Computer Speech and Language (1997) **11,** 187–206

Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models

S. M. Ahadi and P. C. Woodland

Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

Abstract

One problem faced by some model adaptation techniques is that only the parameters of those models which are observed in the adaptation data are updated. Hence, with small amounts of adaptation data most of the system parameters remain unchanged. In this paper, a technique called regression-based model prediction (RMP), which tries to overcome this problem, is presented. This technique tries to adapt the model parameters of a continuous density hidden Markov model set which has insufficient adaptation data when used with maximum *a posteriori* (MAP) estimation. The technique uses the parameters of better estimated models and a set of parameter relationships between the model parameters to update the parameters of models with insufficient adaptation data. The parameter relationships are found by applying linear regression to a number of speaker-specific model sets.

Experiments using both MAP estimation and RMP are presented using the ARPA RM1 continuous speech database and RMP has been found to be useful in improving the system performance with as little as 3 s of adaptation speech. RMP has been shown to consistently improve the results obtained by MAP. When a very large number of adaptation sentences are used the error rates converge towards those of MAP. It is shown that RMP gives an improvement of 8·8% over the baseline error rate with a single adaptation sentence, and 27% with 40 adaptation sentences.

© 1997 Academic Press Limited

1. Introduction

Although modern speaker-independent (SI) continuous speech recognition systems show impressive performance, error rates are still much higher than a well-trained speaker-dependent (SD) system. Speaker adaptation techniques that attempt to adapt the parameters of an SI system to get SD performance with only a small amount of speaker-specific data are therefore of interest. A key issue for such techniques is extracting the maximum information about the new speaker from a limited amount of data.

0885–2308/97/030187+20 $25.00/0/la970031

© 1997 Academic Press Limited

A standard adaptation approach for continuous density hidden Markov model (HMM) systems uses maximum *a posteriori* (MAP) parameter estimation (Lee, Lin & Juang, 1991; Gauvain & Lee, 1994) which combines estimates obtained from the adaptation data with prior parameter estimates from a speaker-independent model. However, in this approach only distributions for which observations occur in the adaptation data are updated. This problem is particularly severe in large vocabulary speaker-independent systems since such a system may contain millions of parameters. This means that techniques such as MAP require a relatively large amount of adaptation data before they are effective. One approach (Leggetter & Woodland, 1995) to this problem trains a small number of regression matrices on the available adaptation data and transforms all the mean vectors in the system using one of these matrices. However, this technique is restricted to fairly broad adjustments to the parameter values, and requires several adaptation sentences before it starts to be effective.

This paper investigates an alternative approach based on learning relationships between the model parameters for training data. Model parameter relationships have previously been used in other adaptation schemes. Stern and Lasry (1987) used inter-model relationships in a MAP estimation approach for speaker adaptation called Extended MAP (EMAP). Furui (1980) introduced a technique based on the use of linear regression in a predictive fashion in order to use a fraction of a proposed vocabulary for training an isolated word recognition system. Another predictive approach to speaker adaptation was investigated by Cox (1993, 1995). It was shown that correlations between different sounds can be used to predict the parameters of some models from those of other models. The technique was applied to a system of single Gaussian HMMs for a small vocabulary isolated word recognition task.

In this paper, RMP for rapid speaker adaptation is introduced. The RMP method can be viewed as an extension of Cox's work to large vocabulary continuous speech recognition using context-dependent mixture Gaussian models containing more orders of magnitude more parameters. In order to accommodate this change in focus the method has been extended and refined. Linear relationships between particular system parameters are computed from speaker-specific models using multiple regression. These parameters are later used in the adaptation phase to update parameter values for distributions not observed in the adaptation data. This updating process uses a small number of well-adapted distributions (source distributions) to predict suitable parameter values for the unseen or poorly adapted distributions (target distributions) using the linear-regression derived relationships. A Bayesian approach is then used to combine the target predictions with prior parameter estimates.

In this paper the MAP estimation technique and the problem of prior parameter estimation are briefly reviewed and the basic ideas behind RMP are presented. This is followed by a description of RMP theory, implementation and a number of improvements over the basic RMP scheme. Finally, there is an experimental evaluation of both MAP and RMP.

2. Bayesian adaptation of CDHMMs

Speaker adaptation normally requires that a system adapt to a new speaker using a limited amount of adaptation data. Maximum *a posteriori* (MAP) estimation (Lee *et al.*, 1991; Gauvain & Lee, 1994) is a standard approach to overcome the grave effects that the use of sparse training data can cause in a maximum likelihood (ML) framework if a large number of parameters are to be estimated. The MAP technique will be briefly discussed in this section.

2.1. MAP estimation of continuous density HMM parameters

Given an adaptation sequence $\mathbf{O}=(\mathbf{o}_1, \ldots, \mathbf{o}_T)$ for a HMM with output distributions consisting of a mixture of Gaussian densities, the MAP re-estimate of the Gaussian means for HMM state i, mixture component k is given by (Gauvain & Lee, 1994)

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

where μ_{ik} is the prior mean, τ_{ik} is a parameter controlling the relative weight of the prior and adaptation data, $\mathbf{o}_t$ is the adaptation data samples and c_{ikt} is the probability of occupying state i and mixture component k at time t given that the model generates the sequence $\mathbf{O}$. New estimates for the other HMM parameters can also be found using MAP (Gauvain & Lee, 1994).

2.2. Prior parameter estimation

A key issue in the use of the MAP technique is the estimation of a set of prior parameters. The use of prior parameters is the basic difference between MAP estimation and other estimation techniques such as ML. However, finding a suitable set of priors has proved to be a rather difficult problem in many cases.

For MAP estimation of continuous density HMM (CDHMM) parameters for the purpose of speaker adaptation, the parameters of a baseline SI system have been used to find the prior parameters (Lee & Gauvain, 1992). For the mean parameters, the prior mean is set to the corresponding SI mean parameter, while the τ value is determined experimentally. This *ad hoc* approach provides an easy and effective method for finding the prior parameters.

One alternative approach to prior estimation was investigated in the application of MAP estimation to discrete and semi-continuous HMMs (Huo & Chan,1992). The prior parameters were estimated in an empirical Bayes framework using the method of moments to derive equations for the prior parameters. In this method, the first few sample moments are equated with the corresponding population moments to obtain as many equations as unknown parameters.

