Computer Speech and Language (1995) **9,** 171–185

Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models

C. J. Leggetter and P. C. Woodland

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

Abstract

A method of speaker adaptation for continuous density hidden Markov models (HMMs) is presented. An initial speaker-independent system is adapted to improve the modelling of a new speaker by updating the HMM parameters. Statistics are gathered from the available adaptation data and used to calculate a linear regression-based transformation for the mean vectors. The transformation matrices are calculated to maximize the likelihood of the adaptation data and can be implemented using the forward–backward algorithm. By tying the transformations among a number of distributions, adaptation can be performed for distributions which are not represented in the training data. An important feature of the method is that arbitrary adaptation data can be used—no special enrolment sentences are needed.

Experiments have been performed on the ARPA RM1 database using an HMM system with cross-word triphones and mixture Gaussian output distributions. Results show that adaptation can be performed using as little as 11 s of adaptation data, and that as more data is used the adaptation performance improves. For example, using 40 adaptation utterances, a 37% reduction in error from the speaker-independent system was achieved with supervised adaptation and a 32% reduction in unsupervised mode.

1. Introduction

In spite of progress in the development of speaker-independent (SI) systems, error rates are still typically two to three times higher than equivalent speaker-dependent (SD) systems (Lee, Lin & Juang, 1991). Since a large amount of speaker-specific data is required for training SD systems they are not suitable for many applications. SI systems model speech from some speakers poorly so it is desirable to use a small amount of the new speaker's speech (adaptation data) to "tune" the SI models to the new speaker, such methods are termed speaker adaptation techniques. The adaptation is said to be supervised if the true transcription of the adaptation data is known and otherwise unsupervised.

Adaptation techniques fall into two main categories—speaker normalization in which

0885–2308/95/020171 + 15 $08.00/0

the input speech is normalized to match the speaker that the system is trained to model, and model adaptation techniques in which the parameters of the model set are adjusted to improve the modelling of the new speaker. An important issue with both approaches is effective operation with a limited amount of adaptation data. For a system with a large number of models and a small amount of adaptation data, some models will not be observed in the data. Some model adaptation techniques [e.g. MAP estimation (Gauvain & Lee, 1994)] only update the parameters of models which are observed in the adaptation data, thus fairly large amounts of adaptation data are generally required.

Here, we propose a model adaptation technique which uses a set of regression-based transforms to tune the hidden Markov model (HMM) mean parameters to the new speaker. Each of the transformations is applied to a number of HMM mean parameters and estimated from the corresponding data. Using this sharing of transformations and data, the method can produce improvements with small amounts of adaptation data.

The method has links with the work on spectral shift transformations (Jaschul, 1982) which attempts to map data from a new speaker onto that from a reference speaker. In that work, uniform and non-uniform frequency shifts and a purely additive transform were considered in a simple spectral phone template. It was found that the frequency shifts were most effective, although they required more data to be estimated. Jaschul (1982) also combined the frequency shift and the additive effect in a single transform. In an attempt to reduce the amount of data needed to estimate the transform a tridiagonal frequency shift matrix was considered, which although successful in halving the amount of data required for estimation, gave poorer results. It was later shown (Hewett, 1989) that the full transform used by Jaschul (1982) gave better performance than a canonical correlation approach (Choukri, Chollet & Grenier, 1986) which projects parameters of both the reference speaker and the new speaker to a new acoustic space where they are similar. These spectral transformation approaches have also shown success in a discrete HMM framework (Class *et al*., 1990), and with individual sound class models or vowel spectra using Gaussian densities (Cox & Bridle, 1989).

Although some approaches consider transformations on a global basis, i.e. transforming all speech vectors/parameters by the same transform (Class *et al*., 1990), it was shown by Jaschul (1982) that gains can be made by using phone-dependent transforms. However, this increases the amount of adaptation data required from the speaker so that there is sufficient data to estimate each transform. A piecewise-linear approach based on phone classes has also been used with discrete HMMs (Bellegarda *et al*., 1992).

Here, we extend the ideas of Jaschul (1982) and Hewett (1989) to adapting the parameters of continuous density HMMs. The parameters of the HMM system are adapted using transforms which are estimated in a maximum likelihood framework. The least squares regression calculation used by Hewett (1989) has been replaced by maximum likelihood estimation taking into account different state distributions. If only a small amount of adaptation data is presented a global transform is used for all models in the system, and if more data is available the number of transforms is increased. This ensures that all model states can be adapted even if no model-specific data is available. We refer to this method as maximum likelihood linear regression (MLLR) adaptation. The statistics used to estimate the transform matrices are generated using a forward–backward alignment of the adaptation data. Hence, the method has clear links with standard Baum–Welch HMM training. A related approach has been developed contemporaneously by Digalakis (Digalakis, Rtischev and Neumeyer, 1995).

