E(\pi_i)]}{Var(\pi_i)} - 1 \right\} E(\pi_i) \qquad (13)$$

Similarly, as the distribution of the state transition probabilities and mixture weights are Dirichlet distributions,

$$\eta_{ij} = \left\{ \frac{E(a_{ij})[1 - E(a_{ij})]}{Var(a_{ij})} - 1 \right\} E(a_{ij}) \qquad (14)$$

$$\psi_{ik} = \left\{ \frac{E(\omega_{ik})[1 - E(\omega_{ik})]}{Var(\omega_{ik})} - 1 \right\} E(\omega_{ik}). \qquad (15)$$

The population moments for the remaining prior parameters can be found in a similar fashion, using the properties of the normal-gamma distribution (i" the case of using diagonal covariance matrix) leading to

$$\alpha_{ikv} = \frac{[E(r_{ikv})]^2}{Var(r_{ikv})} \qquad (16)$$

$$\beta_{ikv} = \frac{E(r_{ikv})}{Var(r_{ikv})} \qquad (17)$$

$$\mu_{ikv} = E(m_{ikv}) \qquad (18)$$

$$\tau_{ikv} = \frac{\beta_{ikv}}{Var(m_{ikv})\, \alpha_{ikv}} \qquad (19)$$

Where all equations are written for every vector element individually and α_{ikv} and β_{ikv}

$$\tilde{r}_{ik}^{-1} = \frac{u_{ik} + \sum_{t=1}^{T} c_{ikt}(o_t - \tilde{m}_{ik})(o_t - \tilde{m}_{ik})' + \tau_{ik}(\tilde{m}_{ik} - \mu_{ik})(\tilde{m}_{ik} - \mu_{ik})'}{\alpha_{ik} - V + \sum_{t=1}^{T} c_{ikt}} \quad (9)$$

where π_i is the initial state probability, α_{ij} is the transition probability from state i to state j and **A** is the $i \times j$ matrix of α's, ω_{ik} is the weight, $\mathbf{m}_{ik}$ is the mean vector and $\mathbf{r}_{ik}$ is the precision matrix of the mixture component k of the state i. Also, in the above equations, η_i, η_{ij} and ψ_{ik} are the components of the parametric vectors of the prior Dirichlet distributions for initial state probability , mixture weights and transition probabilities respectively. τ_{ik}, μ_{ik}, α_{ik} and $\mathbf{u}_{ik}$ are the parameters of the prior normal - Wishart distribution. γ_i is the probability of being in state i initially, given the model and observations, ξ_{ijt} is the probability of being in state i at time t - 1 and state j at time t given the observations and the model, and c_{ikt} is the probability of being in state i and mixture component k at time t given that the model generates the sequence **O**. $\mathbf{o}_t$ is the observation vector at time t and V is the vector size [2].

3 - Prior Parameter Estimation

The prior density defined by (4) was assumed to consist of members of preassigned families of prior distributions . In a pure Bayesian approach, subjective knowledge about the process is also required to enable one to assume a parameter vector of this prior family known. In such cases where the parameters are continuous and multidimensional, however, it is rather difficult to acquire such subjective knowledge [5]. Hence the application of a pure Bayesian approach in this case is extremely difficult.

To alleviate such problems, a few scenarios have been adopted in different research works to perform prior parameter estimation. Gauvain and Lee [2] used the parameters of an already available SI system to estimate their prior parameters in an *ad hoc* fashion. This approach, due to the generality of the SI models, has worked quite well and led to successful results. A fairly similar approach to the previous one, was implemented by Lee *et al.* [6] and consisted of deriving the prior parameters using several *Speaker Dependent* (SD) model sets. Successful results were also reported from the implementation of this approach.

An alternative approach to prior parameter estimation is what is called Empirical Bayes approach and was first introduced as a method in statistical decision problems by Robbins [7]. In this approach a frequency interpretation is given to the prior distribution and some methods such as modified likelihood or moments method are utilised to estimate prior parameters.

In this paper 2 different approaches to prior parameter estimation, which have been applied for the first time to CDHMM - based continuous speech recognisers for the purpose of speaker adaptation, have been discussed and the results are compared to those of speaker adaptation with *ad hoc* prior parameter estimation approach.

3 - 1 - Prior Estimation Using Method of Moments

In the empirical Bayes approach to prior parameter estimation, as stated earlier, a frequency interpretation is given to the

techniques to prior estimation of a CDHMM - based continuous speech recognition system for the purpose of speaker adaptation are discussed.

2 - MAP Estimation of CDHMM Parameters

In speaker adaptation, the system is expected to adapt to a new speaker using limited amounts of adaptation data. Hence, the problem of sparse training data is often faced. This problem, when normal parameter estimation techniques (such as ML) are used, might even lead to substantial system performance degradations [2].

