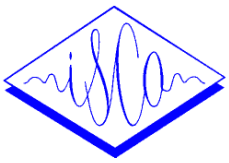

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

HIDDEN MARKOV MODELS USING SHARED VECTOR LINEAR PREDICTORS

B.A. Maxwell & P.C. Woodland
Cambridge University Engineering Department, England

ABSTRACT

It has been previously shown that augmenting a standard HMM with a set of vector linear predictors can improve recognition rates compared with standard HMMs. The set of vector linear predictors associated with each state improve the HMMs ability to model the correlations in real speech data, and help to overcome the HMM state-conditional independence assumption. However, introducing extra parameters into the model requires more training data. This problem can be partly overcome by sharing the predictor parameters between multiple HMM states, and hence more robust, but less specific estimates of the predictor parameters are obtained. This paper develops the theory and implementation of arbitrarily shared vector linear prediction for hidden Markov models. For most predictor offsets, predictors shared across all states of an HMM provide more accurate recognition on both training and test data sets than equivalent HMMs without predictors when evaluated on a British English E-set recognition task.

Keywords: Speech recognition, vector prediction

1. INTRODUCTION

Hidden Markov models (HMMs) are very successful as models for speech recognition. However, the assumptions made by an HMM as to the underlying nature of the speech signal are poor. In particular, it is assumed that each observation vector is conditional on the state that produced it but not on the other observation vectors (the "independence assumption"). Modelling of signal dynamics can be somewhat improved by the now standard practice of including first derivative information as part of the HMM observation vector. Recently, it has been shown that also including higher order derivatives can lead to improved performance. However, this approach leads to non-robust estimates of the higher order derivatives and it becomes impractical to add many more derivative terms and use full covariance matrices in the Gaussian output distribution due to the large number of parameters involved.

An alternative approach to the independence problem is obtained by viewing an HMM as a signal predictor. Each state of a standard (continuous density) HMM predicts the next observation to be the state mean vector, and the prediction error is modelled by the state covariance matrix. However, due to the correlation inherent in real speech, a better prediction of the next observation can be made by taking into account nearby observations. Again the covariance matrix models the prediction error, but the predicted value is computed using vector linear prediction. In effect, the mean vector of each state for the current observation is no longer constant but dependent on other observations. In the work described here, the predictors operate on data at arbitrary offsets from the current speech frame and the predictors themselves can be either diagonal or full matrices.

Vector linear prediction explicitly models correlation in the signal. Wellekens [7] developed a method in which the correlation between the current frame and the previous frame is explicitly modelled. However this model performed poorly [8]. Similar work by Brown [3] also operated on the previous frame, but he reported that it produced worse results than his baseline HMMs. Brown also reported that his 'conditional' Gaussian model method produced a significantly higher average log probability per frame[1] than the baseline HMMs during training despite the poor recognition results. Kenny *et al* [4] describe a system using a prediction model that is, in fact, equivalent to the formulation given in [10] and used here, although the form of the model they present is expressed in terms of slightly different parameters. They also reported results that were worse than a standard HMM that used observation vectors with both static and differential coefficients.

The vector linear prediction method used here is an extension of the work in [10]. The results quoted there, and the results presented here show that vector linear prediction can reduce the error rate on both test and training sets.

While the addition of a vector linear predictor to an HMM gives improved modelling ability, there is an increase in the number of parameters to be estimated from limited training data. The use of state-specific predictors can lead to overtraining. For instance, in [10], it was found that using two full predictors for each state actually increased the error rate on the test set while decreasing the error rate on the training set. One way to reduce the number of parameters is to tie (share) the predictors among a set of HMM states. This allows a more robust estimate of the predictor parameters to

[1]The average log probability per frame is often used as an indicator of how well an HMM matches the training sequences.

be made from the available training set, possibly leading to better test set recognition performance. Using a single predictor shared across all states of an HMM also results in significant memory savings.

In the next sections the theory of HMMs with tied predictors is developed, and then the method is evaluated using speech recognition experiments on a British English E-set recognition task. The results obtained show performance on the test and training sets to be between that of the baseline model without prediction and HMMs with non-shared predictors when diagonal covariance matrices were used. All experimental work used a version of the HTK HMM toolkit modified to include HMMs with arbitrary sharing of vector linear predictors.

2. PREDICTOR MODEL

The formulation of HMMs using shared vector linear predictors is based upon the work presented in [10]. Output distributions are associated with each state.

The output Y_t^i at time t from state i is predicted using

$$Y_t^i = \mu_0^i + \sum_{p=1}^{P} A_p^i (O_{t+q_p} - \mu_{q_p}^i) \tag{1}$$

where there are P predictors associated with i and the pth predictor is at offset q_p. A standard HMM can be obtained from this equation by setting all A_p to the zero matrix.