If CDHMMs with diagonal covariance matrices are used, which is often the case, the use of moment-based prior estimation gives (Ahadi, 1996)

$$\mu_{ikv}=E(m_{ikv}) \tag{2}$$

$$\tau_{ikv}=\frac{1}{\mathrm{Var}(m_{ikv})E(r_{ikv})}, \tag{3}$$

where $\mathrm{E}(m_{ikv})$ and $\mathrm{Var}(m_{ikv})$ are the population expectation and variance of the vth vector element of the mean vector and $E(r_{ikv})$ is the expectation of the vth vector element of the covariance vector. In order to find the estimates of the above parameters, one can replace $E(r_{ikv})$, $E(m_{ikv})$ and $\mathrm{Var}(m_{ikv})$ with their corresponding sample moments.

A classical model training procedure such as the Baum–Welch algorithm can be used to find initial estimates of the model parameters from past observation sets. Ideally, this would estimate a set of speaker-dependent systems which can then be used to find the sample moments of $E(m_{ikv})$, $E(r_{ikv})$ and $\mathrm{Var}(m_{ikv})$.

3. RMP background

The use of regression-based prediction to improve the performance of speech recognition systems has been investigated by both Cox (1993, 1995) and Furui (1980). In this section, the application of such predictive techniques to a CDHMM-based large vocabulary continuous speech recognizer will be discussed.

3.1. Model parameter relationships

When only very limited adaptation data is available even MAP estimation may not be able to give reliable estimates of the system parameters. In these cases, although a small amount of adaptation data is available from a new speaker, it is difficult to exploit.

One approach is to try and find relationships between the parameters of different models for a speaker. If a simple linear relationship between model parameters is assumed, linear regression can be used to estimate the relationships between the acoustic model parameters. Consider a space in which each dimension represents a particular model parameter (e.g. a state output Gaussian distribution mean vector element), then a point would represent the parameter values for a certain speaker and for K speakers (i.e. speaker-dependent systems) there are K points in this space. Assuming that a simple linear relationship exists between these parameters, a line can be found to approximate this set of points.

Since, in practice, not all the points will lie on this line, the actual relationship, written over only a single element of the vector of mean parameters between just two parameters, is of the form

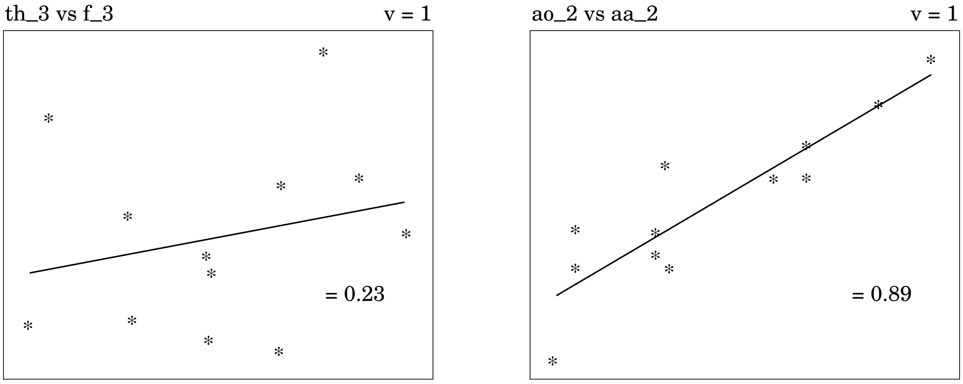
$$y = b_1 x + b_0 + \varepsilon, \tag{4}$$

where y and x represent the two parameters whose relationship is to be found, ε is the error associated with this approximation and b_1 and b_0 are the regression parameters relating these two parameters. If a least squares approach is taken, the parameters are found by minimizing

$$\sum_{k=1}^{K} \varepsilon_k^2 = \sum_{k=1}^{K} (y_k - b_1 x_k - b_0)^2, \tag{5}$$

where K is the total number of regression points (speakers). One of the parameters (here y) will be called the *target* parameter, or an element of the parameter vector which is to be predicted and the other (x) will be called the *source* parameter.

For simple linear regression, a correlation coefficient can be calculated between the target and source parameters, which could be interpreted as an index for the linearity of this relationship. The square of the correlation coefficient is given by (Chatterjee & Price, 1991)



(a) (b)

Figure 1. Scattergrams of mean vector element relationships between the first vector elements of (a) the third states of monophone models for /th/ and /f/ with a correlation coefficient of 0·23 and (b) the second states of monophone models for /ao/ and /aa/ with a correlation coefficient of 0·89.

$$\rho_v^2 = \frac{[\Sigma_{k=1}^{K} (x_{kv} - \bar{x}_v)(y_{kv} - \bar{y}_v)]^2}{\Sigma_{k=1}^{K} (x_{kv} - \bar{x}_v)^2 \, \Sigma_K (y_{kv} - \bar{y}_v)^2,} \tag{6}$$

where ρ_v is the sample correlation coefficient for the vth element of the $\mathbf{y}$ vector and the vth element of the $\mathbf{x}$ vector and $\bar{x}_v$ and $\bar{y}_v$ represent the average of xs and ys over all the data points.

As an example of the application of linear regression, Figs 1(a) and 1(b) display the mean vector element relationships between two potential source and target model components in a monophone single Gaussian HMM system, with five states per model (including entry and exit states), trained on the ARPA RM1 continuous speech corpus. Twelve-speaker model sets were used to construct the scattergrams displayed in these figures. In Fig. 1(a), the relationships are shown for the first mean vector elements of the third states of phone models of /th/ and /f/, and the relationships in Fig. 1(b) are for the first mean vector elements of the second states of phone models of /ao/ and / aa/. As can be seen, in Fig. 1(a), the absolute value of the correlation coefficient among these parameters is small, i.e. the parameters are far from forming an exact linear relationship, while in Fig. 1(b), the absolute correlation coefficient is much higher and the points are closer to the fitted line. Obviously, the number of regression points (speakers) is very important for constructing these relationships, since a larger number leads to a better estimate of regression parameters.

These relationships can be used in the RMP framework. If enough regression data points are available, a linear approximation can be found for the relationships between different parameters of a model set. Once these relationships are found, if only some of the parameters can be estimated directly from the data, the remaining parameters may be estimated using these relationships.

3.2. Use of parameter relationships in continuous speech recognition

Once the relationships between model parameters have been found, they still need to be applied to a continuous speech recognizer. The approach followed by Cox (1992, 1993, 1995) was to apply this technique to a limited vocabulary of isolated words (English alphabet) by allocating one model with 10 states to each word in the vocabulary and dividing the whole set of vocabulary items (models) into two sets and using one set to predict the parameters of the other set. This approach was very successful in adaptng the recognition system to new speakers. However, this technique was designed to cope with an isolated word task and a very limited vocabulary. The models were divided into two groups so that only the adaptation data for source models (a fixed subset of models) was useful and only the other fixed subset of models were predictable. These constraints severely limit the use of this predictive technique for speaker adaptation in large vocabulary continuous speech recognition systems.