Section 2 describes the MLLR approach. The adaptation formulae are derived in Section 3, with special cases detailed in Section 4. Section 5 briefly explains how sets of parameters that use the same transform are tied. The method is evaluated using the experimental setup described in Section 6, and the results are presented in Section 7.

2. Adaptation approach

The MLLR approach to speaker adaptation requires an initial speaker independent continuous density HMM system. MLLR takes some adaptation data from a new speaker and updates the model mean parameters to maximize the likelihood of the adaptation data. The other HMM parameters are not adapted since the main differences between speakers are assumed to be characterized by the means.

Consider the case of a continuous density HMM system with Gaussian output distributions. A particular distribution, s, is characterized by a mean vector, $\boldsymbol{\mu}_s$, and a covariance matrix C_s. Given a parameterized speech frame vector $\boldsymbol{o}$, the probability density of that vector being generated by distribution s is $b_s(\boldsymbol{o})$

$$b_s(\boldsymbol{o}) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(\boldsymbol{o}-\boldsymbol{\mu}_s)'C_s^{-1}(\boldsymbol{o}-\boldsymbol{\mu}_s)}$$

where n is the dimension of the observation vector.

The adaptation of the mean vector is achieved by applying a transformation matrix W_s to the extended mean vector $\boldsymbol{\xi}_s$ to obtain an adapted mean vector $\hat{\boldsymbol{\mu}}_s$

$$\hat{\boldsymbol{\mu}}_s = W_s \boldsymbol{\xi}_s$$

where W_s is an $n \times (n+1)$ matrix which maximizes the likelihood of the adaptation data, and $\boldsymbol{\xi}_s$ is defined as

$$\boldsymbol{\xi}_s = [\omega, \mu_1, \ldots, \mu_n]'$$

where ω is the offset term for the regression ($\omega = 1$ to include an offset in the regression, $\omega = 0$ to ignore offsets).

For distribution s, the probability density function for the adapted system becomes

$$b_s(\boldsymbol{o}) = \frac{1}{(2\pi)^{n/2}|C_s|^{1/2}} e^{-1/2(\boldsymbol{o}-W_s\boldsymbol{\xi}_s)'C_s^{-1}(\boldsymbol{o}-W_s\boldsymbol{\xi}_s)}. \quad (1)$$

Having a separate transform for each Gaussian distribution is equivalent to a complete re-estimation of the means. This leaves the problem of adapting the parameters of unseen distributions unsolved. A more general approach is adopted in which the same transformation matrix is used for several distributions. The transformation is estimated using data from all the associated (tied) distributions, so if some of the distributions are not observed in the adaptation data, a transformation may still be applied. The degree of transformation tying is determined by the amount of adaptation data available. For the case of small amounts of adaptation data a global transformation may be used.

3. Estimation of MLLR regression matrices

MLLR estimates the regression matrices W_s to maximize the likelihood of the adapted models generating the adaptation data. The derivation of the MLLR estimates is given below, making the assumption that each HMM state has a single Gaussian output distribution. The full covariance case has no closed form for the estimation formulae in the tied matrix case, hence only the case of diagonal covariance distributions is considered.

3.1. Definition of auxiliary function

Assume the adaptation data, O, is a series of T observations

$$O = \boldsymbol{o}_1 \ldots \boldsymbol{o}_T.$$

Denote the current set of model parameters by λ and a re-estimated set of model parameters as $\bar{\lambda}$. If all possible state sequences of length T are denoted by the set Θ, the total likelihood of the model set generating the observation sequence is

$$\mathscr{F}(O|\lambda) = \sum_{\theta \in \Theta} \mathscr{F}(O, \theta|\lambda)$$

where $\mathscr{F}(O, \theta|\lambda)$ is the likelihood of generating O using the state sequence θ given the model parameters.

The quantity $\mathscr{F}(O|\lambda)$ is the objective function to be maximized during adaptation. It is convenient to define an auxiliary function $Q(\lambda, \bar{\lambda})$:

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} \mathscr{F}(O, \theta|\lambda) \log(\mathscr{F}(O, \theta|\bar{\lambda})). \tag{2}$$

Model parameters which maximize the auxiliary function also increase the value of the objective function (unless it is at a maximum). Therefore successively forming a new auxiliary function with improved parameters iteratively maximizes the objective function. Baum first proved the convergence of the algorithm (Baum, 1972) which was later extended to mixture distributions and vector observations (Liporace, 1982; Juang, 1985).