The predictor model has three basic parts, the standard mean, μ_0, a set of offset means, μ_{q_p}, which are the means of the observations q_p frames, or time steps, away from the current observation, and the set of predictor matrices, A_p. The O_{t+q_p} are the observations at offset q_p from time t. Note that sharing the A_p between a number of states does not affect the calculation of the output probabilities.

The error between the actual observation O_t and the output of state i is given by (2).

$$E_t^i = O_t - Y_t^i \tag{2}$$

Given the covariance matrix of the prediction error S_i for state i, the probability (density) of observation O_t in state i is given by

$$p(O_t) = \frac{1}{(2\pi)^{\frac{d}{2}} |S_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(E_n^{i^t} S_j^{-1} E_n^i)} \tag{3}$$

3. RE-ESTIMATION EQUATIONS

The maximum likelihood (MLE) re-estimation equations for the vector linear prediction model can be determined by using the methods of Baum and his coworkers [1], [2], [5]. An appropriate auxiliary function is defined and differentiated with respect to the unknown parameters. Each resulting equation is then set equal to zero and simple manipulation then leads to a formula for each parameter. For a complete derivation of the training equations for shared predictors, see [6].

The changes to the HMM structure by the use of vector linear prediction only affects the output probability calculations, and therefore the standard re-estimation formula for training transition probabilities applies to the HMM model with vector linear prediction.

The re-estimation equation for the state means and the offset means used in (1) is given by (4). The variable $\gamma_i(t)$ is the *a posteriori* probability of occupying state i at time t, and the q_p is the offset from the current frame. When $q_p = 0$ the mean is a state mean.

$$\hat{\mu}_{q_p}^i = \frac{\sum_T \gamma_i(t) O_{t+q_p}}{\sum_T \gamma_i(t)} \tag{4}$$

To estimate the predictor matrices, first the covariance matrix, shared between all of the states in the set I_p, between observation vectors at offsets r and s for a set of states is computed

$$C_{rs}^{I_p} = \frac{\sum_{i \in I_p} \sum_T \gamma_i(t) (O_{t+r} - \hat{\mu}_r^i)(O_{t+s} - \hat{\mu}_s^i)^t}{\sum_{i \in I_p} \sum_T \gamma_i(t)} \tag{5}$$

Once these covariance matrices have been found, the new estimates for the P predictor matrices $\hat{A}_p^{I_p}$ are found by solving the matrix equation given by (6), where $\mathcal{B}_I$, $\mathcal{R}_I$, and $\mathcal{Z}_I$ are defined as in (7) and (8).

$$\mathcal{R}_{I_p} = \mathcal{Z}_{I_p} \mathcal{B}_{I_p}^t \tag{6}$$

$$\mathcal{R}_{\mathcal{I}} = \begin{bmatrix} C_{10}^I \\ \vdots \\ C_{P0}^I \end{bmatrix} \quad \mathcal{Z}_{\mathcal{I}} = \begin{bmatrix} C_{10}^I & \cdots & C_{1P}^I \\ \vdots & \ddots & \vdots \\ C_{P1}^I & \cdots & C_{PP}^I \end{bmatrix} \tag{7}$$

$$\mathcal{B}_{\mathcal{I}} = \begin{bmatrix} \hat{A}_1^I & \cdots & \hat{A}_P^I \end{bmatrix} \tag{8}$$

Note that even when the predictors are shared over a set of states, the offset means $\hat{\mu}_{q_p}$ used to calculate the $C_{rs}^{I_p}$ are specific to each state over which the predictor, and the $C_{rs}^{I_p}$, is shared. In theory, the offset means could be shared among several states without affecting the form of the re-estimation equations, but this alternative was not pursued.

To compare the consistency of the shared predictor re-estimation formulas with previous work, the above equations easily reduce to the non-shared case given in [10] when the set of shared states, I_p, consists of only a single state.

The re-estimation equation for the prediction error covariance matrix, $\hat{S}_{I_s}$, for the set of states I_s is given by (9). This follows the form of the standard re-estimation formula for the covariance matrix of a Gaussian output distribution ([5]), but the contributions from each of the shared states in I_s are summed on the numerator and denominator.

$$\hat{S}_{I_s} = \frac{\sum_{i \in I_s} \sum_T \gamma_i(t) \hat{E}_t^i \hat{E}_t^{i^t}}{\sum_{i \in I_s} \sum_T \gamma_i(t)} \tag{9}$$

Note that the $\hat{E}_t$ are based upon the re-estimated values $\hat{\mu}_0$, $\hat{\mu}_{q_p}$, and $\hat{A}_p$.