For the model prediction technique to be applicable to medium to large vocabulary continuous speech recognizers the following issues must be addressed:

- The technique should be applicable to CDHMMs with mixture Gaussian output distributions, which are often used for such tasks.
- Due to the nature of continuous speech, modelling is usually at phone level. A hard division of the models into sources and targets is not possible since the appropriate distribution of phone models in the adaptation data will, in general, be impossible to achieve.

The second factor implies that the best correlated source state distribution cannot always be used to predict target state distribution parameters since adaptation data may not include data for that particular source distribution. Also, both distributions that do not receive any adaptation data and those with small amounts of data should be considered as potential target distributions.

4. RMP theory

In the RMP approach, given a set of models adapted, for example by MAP, and a set of regression parameters relating the sources and the targets, the target parameters can be updated.

The use of the basic simple regression formula [Equation (4)], will lead to a regression-based estimate of the target parameters. However, to obtain a more robust set of target parameters, it may be necessary to use a more sophisticated estimate.

The MAP estimated target parameters can be considered as giving *a priori* knowledge of the final estimate of the target mean parameters, and a Bayesian framework is suitable to combine the parameter estimates from linear regression with this prior. Hence, the problem is to calculate the posterior probabiity $p(m|\mu)$, taking into account the prior density. Note that in this case, μ consists of a single observation, which is the regression-predicted target mean, and the mean of the normal prior density is considered to be the MAP estimated target mean parameter, ζ.

If the distribution of the means are assumed Gaussian, then the final estimate of a mean element m of a target state distribution will be of the following form (Duda & Hart, 1973)

$$\hat{m} = \mu \frac{s_\zeta^2}{s_\zeta^2 + s_\mu^2} + \zeta \frac{s_\mu^2}{s_\zeta^2 + s_\mu^2}, \tag{7}$$

where μ is the regression estimated mean and the associated variance is denoted s_μ^2 and the prior (MAP estimated) mean is ζ and its associated variance is s_ζ^2.

The variance parameter associated with the regression predicted mean consists of two parts. The first part is the variance due to the application of linear regression. If the distribution of mean elements over different SD speakers during the regression parameter calculation state are assumed to be normal, then the estimated sample variance for the target parameters due to linear regression, s_e^2, can be calculated as[1]

$$s_e^2 = s_y^2(1-\rho^2)$$
$$= \frac{1}{K-1}\sum_{k=1}^{K}(y_k-\bar{y})^2 - \frac{b_1}{K-1}\sum_{k=1}^{K}(y_k-\bar{y})(x_k-\bar{x}), \tag{8}$$

where s_y^2 is the sample variance for the target element in the set of SD systems.

There is another component of the total variance of the target parameters due to the fact that the source distribution values used to find targets using regression are not speaker-dependent values but their MAP estimates based on a small amount of adaptation data. This additional variance, s_v^2, due to error in the source distribution estimates, can be calculated by finding the sample variance of the MAP estimated parameters compared to the true SD parameters. For each element of the state distribution mean vector, it is computed as

$$s_v^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left[(x_k-v_k) - \frac{1}{K}\sum_{j=1}^{K}(x_j-v_j)\right]^2, \tag{9}$$

where v_k denotes the MAP estimated source mean elements for the SD speakers, and x_k are the actual SD source elements for these speakers, i.e. the real SD and MAP estimated parameters for all available source speakers are used to estimate s_v^2 for any new speaker. Obviously, these errors in the source parameter elements and the variance due to them will decrease with an increase in the amount of adaptation data, which leads to a better MAP estimate of the source parameters.

The total variance of the target parameter can now be found. The variances calculated in Equations (8) and (9) are combined, noting the dependence on b_1 (Lindley, 1947), to yield:

$$s_\mu^2 = s_e^2 + b_1^2 s_v^2, \tag{10}$$

which is then used in Equation (7).

Next, it is necessary to compute s_ζ^2 which is the variance of the MAP estimated target parameters. The same approach used to find the variance of the source element, s_v^2, in Equation (9), can also be applied to find the variance for a target element, leading to

$$s_\zeta^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left[(y_k-\zeta_k) - \frac{1}{K}\sum_{j=1}^{K}(y_j-\zeta_j)\right]^2. \tag{11}$$

[1] The equation used here is, in fact, an approximation to the variance due to linear regression and in practice, gives a slight underestimate of its value.