3.2. Maximization of auxiliary function

Since only the transformations W_s are re-estimated, only the output distributions b_s are affected so the auxiliary function (2) can be written as

$$Q(\lambda, \bar{\lambda}) = constant + \sum_{\theta \in \Theta} \sum_{t=1}^{T} \mathscr{F}(O, \theta|\lambda) \log b_{\theta_t}(\boldsymbol{o}_t). \tag{3}$$

Defining S as the set of all state distributions in the system, and $\gamma_s(t)$ as the

a posteriori probability of occupying state s at time t given that the observation sequence O is generated

$$\gamma_s(t) = \frac{1}{\mathscr{F}(O|\lambda)} \sum_{\theta \nu \Theta} \mathscr{F}(O, \theta_t = s|\lambda).$$

Equation (3) can then be written as

$$Q(\lambda, \bar{\lambda}) = constant + \mathscr{F}(O|\lambda) \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_j(t) \log b_j(\boldsymbol{o}_t). \tag{4}$$

Expanding $\log b_j(\boldsymbol{o}_t)$ the auxiliary function is

$$Q(\lambda, \bar{\lambda}) = constant - \frac{1}{2} \mathscr{F}(O|\lambda) \times \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_j(t)[n \log(2\pi) + \log|C_j| + h(\boldsymbol{o}_t, j)]$$

where

$$h(\boldsymbol{o}_t, j) = (\boldsymbol{o}_t - \overline{W}_j \boldsymbol{\xi}_j)' C_j^{-1} (\boldsymbol{o}_t - \overline{W}_j \boldsymbol{\xi}_j).$$

The differential of $Q(\lambda, \bar{\lambda})$ with respect to $\bar{W}_s$ is:

$$\frac{d}{d\overline{W}_s} Q(\lambda, \bar{\lambda}) = -\frac{1}{2} \mathscr{F}(O|\lambda) \frac{d}{d\overline{W}_s} \times \sum_{j=1}^{S} \sum_{t=1}^{T} \gamma_j(t)[n \log(2\pi) + \log|C_j| + h(\boldsymbol{o}_t, j)]$$

then completing the differentiation, and equating to zero to find the maximum of $Q(\lambda, \bar{\lambda})$

$$\frac{d}{d\overline{W}_s} Q(\lambda, \bar{\lambda}) = \mathscr{F}(O|\lambda) \sum_{t=1}^{T} \gamma_s(t) C_s^{-1} [\boldsymbol{o}_t - \overline{W}_s \boldsymbol{\xi}_s] \boldsymbol{\xi}_s' = 0$$

hence

$$\sum_{t=1}^{T} \gamma_s(t) C_s^{-1} \boldsymbol{o}_t \boldsymbol{\xi}_s' = \sum_{t=1}^{T} \gamma_s(t) C_s^{-1} \overline{W}_s \boldsymbol{\xi}_s \boldsymbol{\xi}_s'. \tag{5}$$

Equation (5) gives the general form for computing $\overline{W}_s$.

3.3. Re-estimation formula for tied regression matrices

When the regression matrices are tied across a number of distributions (e.g. a global regression matrix) the summations must be performed over all tied distributions. If W_s is shared by R states $\{s_1, s_2 \ldots s_R\}$ Equation (5) becomes:

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)C_{s_r}^{-1}\boldsymbol{o}_t\boldsymbol{\xi}'_{s_r}=\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)C_{s_r}^{-1}\overline{W}_s\boldsymbol{\xi}_{s_r}\boldsymbol{\xi}'_{s_r}. \tag{6}$$

To derive a re-estimation formula for the tied case, Equation (6) is first rewritten as

$$\sum_{t=1}^{T}\sum_{r=1}^{R}\gamma_{s_r}(t)C_{s_r}^{-1}\boldsymbol{o}_t\boldsymbol{\xi}'_{s_r}=\sum_{r=1}^{R}V^{(r)}\overline{W}_sD^{(r)} \tag{7}$$

where $V^{(r)}$ is the state distribution inverse covariance matrix scaled by the state occupation probability,

$$V^{(r)}=\sum_{t=1}^{T}\gamma_{s_r}(t)C_{s_r}^{-1}$$

and $D^{(r)}$ is the outer product of the extended mean vectors

$$D^{(r)}=\boldsymbol{\xi}_{s_r}\boldsymbol{\xi}'_{s_r}.$$

If the right-hand side of Equation (7) is denoted by the $n\times(n+1)$ matrix Y with elements y_{ij}, the individual matrix elements of $V^{(r)}$, W_s and $D^{(r)}$ are denoted by $v_{ij}^{(r)}$, w_{ij} and $d_{ij}^{(r)}$, respectively, then

$$y_{ij}=\sum_{p=1}^{n}\sum_{q=1}^{n+1}w_{pq}\left[\sum_{r=1}^{R}v_{ip}^{(r)}d_{qj}^{(r)}\right].$$