The terms of an expanded version of (9) can be substantially simplified and re-written in terms of (7) and (8).

$$\hat{S}_{I_s} = \hat{C}_{00}^{I_s} - \mathcal{B}_{I_p}\mathcal{R}_{I_s} - (\mathcal{B}_{I_p}\mathcal{R}_{I_s})^t + \mathcal{B}_{I_p}\mathcal{Z}_{I_s}\mathcal{B}_{I_p}^t \quad (10)$$

If the sets $I_p = I_s$ and each consists of a single state, this equation simplifies to the one given by [10].

As the sets I_p and I_s are not necessarily the same, care must be taken during re-estimation of the predictors $\hat{A}_p^{I_p}$ and covariance matrices $\hat{S}_{I_s}$ that the parameters from the correct states are used. Note also that the matrices $\mathcal{R}_I$ and $\mathcal{Z}_I$ are not necessarily equivalent in (10) and (6). $\mathcal{B}_{I_s}$, on the other hand, must consist of the re-estimated predictors as calculated by (6). If the simplified form of the training equations given by (10) is used to calculate $\hat{S}_{I_s}$, then $I_s \subseteq I_p$ in order for S_{I_s} to be properly re-estimated.

Both the predictor and covariance matrices can be either full or diagonal. The matrix elements of the re-estimation equation given by (10) must be adjusted accordingly depending upon the combination used [6].

4. SPEECH DATABASE AND EXPERIMENTAL SETUP

The data set used to evaluate shared linear prediction was a British English E-set ('B', 'C', 'D', 'E', 'G', 'P', 'T', & 'V') used in a multiple speaker mode. This data forms part of a spoken alphabet data set collected and distributed by British Telecom Laboratories. The database contains three utterances from each of 104 speakers (54 male speakers, and 50 female speakers). The data was recorded in a quiet-room at a sampling rate of 20kHz, and then end-pointed semi-automatically. For these experiments, the first two utterances by each speaker were used as the training data (1634), and the third utterance as the test data (824).

The data was analysed using a 27 channel Mel-scaled filter-bank at a 100Hz frame rate and then represented using 12 Mel-frequency cepstral coefficients and their first differentials were then calculated, producing a 24 element observation vector to characterize each frame. Each speech vector was then rotated and scaled so that the average within state covariance matrix was the identity matrix as described in [9]. The means were also adjusted to be, on average, zero.

All of the HMMs used in these experiments were strictly left to right with no skips. The eight consonants were each modelled by an HMM with four emitting states, and a single HMM with seven emitting states was used for the common vowel. Each utterance, therefore, was modelled by 11 states, with the last 7 being shared among all words. It should be noted that while the data sets are identical to those used in [10], the HMMs used have fewer states and hence the results are not strictly comparable.

The HMMs were initialised using uniform segmentation to provide an initial estimate of the means and covariances. Multiple iterations of Baum-Welch re-estimation on the individual HMM models was then used to further estimate the means, covariances, and transition probabilities. Neither of these initial phases affected or trained the predictors or the offset means, which were initialised to zero. Finally, embedded Baum-Welch re-estimation was used for four iterations to train the HMMs with the predictor model implemented as outlined above.

5. EXPERIMENTAL RESULTS

Two baseline HMM sets were developed for comparison purposes, both using a 24 dimensional observation vector and a single Gaussian distribution per state. One model used a diagonal covariance matrix, the other a full covariance matrix. The results for the test and training sets are given in Table 1 [2]. As can be seen, the use of a full covariance matrix significantly improves recognition of both the training and test data.

Cov. Type	Log Prob	Train Err %	Test Err %
diagonal	-30.9	10.8	12.9
full	-27.7	2.9	7.6

Table 1. Average Log Probability per Frame and Recognition Results for the Baseline HMM set

To compare the effects of sharing a diagonal or full predictor among all states of a given HMM, an HMM set was trained and tested for each of the four possible covariance/predictor combinations. Two HMM sets were also trained with non-shared predictors for the diagonal covariance/diagonal predictor and full covariance/diagonal predictor cases. In all cases, following [10] a single predictor was used with an offset of -3. These results are given in Table 2.

Cov	Pred	Log Prob	Train Err	Test Err
diag	shd diag	-25.7	7.6	8.4
diag	ind diag	-25.1	6.0	7.8
diag	shd full	-21.6	6.7	7.9
full	shd diag	-22.6	2.4	5.3
full	ind diag	-22.1	3.1	5.4
full	shd full	-16.1	2.0	5.7

Table 2. Average Log Probability per Frame and Recognition Results for Different Combinations of Covariance and Predictor Types

It is apparent from the results in Table 2 that a shared predictor does provide significant improvement over the baseline models, but not as much improvement as non-shared predictors. The use of a single shared diagonal predictor for all states of each HMM has reduced the test set error rate of the diagonal covariance models by 35% and of the full covariance models by 30%. It may be noted that when a full covariance matrix is used, the shared diagonal predictor actually has a lower error rate than the non-shared diagonal predictors on

[2] All of the average log probabilities per frame given in the tables of results are one iteration behind the HMM set for which the results are reported.

both the test and training data. These results demonstrate that the use of shared vector linear predictors is an efficient means of improving the modelling ability and enhances discriminatory power.

