RAPID SPEAKER ADAPTATION USING MODEL PREDICTION

S.M. Ahadi & P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

A key issue in speaker adaptation is gaining the maximum information from a limited amount of adaptation data. In particular it is important that observations of parameters of (context-dependent) HMMs not occurring in the adaptation data can be updated. In the Regression-based Model Prediction (RMP) approach, sets of speaker-independent linear relationships between different parameters in the HMM set are found from training data. During adaptation, distributions with sufficient adaptation data are used to update the parameters of poorly adapted models using these pre-computed regression-based relationships. The method used Bayesian techniques to combine parameter estimates from different sources. Evaluation on the ARPA Resource Management corpus gave a worthwhile reduction in error rate with just a single adaptation sentence, and that RMP consistently outperforms MAP estimation with the same amount of adaptation data.

1. INTRODUCTION

Although modern speaker independent (SI) continuous speech recognition systems show impressive performance, their error rates are still much higher than a well-trained speaker dependent (SD) system. Speaker adaptation techniques that attempt to adapt the parameters of a speaker independent system to get speaker-dependent 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.

A standard adaptation approach for continuous density HMM systems uses Maximum a Posteriori (MAP) parameter estimation [1, 2] 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 [3] 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 describes a method for rapid adaptation called Regression-based Model Prediction (RMP). Linear relationships between particular system parameters are computed which 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.

Earlier work on regression based prediction for adaptation was reported by Cox [4] who worked on single Gaussian HMMs for a small vocabulary, isolated word recognition task. The RMP method can be viewed as an extension of this work to 1000 word vocabulary continuous speech recognition using context dependent mixture Gaussian models containing orders of magnitude more parameters. To accommodate this change in focus the method has been considerably extended and refined.

2. ADAPTATION ALGORITHM

The RMP algorithm operates in two stages. First conventional MAP estimation is applied to re-estimate the mean vectors given the adaptation data, and then the predictive linear relationships are used to update the target model parameters given the MAP estimated source model parameters.

2.1. Initial MAP Estimation

Initially the forward-backward MAP algorithm [2] is used to estimate an initial set of speaker adapted models using the available source data. This algorithm combines speaker independent prior estimates of model parameters with the adaptation data in an Expectation-Maximization (EM) framework. This approach gives good estimates of model parameters if there is a few samples of adaptation data for that parameter.

In the work reported here, this initial MAP estimation phase is applied to the HMM mean vector only. Given an adaptation sequence $\mathbf{x} = (x_1, \ldots, x_T)$ for an HMM with output distributions consisting of a mixture of Gaussian densities, the MAP re-estimate of the Gaussian means for

0-7803-2431-5/95 $4.00 © 1995 IEEE

HMM state i, mixture component k is [2]

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

where μ_{ik} is the prior mean, τ_{ik} is a parameter controlling the relative weight of the prior and adaptation data, x_t is the adaptation data samples, and c_{ikt} is the probability of being in state i and mixture component k at time t given that the model generates the sequence $\mathbf{x}$.

2.2. Model Prediction

In this stage linear relationships between model mean parameters are used to update target distributions using the MAP-adapted source distributions. The linear relationships between source and target parameters are derived from speaker-dependent sets of models using multiple linear regression and have the form

$$y = a_0 + \sum_{l=1}^{P} a_l x_l \qquad (2)$$

where y is a target mean, the x_l are the values of the corresponding mean elements from the source distributions, the a_l are regression coefficients, and P is the regression order (number of source distributions for each target). This section first describes how the regression coefficients are computed and then how the target distributions are updated.