If all covariances are diagonal, and since D is symmetric, then

$$\sum_{r=1}^{R}v_{ip}^{(r)}d_{qj}^{(r)}=\begin{cases}\Sigma_{r=1}^{R}v_{ii}^{(r)}d_{jq}^{(r)} & \text{when } i=p\\ 0 & \text{when } i \text{ r } p\end{cases}$$

so

$$y_{ij}=\sum_{q=1}^{n+1}w_{iq}g_{jq}^{(i)},$$

where $g_{jk}^{(i)}$ are the elements of the $(n+1)\times(n+1)$ matrix $G^{(i)}$, given by

$$g_{jq}^{(i)} = \sum_{r=1}^{R} v_{ii}^{(r)} d_{jq}^{(r)}.$$

If the left hand side of Equation (7) is an $n\times(n+1)$ matrix denoted by Z with elements z_{ij}, then $Z = Y$ and

$$z_{ij} = y_{ij} = \sum_{q=1}^{n+1} w_{iq} g_{jq}^{(i)}.$$

It should be noted that z_{ij} and $g_{jq}^{(i)}$ are not dependent on $\overline{W}_s$ and both can be computed from the observation vectors and the model parameters. Hence $\overline{W}_s$ can be computed from the system of simultaneous equations

$$\boldsymbol{w}_i' = G^{(i)-1}\boldsymbol{z}_i'$$

where $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ are the i^{th} rows of $\underline{W}_s$ and Z, respectively. These equations can be solved using Gaussian elimination or LU decomposition methods and $\overline{W}_s$ calculated on a row-by-row basis.

The extension of the method to HMMs with mixture Gaussian distributions is straightforward and achieved by substituting mixture component occupation probabilities for state occupation probabilities (Leggetter & Woodland, 1994).

4. Special cases of MLLR

4.1. Least squares regression

Least squares regression computation as used by Hewett (1989) can be shown to be a special case of MLLR. If all the covariances of the distributions tied to the same transformation are the same, Equation (6) becomes

$$\sum_{t=1}^{T}\sum_{r=1}^{R} \gamma_{s_r}(t)\boldsymbol{o}_t\boldsymbol{\xi}_{s_r}' = \sum_{t=1}^{T}\sum_{r=1}^{R} \gamma_{s_r}(t)\overline{W}_s\boldsymbol{\xi}_{s_r}\boldsymbol{\xi}_{s_r}'. \tag{8}$$

If each speech frame is assigned to exactly one distribution (e.g. by Viterbi alignment) so that

$$\gamma_{s_r}(t) = \begin{cases} 1 & \text{if } \boldsymbol{o}_t \text{ is assigned to state distribution } s_r \\ 0 & \text{otherwise} \end{cases}$$

then Equation (8) becomes

$$\sum_{t=1}^{T} \boldsymbol{o}_t \boldsymbol{\xi}'_{\theta_t} \delta_{s\theta_t} = \bar{W}_s \sum_{t=1}^{T} \boldsymbol{\xi}_{\theta_t} \boldsymbol{\xi}'_{\theta_t} \delta_{s\theta_t} \tag{9}$$

where

$$\delta_{s\theta_t} = \begin{cases} 1 & \text{if } \theta_t \text{v} \{s_1 \ldots s_R\} \\ 0 & \text{otherwise.} \end{cases}$$

Defining the matrices X and Y as

$$X = [\boldsymbol{\xi}_{\theta_1} \boldsymbol{\xi}_{\theta_2} \ldots \boldsymbol{\xi}_{\theta_T}]$$
$$Y = [\boldsymbol{o}_1 \delta_{s\theta_1} \boldsymbol{o}_2 \delta_{s\theta_2} \ldots \boldsymbol{o}_T \delta_{s\theta_T}].$$

Substituting X and Y in Equation (9) and rearranging, the estimate of the regression matrix is the standard least squares estimate (Kendall, 1971)

$$\overline{W}_s = YX'(XX')^{-1}.$$

4.2. Single variable linear regression

If all the features in the mean vector are independent, the modification of each mean component can be calculated by simple single variable linear regression. This significantly reduces the number of regression parameters which need to be estimated per regression matrix.