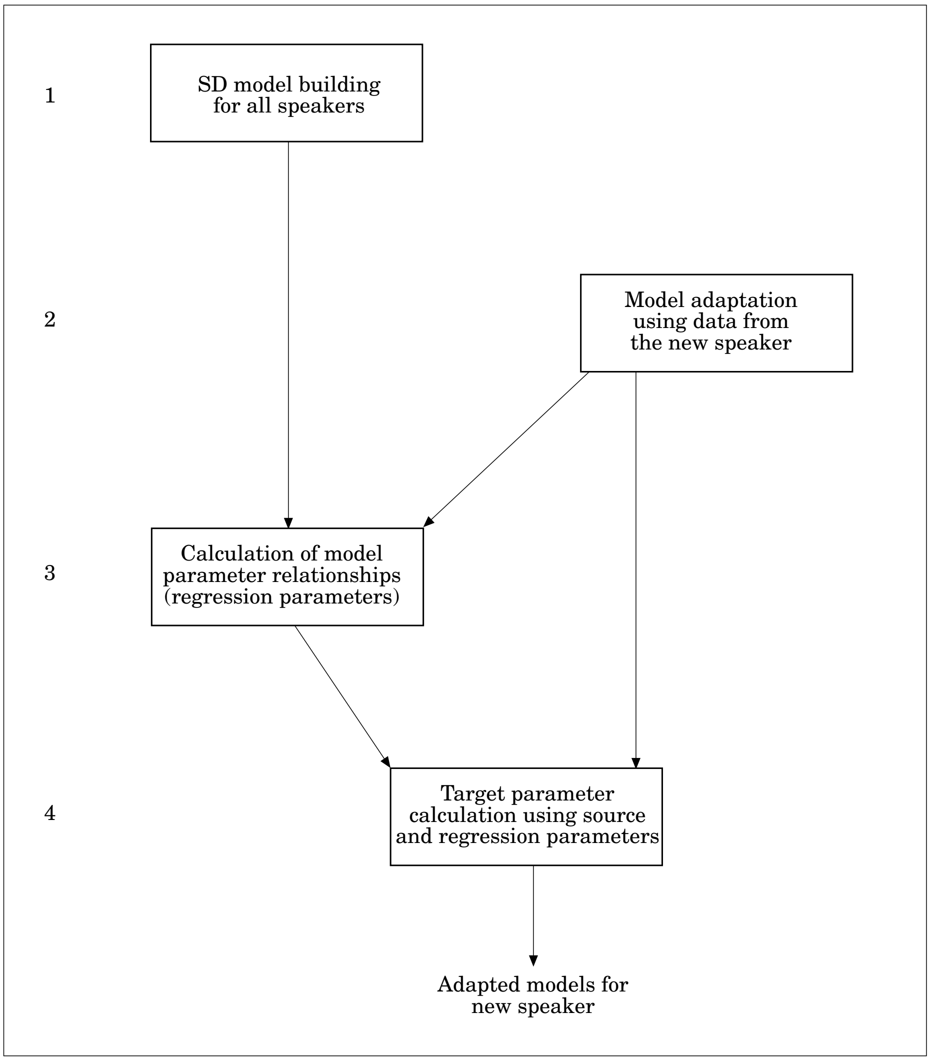

1
SD model building for all speakers
2
Model adaptation using data from the new speaker
3
Calculation of model parameter relationships (regression parameters)
4
Target parameter calculation using source and regression parameters
Adapted models for new speaker

Figure 2. Basic system block diagram for an RMP-based speaker adaptation system.

Finally, using the variances s_μ^2 and s_ζ^2 and the target component mean values found by regression with the MAP estimated element as a prior, the final estimation of the target parameter vector elements can be found by Equation (7).

5. RMP implementation

5.1. Overall approach

Figure 2 shows the basic system block diagram of the RMP-based speaker adaptation system. The first step is to find the model parameter relationships. A number of speaker-specific model sets are needed for this purpose. A larger number of speakers leads to more accurate estimation of the regression parameters which will lead to improved model prediction.

In the second step, the adaptation utterances from a new speaker are used in a model adaptation framework (e.g. MAP) and all the model parameters for which enough adaptation data is available are updated using the appropriate adaptation data.

In the third step, the correlation parameters between all source and target state distributions of interest (using SD parameters) are calculated and the best correlated source distribution for each target is found. Then, for all these best correlated sources and targets, all required regression parameters together with some other parameters needed for later calculations are calculated and saved. During the fourth step, all the regression-related parameters and the initially adapted models are used to further adapt target state distribution parameters using the source distribution parameters giving a new set of target parameters for the system.

5.2. *SD model building*

A number of speaker-dependent systems must be available before the regression parameters can be calculated. Baum–Welch re-estimation is usually used to train such systems in a maximum likelihood framework, using as much training data as is available from each speaker. Gauvain and Lee (1992) achieved similar results using both MAP and ML estimation approaches for training SD systems. Huang *et al.* (1993) have reported better performance compared to standard SD training, using MAP estimation, unless the number of training sentences is very large. Although MAP estimation results asymptotically approach ML estimation results with very large amounts of training data it is unclear what amount of data should be considered large enough. This is mostly a matter of system architecture and number of system parameters, and so in these experiments, an iterative MAP estimation approach (Ahadi, 1996) was used to train the K speaker-dependent systems.

5.3. *Preliminary model adaptation*

The model parameters for which there is adaptation data are first updated. MAP estimation is used for this purpose using the priors obtained from a speaker-independent system. Since the RMP approach adapts only the mean vectors, only the mean values, as in Equation (1), are estimated.

In order to be able to separate source and target states the amount of adaptation each state distribution has received needs to be known. The parameter c_{it} in Equation (1) is the *state occupation probability* at time t. Thus, $\Sigma_{t=1}^{T} c_{it}$ will represent the total number of times the state was occupied, i.e. the amount of adaptation data each HMM state has received during this model adaptation process. These state occupation "counts" are also calculated and saved during this MAP estimation stage.

5.4. *Regression parameter calculation*

For regression and other associated parameter calculations, K speaker-dependent sets of models, trained in step 1 of Fig. 2, are used. Also, the state occupation counts are used to determine the suitable state distributions of the models which could be used as sources and targets by applying a threshold. After partitioning the whole set of state distributions into two, the squared-correlation coefficients between all mean vector elements of the source distributions with the corresponding vector elements of the target distributions are found using Equation (6) and averaged over all the vector elements for each pair of distributions. This provides a set of averaged squared-correlation coeffficients for each target element with all the source elements. The best of these squared-correlations for each target component are found and the corresponding source component is marked as the best matching source component.

At this stage, to prevent weakly correlated matches between source and target state distributions taking place, a correlation coefficient threshold, T_c, is also applied. Thus, any matches with averaged squared-correlations less than the threshold are rejected and not used in the regression.

The next step is the calculation of the regression parameters. For the case of simple linear regression, the regression parameters are calculated by applying the least squares method which leads to the following equations

$$b_1 = \frac{\Sigma_{k=1}^{K}(x_k - \bar{x})(y_k - \bar{y})}{\Sigma_{k=1}^{K}(x_k - \bar{x})^2} \tag{12}$$

and

$$b_0 = \bar{y} - b_1\bar{x}. \tag{13}$$

The above parameters are calculated for each mean vector element of each target state distribution with the corresponding source state distribution.

5.5. *Target parameter prediction*

The last stage in the application of the RMP technique to speaker adaptation is the prediction of target parameters. This is done as described in Section 4. The variances required at this stage are found using SD models. The parameters needed for these calculations are not only the speaker-dependent mean parameters, but also the mean parameters of the MAP estimated models for all the SD speakers with similar adaptation data. These calculations are performed, and the resultant variance values saved, in the previous stage of RMP (step 3 in Fig. 2). Although not shown in the figure, a MAP estimation stage is also needed for all K SD speakers with the same adaptation data used from the new speaker so that the variance calculations can be carried out.

6. Improved RMP adaptation

In this section a number of changes to the basic RMP approach are discussed.

6.1. *Mixture Gaussian modelling*

The method discussed so far has not included mixture distributions in its scope. A simple extension to the case of mixture distributions is to treat all the mixture components of a single HMM state as parts of that state and still continue to find inter-state relationships. In this case, a target HMM state of M mixture components of V vector elements each would have $M \times V$ elements and the regressions are formed using all these elements and the corresponding ones from the states of other HMMs. Although such an approach can satisfy the basic idea of RMP, it is restricted by inter-state relationships. In practice, however, it has been found that basing relationships at a mixture component level can lead to superior performance, i.e. each Gaussian distribution is individually related to another one based on the average correlation between their mean vector elements. This also necessitates the use of mixture component occupation counts for source–target separation.