First a set of N speaker dependent models are trained. It is assumed that the parameters of these models are exact for the speakers that they represent. Then, the regression coefficients a_l can be found from solving the following matrix equation [5]

$$\mathbf{U}\mathbf{a} = \mathbf{V} \qquad (3)$$

where $\mathbf{U}$ is a $P \times P$ matrix, and $\mathbf{a}$ and $\mathbf{V}$ are $P \times 1$. The matrix elements are given by

$$U_{kl} = \sum_{i=1}^{N}(x_{ki} - \bar{x}_k)(x_{li} - \bar{x}_l) \qquad (4)$$

$$V_l = \sum_{i=1}^{N}(y_i - \bar{y})(x_{li} - \bar{x}_l) \qquad (5)$$

where y and x are the target and source distribution parameters. The value of a_0 can be found using $\bar{x}_l$, $\bar{y}$, and a_l in (2).

The squared multiple correlation coefficient between the target y, and the regression predicted target value y' can be found by

$$R^2_{yy'} = \frac{\sum_{l=1}^{P} a_l \sum_{i=1}^{N}(y_i - \bar{y})(x_{li} - \bar{x}_l)}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad (6)$$

where N is the number of speakers (SD model sets) taking part in the regression.

Given the source parameter values and a set of regression coefficients, an estimate of the target value can be found by (2). In order to combine this estimate with the estimate, for the target parameter, derived from the initial MAP phase using a Bayesian approach (with the initial MAP estimate as a prior), the variance of the predicted target parameter is required.

Assuming that the mean element values are normally distributed over different speakers, then the estimated sample variance due to multiple regression can be calculated as [6]

$$\begin{aligned} s_e^2 &= s_y^2(1 - R^2_{yy'})\frac{N-1}{N-P-1} \\ &= \frac{1}{N-P-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 \\ &\quad - \frac{1}{N-P-1}\sum_{l=1}^{P} a_l \sum_{i=1}^{N}(y_i - \bar{y})(x_{li} - \bar{x}_l) \end{aligned} \qquad (7)$$

where s_y^2 is the sample variance for the target element in the set of SD systems. The quantity s_e^2 is the estimated variance of the target elements predicted using the regression parameters (a_l) and particular source distribution parameter values. However, there is additional variance in the target values due to the fact that the source distribution values are not exact (i.e. not the true SD values), but are based on the initial MAP estimate which uses a relatively small amount of adaptation data.

In order to find the additional variance due to the errors in the source distribution parameter estimates, the sample variance of the MAP estimated parameters compared to the true (i.e. SD) values is computed for each element of each source distribution as follows

$$s_{\nu_l}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[(x_{li} - \nu_{li}) - \frac{1}{N}\sum_{j=1}^{N}(x_{lj} - \nu_{lj})\right]^2 \qquad (8)$$

where ν_{li} stand for the MAP estimated source elements. This extra variance due to the source distribution estimate error will decrease as more adaptation data is used.

Therefore the total variance for the regression predicted target element, assuming, for the sake of simplicity, that source distributions are independent is

$$s_\mu^2 = s_e^2 + \sum_{l=1}^{P} a_l^2 s_{\nu_l}^2. \qquad (9)$$

Given the variance of the predicted target element, a Bayesian approach to combine the regression-predicted element with the prior estimate from the initial MAP estimation phase can be used. Assuming the distribution of the means to be Gaussian, and using the concept of conjugate priors the final estimate for any mean element of the target, m, can be found by [7]

$$m = \frac{\mu s_\zeta^2 + \zeta s_\mu^2}{s_\zeta^2 + s_\mu^2} \qquad (10)$$

where ζ and s_ζ^2 are initial MAP estimated mean for the target and the variance associated with it, and μ and s_μ^2 are the regression predicted mean and its variance respectively.

Finally, the sample variance for the initial MAP estimated priors in (10) can be found using the SD model set to compute the variance of the target initial MAP estimated means:

$$s_\zeta^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left[(y_i-\zeta_i) - \frac{1}{N}\textstyle\sum_{j=1}^{N}(y_j-\zeta_j)\right]^2 . \quad (11)$$

2.3. RMP Implementation

The RMP adaptation algorithm, as described, would be computationally expensive unless some constraints are applied to it. For example, a 3-state triphone HMM set with 2 mixture components per state, in our implementation, contains more than 3000 mixture components. This means that about 350 million relationships are required before the best correlated distributions can be found. Even larger numbers of relationships must be found when more Gaussians are used.