Rewriting the transformation matrix $\overline{W}_s$ for this case as a vector $\hat{\boldsymbol{w}}_s$

$$\overline{W}_s = \begin{pmatrix} w_{1,1} & w_{1,2} & 0 & \ldots & 0 \\ w_{2,1} & 0 & w_{2,3} & \ldots & 0 \\ \vdots & & & & \vdots \\ w_{n,1} & \ldots & \ldots & 0 & w_{n,n+1} \end{pmatrix}, \hat{\mathbf{w}}_s = \begin{pmatrix} w_{1,1} \\ \vdots \\ w_{n,1} \\ w_{1,2} \\ \vdots \\ w_{n,n+1} \end{pmatrix},$$

and defining an $n \times 2n$ matrix D_s made up of elements of the extended mean vector $\boldsymbol{\xi}$

$$D_s = \begin{pmatrix} \omega & 0 & \ldots & \ldots & 0 & \mu_1 & 0 & \ldots & & 0 \\ 0 & \omega & 0 & \ldots & \ldots & 0 & \mu_2 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & \vdots \\ 0 & \ldots & 0 & \omega & 0 & \ldots & \ldots & 0 & \mu_{n-1} & 0 \\ 0 & \ldots & \ldots & 0 & \omega & 0 & \ldots & \ldots & 0 & \mu_n \end{pmatrix}$$

then the maximization of the objective function leads to the following formula for the regression parameters in the tied regression matrix cases:

$$\hat{\boldsymbol{w}}_s = \left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_s(t)D'_{s_r}C_{s_r}^{-1}D_{s_r}\right]^{-1}\left[\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_{s_r}(t)D'_{s_r}C_{s_r}^{-1}\boldsymbol{o}_t\right].$$

Thus only one matrix inversion is required to calculate W and if the offset term is zero ($\omega = 0$) all matrices can be reduced to diagonal matrices making computation trivial.

5. Defining regression classes

When regression matrices are tied across mixture components, each matrix is associated with many mixture components. This is achieved by defining a set of regression classes where each class contains all the mixture components associated with the same regression matrix.

For the tied approach to be effective it is desirable to put all the mixture components which will use similar transforms into the same class. However, since we have no *a priori* knowledge of the transforms some other criteria must be used. Thus we make an assumption that mixture components with similar parameter values will change in a similar manner. By assigning mixture components, representing similar acoustic phenomena to the same regression class, they will be transformed using the same regression matrix.

Two approaches for defining regression classes were considered: one based on broad phonetic classes and the other based on clustering of mixture components. In the first case all mixture components in any model representing the same broad phonetic class (e.g. fricatives, nasals, etc.) were placed in the same regression class. In the clustering approach the mixture components were compared using a likelihood measure and similar components placed in the same regression class. Preliminary results showed little difference between the two approaches when a small number of classes is used, but the data driven approach was found to be more appropriate for defining large numbers of classes.

6. Experimental set-up

The ARPA Resource Management RM1 database was used to evaluate MLLR speaker adaptation. The speech was coded into 25 ms frames, with a frame advance of 10 ms. Each frame was represented by a 39 component vector consisting of 12 MFCCs plus log energy, and their first and second time derivatives.

A set of SI models was trained on the SI portion of the database (3990 utterances), using standard Baum–Welch maximum likelihood estimation. The models were state clustered cross-word triphones, containing a total of 1778 states. Each state had six mixture components for the output distribution. All training and data preparation was performed using the HTK HMM toolkit (Young, Woodland & Byrne, 1993).

The 48 phone CMU set was used for labels and a single pronunciation for each word was derived from a dictionary (Lee, 1988). This set of SI models gives a 2·5%

word error rate on the RM Feb'91 SI test set (Young, Odell & Woodland, 1994) and was used as the initial system for the adaptation experiments.

The SD portion of the database was used for all adaptation experiments. This consists of 600 training and 100 test utterances from each of 12 different speakers (seven male and five female). Portions of the training data were used as adaptation data.

Adaptation was implemented by using a forward–backward alignment of the adaptation data to assign frames to regression matrices as previously described. All recognition tests were performed using a dynamic network decoder (Odell, Valtchev, Woodland & Young, 1994) with the standard RM word-pair grammar (perplexity 60).

To enable a comparison with an equivalent speaker dependent system the SI models were retrained for each speaker using all 600 training files using Baum–Welch estimation, and tested on the speaker dependent test set. The SI system gave 4·3% word error and the SD system 1·8% word error (averaged over all 12 speakers).

For each adaptation experiment the number of regression matrices was pre-determined. The mixture components were assigned to the regression classes using a clustering procedure. Each mixture component was initially assigned to an individual class. The two closest classes were combined to create a single class, and the process repeated until the desired number of classes was obtained. The distance measure used was based on the change in likelihood of the class data being generated by the classes when the classes are combined as opposed to being kept separate. The smaller the change the closer the distributions are, so these components are placed in the same class.