6.2. Multiple regression

So far, simple linear regression has been used for all parameter predictions and correlation calculations. Multiple regression can be more effective if enough data has been provided for robust estimation of regression parameters. For a multiple linear regression, Equation (4) becomes

$$y = b_0 + \sum_{l=1}^{P} b_l x_l + \varepsilon, \tag{14}$$

where P is the order of regression (number of source distributions for each target) and b_l and x_l are regression coefficients.

One issue in this case is the appropriate regression order. One possibility could be the use of a full regression matrix of the size V, so that

$$\boldsymbol{y} = \boldsymbol{b}_0 + \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varepsilon}. \tag{15}$$

The above equation can be interpreted as a transformation relating each element of the target distribution mean vector to all the elements of the source. Such transformations, as introduced in Leggetter and Woodland (1995), can be successful in certain applications, but require larger amounts of data for parameter estimation due to the larger numbers of parameters. In the work presented here, every element of the target vector was related to similar elements of several different source components, not different elements of a single source component. The application of multiple regression in the case of RMP is highly dependent on the number of SD speakers (regression points) available, which is an important limiting factor. However,it has been found that even with a relatively small number of SD speakers multiple regression can still be useful and improve the system performance.

In this case, Equation (14) can be used to find a linear relationship between the target component vector element and multiple source elements. The regression coefficients, in this case, can be calculated by solving the following matrix equation (Chatterjee & Price, 1991)

$$\boldsymbol{U}\boldsymbol{b} = \boldsymbol{w}, \tag{16}$$

where $\boldsymbol{U}$ is a $P \times P$ matrix and $\boldsymbol{b}$ and $\boldsymbol{w}$ are $P \times l$ vectors. The elements of $\boldsymbol{U}$ and $\boldsymbol{w}$ are given by

$$U_{nl}=\sum_{k=1}^{K}(x_{nk}-\bar{x}_n)(x_{lk}-\bar{x}_l) \tag{17}$$

$$w_l=\sum_{k=1}^{K}(y_k-\bar{y})(x_{lk}-\bar{x}_l) \tag{18}$$

and the value of b_0 is found by

$$b_0=\bar{y}-\sum_{l=1}^{P}b_l\bar{x}_l. \tag{19}$$

The squared multiple correlation coefficient between the target y, and the regression predicted target value y', found using the regression formula, is given by

$$R^2_{yy}{}'=\frac{\Sigma_{l=1}^{P}b_l\,\Sigma_{k=1}^{K}(y_k-\bar{y})(x_{lk}-\bar{x}_l)}{\Sigma_{k=1}^{K}(y_k-\bar{y})^2.} \tag{20}$$

In this case, the variances introduced in Section 5.5 should be calculated taking into account multiple source distributions. Thus, the estimated sample variance due to multiple regression, assuming once again that the mean element values are normally distributed over different speakers, can be written as (Dunn & Clark, 1987)

$$\begin{aligned}s_e^2&=s_y^2(1-R^2_{yy}{}')\frac{K-1}{K-P-1}\\&=\frac{1}{K-P-1}\left[\sum_{k=1}^{K}(y_k-\bar{y})^2-\sum_{l=1}^{P}b_l\sum_{k=1}^{K}(y_k-\bar{y})(x_{lk}-\bar{x})\right].\end{aligned} \tag{21}$$

Also, the additional sample variance in target parameters due to the errors in the source distribution parameter estimates are calculated for each element of each source distribution as follows

$$s_{v_l}^2=\frac{1}{K-1}\sum_{k=1}^{K}\left[(x_{lk}-v_{lk})-\frac{1}{K}\sum_{j=1}^{K}(x_{lk}-v_{lk})\right]^2, \tag{22}$$

where v_{lk} denote the MAP estimated source elements. Therefore the total variance for the regression predicted target value, assuming for the sake of simplicity that the source distributions are independent, can be written as

$$s_\mu^2=s_e^2+\sum_{l=1}^{P}b_l^2 s_{v_l}^2. \tag{23}$$

In this case, the sample variance for the initial MAP estimated target parameters is the same as in Equation (11) and, finally, Equation (7) should be used to find the final estimate for any target mean vector element.

6.3. *Dynamic setting of regression order*

In applying multiple regression to RMP, the order of regression is left free to change dynamically. This is due to the fact that as a result of applying a correlation threshold to the process of finding matching targets and sources, in some cases, there may not be P source components available with averaged squared-correlation coefficients higher than the correlation threshold for a single target. In such cases, that target component is allowed to use a lower order of regression equal to the number of source components available for that target, satisfying the above condition. This approach leaves targets free to have any number of source components up to a maximum of P, hence increasing the chance for all targets to find at least one source component for the purpose of prediction.

6.4. *Multiple thresholds for distribution separation*

The application of a state or mixture occupation threshold for dividing distributions into two groups of sources and targets has been discussed. Since the threshold divides the whole set of components into two sets of source and target distributions, it is very likely that components close to the threshold have very similar occupation counts, while some of them are used as sources and others as targets. Also, there is no limit during the whole adaptation process as to which source component is used to estimate the parameters of which component. Consequently, some source components with the occupation counts very close to the threshold might be used to estimate some target components which are themselves not too far from the threshold. This may result in source–target pairs with similar occupation counts which will not lead to reliable target estimates.

To overcome this problem and in order to have more reliable estimates, an occupation count gap is introduced between source and target parameters by using two thresholds known as higher thresholds (T_H) and lower threshold (T_L). Therefore, some of the components between these two thresholds, for the sake of reliability, are left unused in RMP and the components with lower occupation counts than T_L are used as targets and those with higher occupation counts than T_H as sources, guaranteeing that there exists a minimum difference of occupation counts between source and target components.

6.5. *Other changes to the basic approach*

RMP, as presented above, requires a large number of calculations, especially in the stage of regression parameter calculation. For a system with S source components and T target components, $T \times S$ correlation calculations need to be carried out before finding which sources are appropriate for each target.[2] Then about T regression and variance calculations are needed, if all the targets are to be estimated using available

[2] In fact, the total number of correlation calculations would be much higher still if the best N sources were selected for each target via calculation of multiple correlation coefficients.

sources. While the second set of calculations are almost unavoidable, the first set, which usually constitute the main part of calculations, can be reduced.

Investigations on the relationships between the distributions belonging to different phones, and/or different states of similar phones, reveal that the number of relationships with high correlations between distributions from different model state positions are few and far between. Thus, a constraint can be set to limit the correlation calculations to only those distributions of similar states from different models. This condition can reduce the amount of computation required at this stage to a third, without causing a major degradation to the overall result.