In order to reduce this computation to be more manageable, and also to separate the better-trained source distributions from the more poorly-trained targets, the search for source distributions for each target is restricted. Target distributions are distinguished from the sources by having occurred less than a prespecified threshold number of times in the adaptation data, while source distributions must have received sufficient adaptation; the source and target distributions must occur in the same HMM state position, and must be from the same broad speech class (vowel or consonant). The source distributions, subject to the above restrictions, are then chosen for each target to have the maximum average squared correlation coefficient. A threshold is also set on the correlation coefficients to prevent weak relationships from being used in adaptation.

3. EXPERIMENTS AND RESULTS

The RMP approach has been evaluated on the ARPA RM1 continuous speech database. The HTK HMM toolkit was used for all model building and recognition, and special tools were written for forward-backward MAP estimation and model prediction. The database was parametrised using 12 Mel frequency cepstral coefficients, normalised log energy and the first and second differentials of these parameters. A baseline SI state-clustered word-internal triphone gender independent system was trained using the 109-speaker SI training data. The state clustering procedure used the decision tree method described in [8]. 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, and all word error rates 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. For 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. 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. 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. Table 1 shows both ML SD results (under SD) and MAP SD results (under MAP with 600 training sentences). All models built using MAP estimation set the τ_{ik} parameter in (1) to the fixed value of 10. Due to the limited number of speakers in the speaker dependent portion of the RM database, during testing for each speaker, the data from all other 11 speakers was used to build the SD models and compute regression parameters. Hence, this reduces the number of SD models available to estimate the regression parameters, as well as requiring both stages of adaptation to be carried out for each individual speaker. These problems both could be avoided if a larger number of speakers was available in the SD portion of the database. If all the speakers in the SD model set are used for the regression calculations, including the speaker currently under test, a very small further decrease in word error rate is obtained.

	Number of Adaptation Sentences					
	0	1	10	40	100	600
	Single Gaussian Models					
SI	10.21	-	-	-	-	-
SD	-	-	-	-	-	4.63
MAP	10.21	10.54	9.95	7.96	6.64	4.07
RMP	10.21	9.42	8.37	7.06	6.15	4.07
	2 Component Mixture Models					
SI	7.87	-	-	-	-	-
SD	-	-	-	-	-	3.13
MAP	7.87	7.76	7.15	6.35	5.29	2.82
RMP	7.87	7.31	6.72	5.90	4.98	2.82
	6 Component Mixture Models					
SI	5.92	-	-	-	-	-
SD	-	-	-	-	-	2.14
MAP	5.92	5.99	5.74	5.09	4.09	2.06
RMP	5.92	5.44	5.19	4.56	3.86	2.06

Table 1. Percent word error rates obtained with SD, SI, MAP, and RMP for single Gaussian, two component, and six component Gaussian mixture models. All results are averaged over all 12 available test speakers.

In the experiments reported here, there were two source distributions for each target. After finding suitable source distributions, the correlations, regression coefficients and variances were computed using the relationships given in (7), (8), and (11). Finally, for each target distribution, the initial MAP estimated values for each parameter were combined with the regression-based predicted values using (10).

Experiments were performed to compare the effectiveness of the RMP approach with a standard MAP approach for various numbers of adaptation sentences and model sets with single Gaussian, two component Gaussian mixtures and six component mixtures. Table 1 gives word error rates with no adaptation data (i.e. SI performance), and for points between one and 600 adaptation utterances. With 600 utterances the variance of the MAP estimate computed by (11) is zero, and hence the MAP and RMP values are