To examine the effect of using different offsets with a single predictor, experiments were performed on HMMs with a diagonal covariance matrix and a shared diagonal predictor for all offsets between -8 and +8. The specific values are listed in Table 3. For this task and HMM structure, the offsets from +3 to +7 provide the best recognition results. [3]

Offset	Log Prob	Train Err %	Test Err %
-8	-29.8	9.7	10.7
-7	-29.3	9.7	10.8
-6	-28.8	8.8	10.0
-5	-28.1	8.1	9.0
-4	-27.1	7.6	8.5
-3	-25.7	7.6	8.4
-2	-22.9	9.0	9.5
-1	-13.9	16.0	16.0
1	-13.7	12.8	15.7
2	-22.9	8.5	9.5
3	-25.7	7.3	8.1
4	-27.2	7.0	7.8
5	-28.1	7.2	8.4
6	-28.8	7.4	8.9
7	-29.3	7.5	8.0
8	-29.6	7.7	9.1

Table 3. Average Log Probability per Frame and Recognition Results for HMM sets with Diagonal Covariance Matrices and Diagonal Globally Shared Predictors at Different Offsets

The results in Table 3 show that the best negative is -3 while the best positive offset is +4. Note that in the case of a ± 1 predictor, especially the -1 case, the average log probability—usually an indicator of how well the model fits the data—is extremely high compared to the other offsets. This is because all speech utterances are highly correlated between adjacent frames (and that the original observation data includes differential information). By using a ± 1 offset the value of the next observation vector can be accurately predicted by a predictor that is very similar for all states. This decreases the discrimination ability since each model fits a large subset of the utterances with a high probability. Several sets of predictor pairs were also chosen and tested, but no pair of predictors outperformed the better single predictor results given above. Further details of these experiments are given in [6].

6. CONCLUSIONS

The theory and implementation of HMMs which used arbitrarily shared vector linear predictors has been discussed. The recognition results show that the use of shared predictors can reduce both the test and training set error rates compared to HMMs without predictors. A single diagonal predictor per HMM decreased the test set error rate by 35% when a diagonal covariance matrix was used and by 30% with a full covariance matrix, and the test set error rates are broadly similar to those obtained when using an individual predictor per state. This is because the parameters are estimated more robustly through parameter sharing, although the resulting values are less specific. It was also shown that the predictor offsets must be carefully chosen to maximise recognition performance. In particular offsets of either -3 and -4 or offsets between +3 and +7 give good performance, although it should be noted that the optimal offsets are probably task dependent.

Overall, shared vector linear prediction has been shown to be an efficient way of increasing recognition performance of HMMs without significantly increasing memory use or computational loads.

[3] Results are not shown in Table 3 for offset 0 since this corresponds to perfect knowledge of the current speech vector through an identity predictor. This results in no discrimination ability as well as an infinite output probability density!

REFERENCES

[1] Baum L.E., Petrie T., Soules G. & Weiss N. (1970). A maximization technique occurring in the statistical analysis of probablistic functions of Markov chains. *Ann. Math. Stat.*, **41**, 164-171.

[2] Baum L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities*, **3**, 1-8.

[3] Brown P.F. (1987). The acoustic modelling problem in automatic speech recognition. IBM Technical Report No. RC 12750.

[4] Kenny P., Lennig M. & Mermelstein P. (1990). A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Trans. ASSP*, **38**, 220-225.

[5] Liporace L. A. (1982). Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Trans. Information Theory*, **28**, 729-734.

[6] Maxwell B.A. (1992). Hidden Markov models using shared and global vector linear predictors. M.Phil Thesis Cambridge University.

[7] Wellekens C. (1987). Explicit time correlation in hidden Markov models for speech recognition. *Proc. ICASSP'87*, 384-386, Dallas.

[8] Wellekens C. (1992). Personal Communication.

[9] Woodland P.C. & Cole D.R. (1991). Optimising hidden Markov models using discriminative output distributions. *Proc. ICASSP'91*, 545-548, Toronto.

[10] Woodland P.C. (1992). Hidden Markov Models Using Vector Linear Predictors and Discriminative Output Distributions. *Proc. ICASSP'92*, San Francisco.

[11] Young S.J. (1992) HTK V1.4 User, Reference & Programmer Manual. Cambridge Univsersity Engineering Dept, Speech Group, August 1992.