Another point to consider is that most of the relationships usually take place within fairly broad phonetic groups. Hence, there is usually no need to look for inter-group relationships. This can also be very helpful in reducing the computation further. A very broad classification could be the division of the phones into vowel and consonant groups. Also, more specific phonetic groups can be used, but increasing the number of such groups may lead to performance degradation to some extent.

7. Experimental evaluation

This section concentrates on the experimental evaluation of the RMP and MAP adaptation techniques. The ARPA resource management database RM1 was used for this evaluation, and the HTK HMM toolkit was used for all model building and recognition. Special tools were written for forward–backward MAP estimation and model prediction.The RM1 database was parametrized using 12 Mel frequency cepstral coefficient, normalized log energy and the first and second differentials of these parameters.

The phone set and dictionary used were those produced by CMU and listed in Lee (1989). The basic phone set consists of 47 phone symbols plus silence and the dictionary defines a single pronunciation for each of the 991 words used in the RM task. For each phone, a five-state HMM (including the non-emitting entry and exit states) with a left-to-right topology and mostly without skip transitions, were defined. The baseline SI state-clustered word-internal triphone gender-independent system was trained, using the RM SI-109 data, by several iterations of the embedded Baum–Welch re-estimation procedure. The state clustering procedure used the decision tree method described in Woodland, Odell, Valtchev and Young (1994). Both single Gaussian and mixture Gaussian versions of this SI system were trained and these SI model sets were used as a base for all experiments.

For building SD model sets, 600 RM SD training sentences were used. All word error rates were computed using 100 test sentences for each speaker and then averaged over the 12 RM SD speakers. The standard word-pair grammar was used for all recognition tests. SD systems were trained using both maximum likelihood (ML) estimation with the SI set as the initial models and MAP estimation with the SI system used both as initial models and for prior parameter calculation. An iterative MAP approach was followed in the latter case (Ahadi, 1996). The MAP SD systems give lower error rates than the ML models and were used to find the correlations between models and hence the regression parameters.

For the adaptation experiments, a portion of the SD training data from each speaker was used as adaptation data with the sentences taken in order from the database. All adaptation experiments reported were carried out in batch (or static) supervised mode.

Table I. Percentage of the system mean parameters (six-component mixture Gaussian word-internal triphone models) receiving any adaptation data with different numbers of adaptation sentences

Adaptation sentences	Percentage adapted means
1	5·6
10	31·8
40	64·3
100	82·4
600	97·0

7.1. Evaluation of RMP

As pointed out in Section 3, the RMP method is designed to improve the performance of a HMM system already adapted to a speaker, using a MAP or similar model parameter adaptation technique. As noticed in the previous section, the asymptotic convergence of the performance of MAP estimation, together with its ability to provide improvements with fairly small amounts of training data, makes it a desirable technique for speaker adaptation. However, RMP is designed to further improve the performance of such MAP estimated systems, especially when only a very small amount of adaptation data is available.

As an example, the percentage of system parameters receiving any adaptation in a MAP estimated six-component mixture Gaussian tied-state word-internal triphone system is presented in Table I. This shows that a large number of model parameters do not receive any adaptation data with small numbers of adaptation sentences. Furthermore, even among the small number of adapted parameters in such cases, the amount of adaptation data a parameter receives is small. Thus, in the case of a small number of adaptation utterances, only a very small percentage of the system parameters receive the minimum amount of adaptation data needed for improving performance. However, for the same system adapted using RMP, 96% of the mean parameters are updated after just a single adaptation sentence.

The RMP adaptation algorithm has been applied to several context-dependent systems to assess its performance in several different conditions. These include single- and six-component mixture Gaussian tree-based state-clustered word-internal triphone systems. Here, due to the tying of the HMM states, the prediction algorithm is applied to the tied parameters.

For these experiments, a correlation coefficient threshold of 0·4 and a maximum regression order of 2 is used. These have been found to perform well under the above test conditions. The 12 SD speakers available are used in a “leave one out” manner, so that in any one pass, one speaker could be used for test purposes and the remaining 11 for regression parameter calculations. This procedure, which is utilized due to the limited number of speakers available, which should be used for both regression parameter calculations and adaptation and test purposes, has been repeated for all 12 speakers so that an average of the results could be obtained.

The results of the application of RMP to single- and six-component mixture Gaussian triphone systems are shown in Fig. 3, together with the corresponding MAP, SI and SD results. Note that all reported MAP and SD results are obtained using models trained by running a single iteration of either the Forward–Backward MAP estimation (for MAP) or the Baum–Welch algorithm (for ML), updating only the Gaussian mean parameters, and hence are different from the iterative MAP and ML estimated SD systems mentioned in Section 5. The baseline SI systems have average word error rates of 10·2% and 5·9% for single- and six-component Gaussian models, respectively.

For MAP estimation purposes, for each system, the prior parameters are estimated using the corresponding baseline SI system parameters in the same manner introduced in Section 2, using a value of $\tau_{ik}=10$ in all the experiments. The thresholds, T_H and T_L, used in RMP evaluations are set to different values, depending on the number of adaptation sentences and the type of system in use, which have a direct influence on the mixture occupation values obtained during MAP estimation. Usually, the thresholds were set so that the RMP uses only about 5–15% of the distributions as sources and about 75–90% of them as targets. Further details of the implementation are given in Ahadi and Woodland (1995).

The results shown in Fig. 3 indicate that a worthwhile improvement can be obtained with the application of RMP to context-dependent models. It is shown that in the case of only one adaptation utterance, where no improvement can be expected from MAP estimation, RMP results in an 8% reduction in word error rate. The RMP performance always remains better than MAP estimation as more adaptation sentences become available, although the size of the improvement relative to that of MAP gradually reduces. This could be expected since with more adaptation data, the number of untrained or poorly trained parameters is gradually reduced.

With a very large number of adaptation sentences it can be seen, by inspecting Equations (7) and (11), that s_ζ^2 would be zero which leads to a performance equal to MAP estimation. Hence, RMP displays two desirable characteristics: the asymptotic performance of the MAP estimation with a large number of adaptation sentences and very fast speaker adaptation performance. It also continuously outperforms MAP estimation with different numbers of adaptation utterances.

7.2. Reducing adaptation computation

The RMP technique, as discussed above, can be computationally expensive. The main drawback is the need to calculate the regression parameters for every new speaker because of the use of the speaker's mixture occupation counts for deciding on the allocation of the source and target distributions. Due to the size of the system used, the thresholds set and the number of SD models used, this can be costly. In fact, for any new speaker, given the number of adaptation sentences to be used, in a supervised adaptation scheme, the calculation involves the use of the same basic SD and MAP model sets for SD speakers. The only difference would be the difference in the mixture occupation counts of Gaussian components for different speakers.