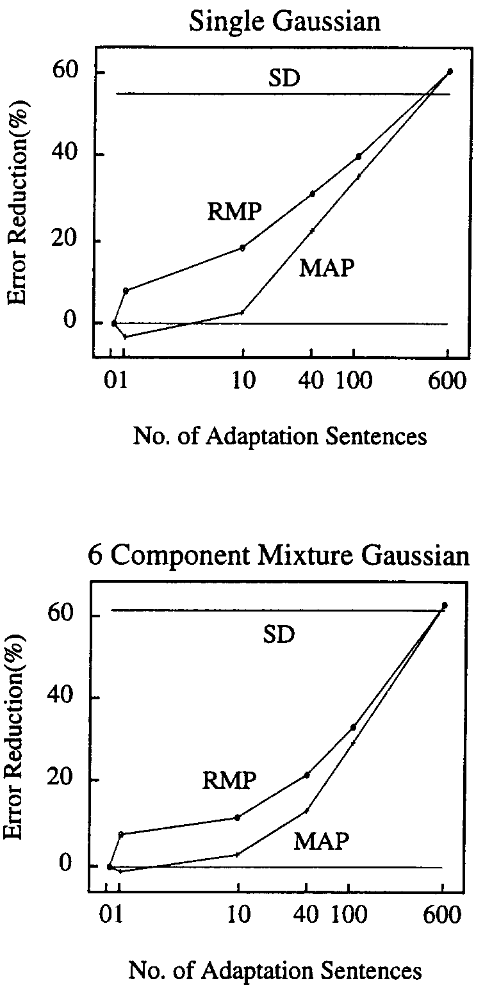

Figure 1. Percentage of word error rate reduction obtained for single Gaussian and six component mixture Gaussian models versus the number of adaptation sentences for SD, MAP, and RMP averaged over all 12 speakers.

identical. For small amounts of adaptation data the RMP approach leads to a significant reduction in error compared to both the speaker independent systems and MAP.'It is of interest that the improvement using only one adaptation sentence, which is on average about 3 seconds in duration, leads to an 8% reduction in word error rate for the 6 component mixture Gaussian system. This corresponds to about 13% of the error reduction of the speaker dependent system, and is shown in the Figure 1. It should also be of interest that in this case, the average result obtained using MAP estimation does not show any improvement over SI result. Another point to be noted here is that in the 6 mixture one adaptation sentence RMP tests, the results obtained for all individual speakers except one, either outperformed or were equal to SI or MAP, which is a good indication of robustness of this technique with very small amounts of adaptation data. As can be seen in the Figure 1, the amount of the error reduction obtained with RMP is consistently above that for MAP estimation for all system configurations and amounts of adaptation data below 600 sentences. It should be emphasised that in this figure, the horizontal axes are non-linear to show the effects of using small amounts of adaptation data.

4. CONCLUSIONS

A technique for rapid speaker adaptation has been presented that exploits correlations between system parameters to predict suitable values for model parameters not observed in the adaptation data. Experiments show that with only one adaptation sentence significant reductions in error rates can be achieved, and that with larger amounts of adaptation data the technique converges to speaker dependent performance.

ACKNOWLEDGEMENTS

S.M. Ahadi is supported by a studentship from the Iranian Ministry of Culture and Higher Education.

REFERENCES

[1] C-H. Lee, C-H. Lin, & B-H. Juang (1991). A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models. *IEEE Trans. Sig. Proc.*, Vol. 39, No. 4, 806–814.

[2] J-L. Gauvain & C-H. Lee (1994). Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. SAP*, Vol. 2, No. 2, 291–298.

[3] C.J. Leggetter & P.C. Woodland (1994). Speaker Adaptation of Continuous Density HMMs Using Linear Regression. *Proc. ICSLP'94*, Yokohama.

[4] S.J. Cox (1993). Speaker Adaptation Using a Predictive Model. *Proc. Eurospeech'93*, Vol. 3, 2283–2286, Berlin.

[5] S. Chatterjee, & B. Price (1991). *Regression Analysis by Example.* John Wiley & Sons, New York, 2nd Edition.

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

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

[8] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young (1994). Large Vocabulary Continuous Speech Recognition Using HTK. *Proc. ICASSP'94*, Vol. 2, 125–128, Adelaide.