The main difference between MAP and ML estimation is the use of a prior distribution of parameters. If $\mathbf{O} = \{\mathbf{o}_1, ..., \mathbf{o}_T\}$ is a sequence of observations with a p.d.f. $P(\mathbf{O})$ and λ is the parameter set defining the distribution, given a sequence of training data $\mathbf{O}$, λ is to be estimated. The maximum likelihood estimate for λ, if λ is assumed fixed and unknown, can be found by

$$\frac{\partial}{\partial \lambda} P(o_1, ..., o_T|\lambda) = 0 \qquad (1)$$

However, if λ is assumed random with *a priori* distribution function $P_0(\lambda)$, then the MAP estimate for λ is found by solving

$$\frac{\partial}{\partial \lambda.} P(\lambda\, |o_1, ..., o_T) = 0 \qquad (2)$$

Using Bayes theorem, one can write

$$P(\lambda\, |o_1, ..., o_T) = \frac{P(o_1, ..., o_T|\lambda)\, P_0(\lambda)}{P(o_1, ..., o_T)} \qquad (3)$$

Thus, the MAP estimation procedure involves a prior distribution function $P_0(\lambda)$ for the random parameter λ.

Gauvain and Lee [1] have shown that the MAP estimation technique is applicable to the parameters of CDHMMs with mixture output distributions. However, due to statistical limitations, similar to ML estimates, a local maximisation of likelihood function for the observed data should be carried out. This, in the case of ML estimates is usually done using the EM algorithm. Dempster *et al.* [3] have shown that the same algorithm can also be applied to the case of MAP estimation.

Assume a joint prior density of the form

$$G(\lambda) \propto \prod_{i=1}^{N} \left[\pi_i^{\eta i - 1}\, g(\theta_i) \prod_{j=1}^{N} \alpha_{ij}^{\eta ij - 1} \right] \qquad (4)$$

for all the HMM parameters, where $g(\theta_i)$ is the joint conjugate prior density for the parameters of the mixture Gaussian distribution and consists of Dirichlet and normal Wishart distributions [1] [2] [4]. The equations for the MAP estimations of the parameters of a hidden Markov model with $\lambda = (\pi, A, \{\omega_{ik}, \mathbf{m}_{ik}, \mathbf{r}_{ik}\}\ i = 1, ..., N, k = 1, ..., K)$, are derived as

$$\tilde{\pi}_i = \frac{\eta_i - 1 + \gamma_i}{\sum_{j=1}^{N} (\eta_j - 1) + \sum_{j=1}^{N} \gamma_j} \qquad (5)$$

$$\tilde{\alpha}_{ij} = \frac{\eta_{ij} - 1 + \sum_{t=1}^{T} \xi_{ijt}}{\sum_{j=1}^{N} (\eta_{ij} - 1) + \sum_{j=1}^{N} \sum_{t=1}^{T} \xi_{ijt}} \qquad (6)$$

$$\tilde{\omega}_{ij} = \frac{\psi_{ik} - 1 + \sum_{t=1}^{T} c_{ikt}}{\sum_{k=1}^{K} \psi_{ik} - K + \sum_{k=1}^{K} \sum_{t=1}^{T} c_{ikt}} \qquad (7)$$

$$\tilde{\mathbf{m}}_{ik} = \frac{\tau_{ik}\, \mu_{ik} + \sum_{t=1}^{T} c_{ikt}\, \mathbf{o}_t}{\tau_{ik} + \sum_{t=1}^{T} c_{ikt}} \qquad (8)$$

On Prior Parameter Estimation for Bayesian Adaptation of Continuous Density Hidden Markov Models

S. M. Ahadi
Assistant Professor
Electrical Engineering Department
Amirkabir University of Technology

Abstract

Prior parameter estimation has proved to be a rather difficult task in Bayesian adaptation. However, it plays a fundamental role in the process. As Bayesian adaptation is known as a successful candidate for the speaker adaptation task of CDHMMs, finding better priors is of great importance for such applications. Two methods of prior parameter estimation for CDHMMs have been introduced in this paper, and their implementation in a continuous speech recognition system has been discussed. Both techniques have shown capabilities to improve the adaptation results up to more than double the improvements obtained from normal Bayesian adaptation.

1 - Introduction

The application of Bayesian adaptation to the estimation problem of the parameters of *Continuous Density Hidden Markov Models* (CDHMMs) has recently attracted much attention. This technique, also known as Maximum *a Posteriori* (MAP) estimation in such cases, has been found useful in several aspects of speech recognition such as speaker adaptation, context dependent model building, etc. [1]. The reported results show a noticeable improvement over normal training techniques such as *maximum likelihood* (ML) estimation especially under the sparse training data condition.

Although several different successful applications of this technique to such areas in speech recognition have been reported, it still faces an important problem, i.e. the estimation of prior parameters. The application of the prior parameters in the process of estimation is known as the main difference between MAP and ML. In fact, the improvement obtained by MAP estimation in sparse data conditions is mainly due to the use of suitable prior parameters. Hence, finding these suitable parameters is of great importance.

The normal approach to this problem is via an *ad hoc* method which usually estimates the prior parameters of a speech recognition system as functions of the parameters of already available system(s), e.g. a *Speaker Independent* (SI) system [2].

In this paper, several approaches to prior parameter estimation have been introduced and the results of the application of such