A somewhat simpler approach would be to use a general set of mixture occupation counts for all the speakers. This could be found, say, by averaging the mixture occupation counts from several speakers. Then, for a given number of adaptation utterances, a global regression parameter calculation can be carried out using all the SD models available, plus all the corresponding MAP estimated models and the global mixture occupation counts. In this case, the models of the speaker under test are not included

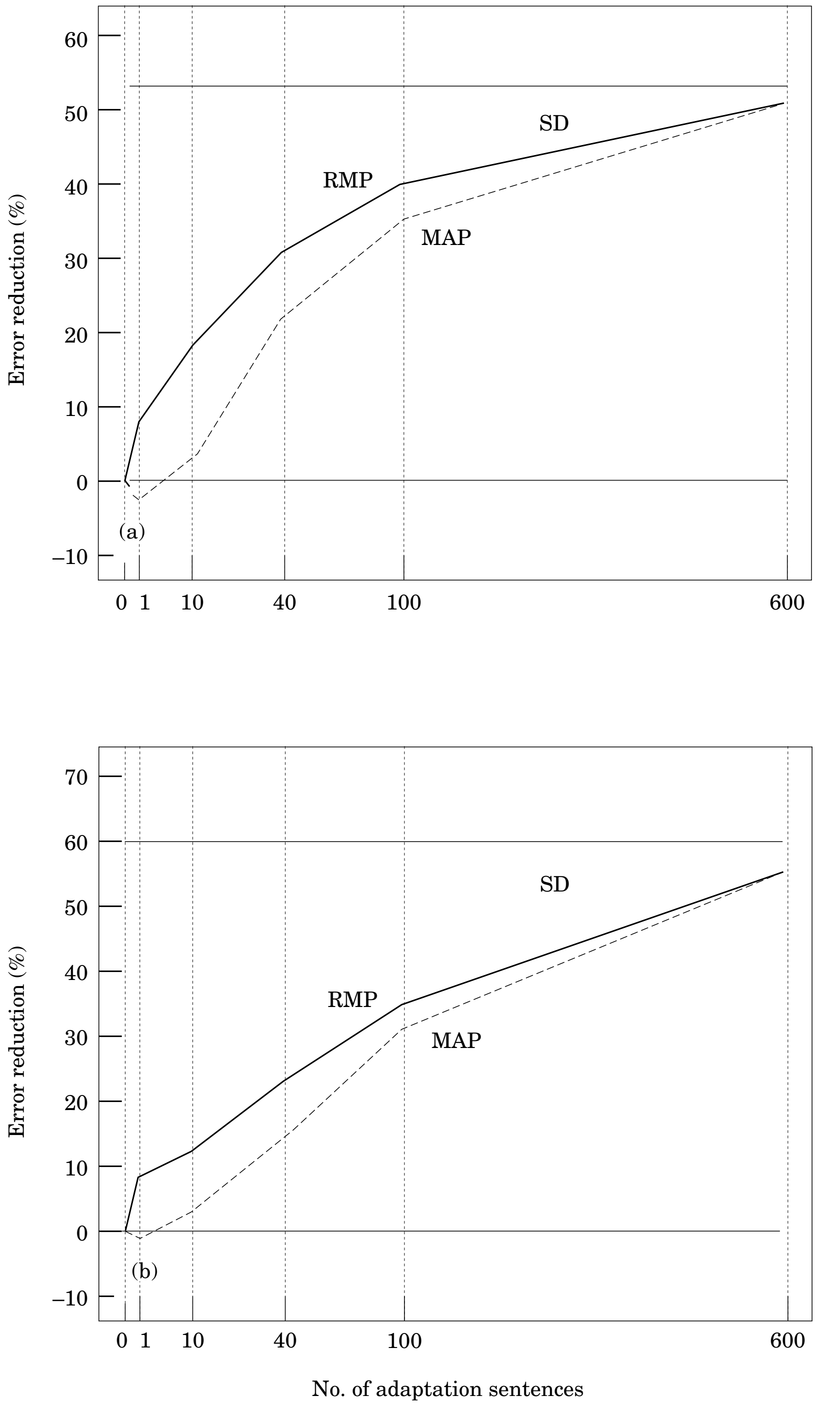


Figure 3. A comparison between percentage error rate reductions, relative to SI, obtained by the application of RMP and MAP methods for speaker adaptation of (a) single Gaussian and (b) six-component mixture Gaussian triphone systems.

Table II. Comparison of the word error rates of RMP and RMP-AMO for single Gaussian and six-component mixture triphone systems

	Single Gaussian triphones		Six-component triphones	
Adaptation sentences	RMP	RMP-AMO	RMP	RMP-AMO
0	10·21	10·21	5·92	5·92
1	9·42	9·42	5·44	5·60
10	8·37	8·31	5·19	5·31
40	7·06	7·13	4·56	4·62
100	6·15	6·04	3·86	3·98
600	5·02	5·02	2·34	2·34

in the set of SD models. These globally calculated parameters can then be easily used to adapt the models of any new speaker. Note that, once the regression parameters are available, the updating of model parameters for a new speaker is a simple and fast task which does not involve much computation.

Table II includes a comparison between the results obtained on the single-component