7. Evaluation

A number of experiments have been performed to investigate the merits of the approach. These include a comparison between using diagonal and full regression matrices; an investigation into the amount of adaptation data required; unsupervised implementation of MLLR; and a comparison using the least squares regression criteria.

All adaptation was performed in a static supervised manner using labelled adaptation data unless otherwise indicated. Only one iteration of adaptation was performed in all cases since preliminary tests indicated that further iterations had very little effect (i.e. the frame-state alignment for the adapted models is very similar to the SI frame-state alignment). In all cases the silence models were not adapted and the silence data was ignored during adaptation. For all experiments the offset term for the regression calculation was set to one ($\omega = 1$).

7.1. Comparison of full and diagonal regression matrices

Using 40 adaptation utterances, MLLR adaptation was performed while varying the number of regression classes using either full or diagonal regression matrices.

Both approaches give an improvement over the initial model set, but the effect of the diagonal matrices is limited (Fig. 1). The full matrices give a substantial improvement when using just one or two classes, but as the number of classes is increased the amount of data allocated to each class becomes small and the matrices are poorly estimated. Thus the performance falls away rapidly. With diagonal matrices, as more classes are used the performance gradually increases; however, this effect is very small and using 300 diagonal matrices is only 0·2% absolute better than using 40 diagonal regression

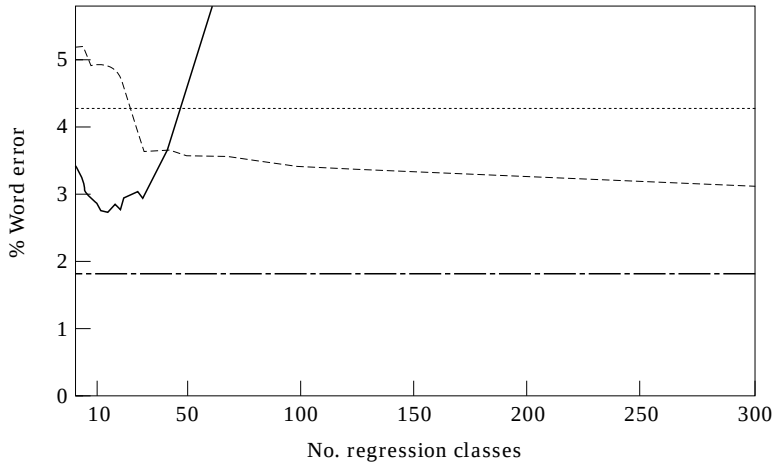

Figure 1. Full matrix vs. diagonal matrix maximum likelihood linear regression using 40 adaptation utterances. (·····), Speaker independent; (—·—·—·), speaker dependent; (——), full adapted; (– – –), diagonal adapted.

matrices. A diagonal matrix has $2n$ parameters while a full matrix has $n(n+1)$ parameters, thus the amount of data needed to estimate a diagonal regression matrix is much smaller than that of a full matrix. This indicates that more classes can be used with the same amount of data. The results show that increasing the number of classes does improve performance, but the performance never reaches that of the full regression matrix. Although an examination of full regression matrices showed the main diagonal to be quite dominant, it is clear that the off-diagonal terms relating the interdependencies between components is important.

The single variable regression used for the diagonal matrix is clearly not powerful enough to capture many different variations within a class. When using 300 classes of diagonal matrices the number of adaptation parameters is equivalent to that using 15 full matrices, yet the performance is significantly poorer (3·2 vs. 2·7%). Using a large number of classes can lead to problems of data sufficiency for robust estimates of each class, and therefore adaptation using a small number of full matrices is superior to using many diagonal matrices. There are also advantages for unsupervised adaptation since using a small number of classes is more likely to generate more robust estimates for the adaptation parameters in the presence of some errors in the labelling of the adaptation data.

7.2. Small amounts of adaptation data

To assess how the amount of adaptation data affects performance, a global regression matrix was estimated while varying the number of adaptation utterances.

Using a full regression matrix gives an improvement after just three adaptation utterances (equivalent to an average of 11 s of speech per speaker), and as more adaptation data is presented the performance gradually improves (Fig. 2). When a sufficient amount of data has been used to estimate the regression matrix further addition of data has little effect. By increasing the number of classes as the amount of

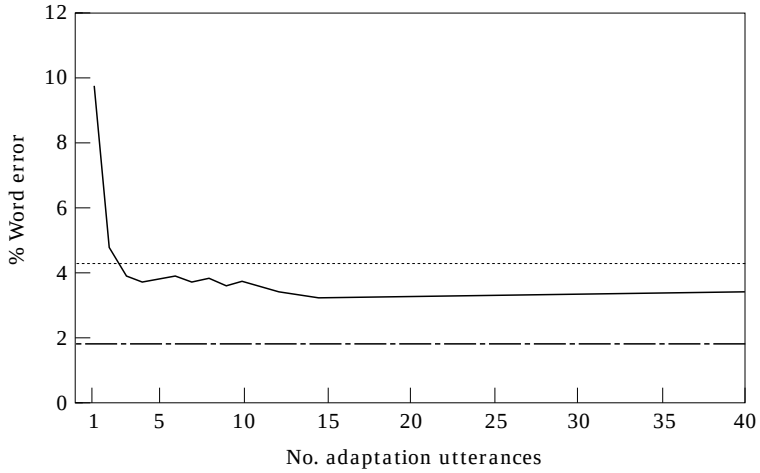

Figure 2. Full matrix maximum likelihood linear regression using global regression class. (·····), Speaker independent; (—·—·—·—), speaker dependent; (——), speaker adapted.

data increases while ensuring sufficient adaptation data per regression class, improved performance can be achieved. The method performs poorly when there are too many classes and not enough data. In these cases the assignment of data to each class is insufficient for a robust estimate for the regression, and in extreme cases the accumulated matrices to invert are very close to being singular (due to linear dependence), resulting in computational errors.

7.3. Comparison of supervised and unsupervised adaptation

In the previous experiments the adaptation has been supervised, i.e. the true transcription of the adaptation data was known. Unsupervised adaptation may be implemented by first generating the adaptation data transcription by an initial recognition pass. This information is then used in the forward–backward alignment of data in adaptation. In experiments to compare supervised and unsupervised adaptation different quantities of adaptation data were used and an appropriate number of regression classes selected by experiment. The best supervised and unsupervised adaptation results for the given data are shown in Fig. 3.

Supervised adaptation performs marginally better in all cases, but the difference is small: for 40 utterances the difference is 0·2% absolute; with 100 utterances 0·2%; and 600 utterances 0·1%. This small difference is partly due to the accuracy of the SI models which gives a reasonable alignment in most cases, and also due to the broad tying of classes which reduces the effect of misaligned frames due to errors in the adaptation labelling.

As the number of adaptation files increases the performance also improves and gradually tends towards the performance of an SD system. In the limit, with sufficient data and one class per mixture component, the performance should be equivalent to that achieved by Baum–Welch estimation since the new means are estimated by the same optimization criteria.

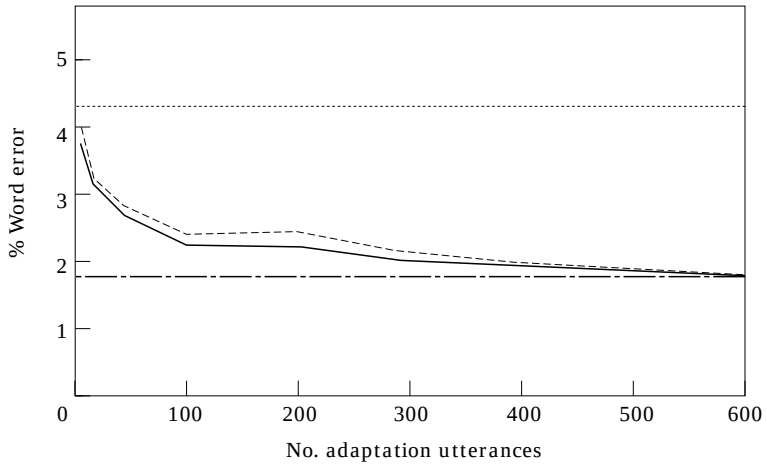

Figure 3. Supervised vs. unsupervised adaptation using maximum likelihood linear regression. (·····), Speaker independent; (—-—-—-), speaker dependent; (——), supervised adapted; (– – –), unsupervised adapted.

TABLE I. Comparison of adaptation using least squares estimation and maximum likelihood linear regression (MLLR) estimation using Viterbi and forward–backward alignment of the data % word error

No. adaptation utterances	No. of classes	Least squares adaptation	MLLR adaptation	
			Viterbi	Forward–backward
5	1	4·0	3·8	3·8
10	1	3·8	3·8	3·7
15	1	3·4	3·3	3·2
40	1	3·5	3·4	3·4
40	5	3·5	3·2	3·0
40	10	3·3	3·0	2·9
40	15	3·5	2·8	2·7
40	20	3·8	3·0	2·8

7.4. Comparison of MLLR to least squares regression

Experiments were performed using the standard least squares regression (as discussed in Subsection 4.1) to estimate the regression matrix, and compared to the MLLR method. Exactly the same experimental set-up was used, the only difference being the adaptation phase. The adaptation data was aligned using a Viterbi alignment of the correct labels (i.e. each frame was assigned to a single mixture component). The effect of including the variance in the estimation of the transform was also examined by generating MLLR transforms using a Viterbi alignment.