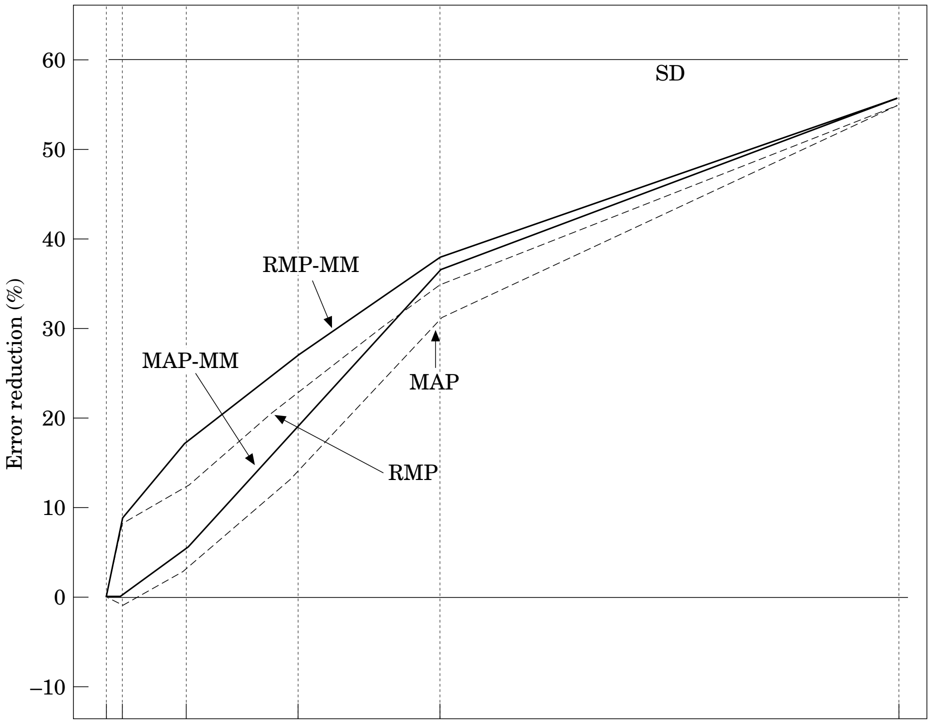


Figure 4. A comparison of the improvements obtained over SI system performance using the method of moments for estimating the prior parameters of MAP estimation and corresponding RMP adaptation (marked MAP_MM and RMP_MM) with previous MAP and RMP results on a six-component mixture Gaussian triphone system.

and six-component mixture triphone systems, using the standard RMP method and this method, identified as *RMP with Averaged Mixture Occupations* (RMP-AMO). These results emphasize the fact that the use of averaged mixture occupations does not involve a significant degradation in the performance of RMP adaptation. A reduction of more than 90% in the adaptation computation time results from the application of RMP-AMO in comparison to RMP, if the adaptation is to be carried out on 10 new speakers.

7.3. RMP adaptation with moment-based MAP

An empirical Bayes method of prior estimation using the method of moments (see Section 2.2) can be used as an alternative to the *ad hoc* method and might lead to some improvements in the overall performance of the adapted system. Since in RMP, the MAP estimated system is used as the basic system, it is believed that having better priors can lead to better source models and hence better estimates of other model parameters calculated by RMP. This provided the motivation for applying the RMP method to the MAP estimated models with moment estimated priors. The results of these experiments, together with the previous results of MAP and RMP for the six-component mixture Gaussian triphone system, are shown in Fig. 4.

The results of MAP estimation, as expected, are improved in comparison with the *ad hoc* approach and almost the same effect can be observed on the RMP results. Note that the improvement in the MAP result for one adaptation sentence is insignificant and, hence, the RMP result is also similar. However, for a larger number of adaptation sentences, better MAP models have helped to improve RMP results. It is also noticed that for a very large number of sentences, the difference between MAP and RMP results is reduced.

Another point in the application of the moment's method is that due to the availability of several SD model sets trained to be used in the RMP method, the calculation of the prior parameters needed in this case is straightforward.

8. Conclusions

A new method for adaptation of CDHMMs, called RMP, has been introduced. In this method, a predictive approach is used to estimate the parameters of unadapted or poorly adapted models resulting from initial MAP-based adaptation. The model relationships used for this purpose were found by application of linear regression to the parameters of several speaker-specific systems.

A number of experiments using the RMP method for speaker adaptation were presented. The results of these experiments demonstrate the ability of RMP to further adapt a baseline model set to the speech of a new speaker with especially notable improvements coming from a very small number of adaptation sentences. RMP improved the results of MAP estimation over a range of adaptation sentences and, for a large number of sentences, RMP converges towards the MAP and SD performance.

It has been shown that RMP can be made to be more computationally efficient by using averaged mixture occupations in the regression calculations. This eliminates the need for running the regression calculations each time adaptation is to be carried out and substantially reduces the amount of computation. For mixture Gaussian models, a small part of the reduction in error rate is lost, while for single Gaussians almost the same performance is obtained.

The moment's method of prior parameter estimation has also been used to improve the initial MAP estimation process and has also led to an improvement in RMP performance. However, this method is found to contribute more to the MAP estimation process than to RMP.

S. M. Ahadi was supported by a studentship from the Iranian Ministry of Culture and Higher Education and the CVCP ORS award. The authors would like to thank two anonymous reviewers for their helpful comments.

References

Ahadi, S. M. (1996). *Bayesian and Predictive Techniques for Speaker Adaptation.* Ph.D. Thesis, Cambridge University Engineering Department.

Ahadi, S. M. & Woodland, P. C. (1995). Rapid speaker adaptation using model prediction. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1. Detroit, pp. 684–687.

Chatterjee, S. & Price, B. (1991). *Regression Analysis by Example*, 2nd edn. John Wiley and Sons, New York.

Cox, S. J. (1992). Speaker adaptation in speech recognition using linear regression techniques. *Electronic Letters* **28**(22), 2093–2094.

Cox, S. J. (1993). Speaker adaptation using a predictive model. *Proceedings of Eurospeech*, Vol. 3, Berlin, pp. 2283–2286.

Cox, S. J. (1995). Predictive speaker adaptation in speech recognition. *Computer Speech and Language* **9**, 1–17.

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

Dunn, O. J. & Clark, V. A. (1987). *Applied Statistics: Analysis of Variance and Regression.* John Wiley and Sons, New York.

Furui, S. (1980). A training procedure for isolated word recognition systems. *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-28**(2), 129––136.

Gauvain, J.-L. & Lee, C.-H. (1992). Bayesian learning for hidden Markov model with Gaussian mixture observation densities. *Speech Communication* **11**, 205–213.

Gauvain, J.-L. & Lee, C. H. (1994). Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing* **SAP-2**(2), 291–298.

Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F. & Rosenfeld, R. (1993). The SPHINX-II speech recognition system: an overview. *Computer Speech and Language* **7**(2), 137–148.

Huo, Q. & Chan, C. (1992). Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. Technical Report, Department of Computer Science, University of Hong Kong.

Lee, K.-F. (1989). *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, Boston.

Lee, C.-H. & Gauvain, J.-L. (1992). A study on speaker adaptation for continuous speech recognition. *Proceedings of the ARPA Continuous Speech Recognition Workshop*. Stanford, pp. 59–64.

Lee, C.-H., Lin, C.-H. & Juang, B.-H. (1991). A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing* **SP-39**(4), 806–814.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language* **9**, 171–185.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society* **B-9**, 218–244.

Stern, R. M. & Lasry, M. J. (1987). Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-35**(6), 751–762.

Woodland, P. C., Odell, J. J., Valtchev, V. & Young, S. J. (1994). Large vocabulary continuous speech recognition using HTK. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. II. Adelaide, pp. 125–128.

(*Received 24 August 1996 and accepted for publication 2 July 1997*)