The comparison (Table I) shows that MLLR gives a lower error rate than least squares optimization. The difference is due mainly to the inclusion of variance information in the estimation of the transform since the two alignment strategies give similar MLLR results. The least squares approach made the invalid assumption that each distribution

within the same regression class has the same covariance matrix. As the number of classes increases, the difference between least squares and MLLR performance increases showing the importance of correctly incorporating the covariance matrices into the regression matrix calculation.

8. Conclusion

A new speaker adaptation method for continuous density HMMs called maximum likelihood linear regression (MLLR) adaptation has been described. The method adapts a set of SI models to a specific speaker by applying a set of linear transformations to the Gaussian mean vectors. Each transformation is used for a number of Gaussian distributions, and the number of transformations is determined by the amount of adaptation data available. The parameters of the transformation matrices are estimated to maximize the likelihood of the speaker specific data.

MLLR can be seen as an extension to least squares regression. The assumption of equal covariances has been removed, and the adaptation data can be assigned to distributions on a probabilistic basis. The improvement is reflected in the experiments which show that MLLR is more effective than least squares, especially when more regression classes are used.

The method has been evaluated on the ARPA Resource Management corpus using a mixture Gaussian cross-word triphone system. Experiments have shown that full regression matrices are more effective than diagonal matrices, and that improvements in recognition rates can be achieved using only a few seconds of adaptation data. As more data is used for adaptation a corresponding increase in performance can be achieved by matching the number of regression transformations in the system to the quantity of data available. The performance tends towards that of an SD system as more data is used. Implementing the method in an unsupervised manner gives only a marginal degradation in performance from a supervised approach.

It has been shown that MLLR can be applied to continuous density HMMs with a large number of Gaussians and is effective with small amounts of adaptation data. Furthermore, since any data can be used for adaptation, it is possible to integrate MLLR transparently into a recognition system in an incremental unsupervised adaptation mode.

References

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1–8.

Bellegarda, J. R., de Souza, P. V., Nadas, A. J., Nahamoo, D., Picheny, M. A. & Bahl, L. R. (1992). Robust speaker adaptation using a piecewise linear acoustic mapping. *Proceedings of the ICASSP* **1**, 445–448.

Choukri, K., Chollet, G. & Grenier, Y. (1986). Spectral transformations through canonical correlation analysis for speaker adaptation in ASR. *Proceedings of the ICASSP* **4**, 2659–2662.

Class, F., Kaltenmeier, A., Regel, P. & Trottler, K. (1990). Fast speaker adaptation for speech recognition systems. *Proceedings of the ICASSP* **1**, 133–136.

Cox, S. J. & Bridle, J. S. (1989). Unsupervised speaker adaptation by probabilistic spectrum fitting. *Proceedings of the ICASSP* **1**, 294–297.

Digalakis, V., Rtischev, D. & Neumeyer, L. (1995). Fast speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing* (in preparation).

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* **2**, 291–298.

Hewett, A. J. (1989). Training and speaker adaptation in template-based speech recognition. PhD Thesis, Cambridge University.

Jaschul, J. (1982). Speaker adaptation by a linear transformation with optimised parameters. *Proceedings of the ICASSP* **3**, 1657–1670.

Juang, B.-H. (1985). Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. *AT&T Technical Journal* **64**, 1235–1249.

Kendall, M. G. (1971). *The Advanced Theory of Statistics*, Vol. 2. Charles Griffin & Company, London.

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* **39**, 806–814.

Lee, K.-F. (1988). Large vocabulary speaker independent speech recognition: The SPHINX system. PhD Thesis, Carnegie Mellon University.

Leggetter, C. J. & Woodland, P. C. (1994). Speaker adaptation using linear regression. Technical Report, CUED/F-INFENG/TR.181, Cambridge University Engineering Department.

Liporace, L. A. (1982). Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory* **IT-28**, 729–734.

Odell, J. J., Valtchev, V., Woodland, P. C. & Young, S. J. (1994). A one pass decoder design for large vocabulary recognition. *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, Princeton NJ, pp. 405–410.

Young, S. J., Odell, J. J. & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, Princeton NJ, pp. 307–312.

Young, S. J., Woodland, P. C. & Byrne, W. J. (1993). *HTK—Hidden Markov Model Toolkit, Version 1.5*, Cambridge University Engineering Department and Entropic Research Laboratories Inc.

(*Received 1 December 1994 and accepted for publication 1 February 1995*)