SPEAKER ADAPTATION OF CONTINUOUS DENSITY HMMS USING MULTIVARIATE LINEAR REGRESSION

C.J. Leggetter *P.C. Woodland*

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

ABSTRACT

A method of speaker adaptation for continuous density mixture Gaussian HMMs is presented. A transformation for the component mixture means is derived by linear regression using a maximum likelihood optimisation criteria. The best use is made of the available adaptation data by defining equivalence classes of regression transforms and tying one regression matrix to a number of component mixtures. This allows successful adaptation on any amount of adaptation data. Tests on the RM1 database show that successful adaptation can be achieved with only 11 seconds of speech, and performance converges towards that of speaker dependent training as more adaptation data is used.

1. INTRODUCTION

Recent advances in speech recognition research have resulted in high performance speaker independent recognition systems [6]. These systems perform well because they use large amounts of training data to provide detailed modelling of speech patterns. Even with good speaker independent systems some speakers are modelled poorly, and it is clear that improvements could be made by tuning the system parameters to improve the modelling of an individual speaker. To train a full speaker dependent (SD) system would require a large amount (hours) of training data from the speaker, and even then some speech phenomena may not be present. The modelling of speech phenomena that are not present in specific speaker data (e.g. unseen triphone contexts) is a major problem and is a source of many errors.

In many cases, retraining the whole system for a new speaker is undesirable, and hence methods of adapting recognition systems to a specific speaker are of great interest. An ideal adaptation technique would use any available data to adapt the system, with the performance of the adapted system improving as more data is used. A variety of adaptation approaches have previously been reported but most algorithms (e.g. MAP estimation [4]) either require a reasonably large amount of speaker data for successful adaptation or are applicable only to discrete density HMMs.

Here, a method of adaptation for continuous density HMMs is presented which is aimed at performing adaptation using limited data from the new speaker. Adaptation is performed on small amounts of data by capturing general characteristics of the speaker with respect to the original system. As further data becomes available, more specific speaker effects can be captured. The method can be viewed as a generalisation of the spectral mapping approach [2].

The method uses an initial set of good speaker independent (SI) models and adapts the model parameters to the new speaker by transforming the mean parameters of the models with a set of linear transforms. The transformations are found using a maximum likelihood criteria which is implemented in a similar fashion to the standard ML training algorithms for HMMs. By using the same transformation across a number of distributions and pooling the transformation training data maximum use is made of the adaptation data. This allows the parameters of all state distributions to be adapted. Results are presented on the 1000 word ARPA Resource Management RM1 database using a continuous density Gaussian mixture HMM system with cross-word triphone models.

2. ADAPTATION APPROACH

Each state in a continuous density Gaussian mixture HMM has an output distribution made up of a number of mixture component densities. A state with m mixture components can be expanded to m parallel single mixture component states. Thus the case of single mixture component states is described, and the extension to multiple mixture components is straightforward.

The probability density of state j generating a speech observation vector $\mathbf{o}$ of dimension n is

$$b_j(\mathbf{o}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{o}-\mu_j)'\Sigma_j^{-1}(\mathbf{o}-\mu_j)} \qquad (1)$$

where μ_j and Σ_j are the mean and covariance respectively of the output distribution of state j. The adaptation procedure is based on re-estimating the means of the state distributions using a linear transform of the existing mean. Thus it is assumed that in adapting from the SI system to the speaker adapted (SA) system the state transition probabilities and the covariances of the state distributions do not change.

The SI means are mapped to the unknown SD means ($\hat{\mu}_j$) by a linear regression transform estimated from the adaptation data:

$$\hat{\mu}_j = W_j \nu_j$$

where W_j is the $n \times (n+1)$ transformation matrix and ν_j is the extended mean vector:

$$\nu_j = [1, \mu_{j_1}, \ldots, \mu_{j_n}]'$$

If an individual regression matrix is used for each state using small amounts of adaptation data will result in very poor estimates of the matrices. Thus, each regression matrix is associated with many state distributions and estimated from the combined data. Tying transform matrices in this manner is similar in essence to the tying of states or mixtures [7] [8] which makes parameter estimates more robust. The process of tying many state distributions to one transform averages the transforms required for each component mean and produce a general transformation for all the tied components. This captures the general speaker characteristics for that class of sounds, and by tying the matrices in a manner such that distributions requiring similar transforms are associated with the same matrix, the regression transform can be effective even for states which are not seen in the adaptation data.

With complete tying, a single global regression matrix is used and associated with all distributions. At the opposite extreme, if a separate transform is used for each distribution in the system it can be shown that this is equivalent to performing a complete re-estimation of the model means using the adaptation data. In this case the problem of adapting the unseen distributions is not solved, and the best estimates for such parameters are from the SI system. Thus a compromise between the two extremes must be found such that the means of all distributions can be adapted well given the amount of adaptation data available.

3. REGRESSION TRANSFORM

The regression transformation is estimated using a maximum likelihood optimisation criteria which is consistent with standard HMM training methods.

Using the transform W_j for state j and again considering the case of a single Gaussian mixture component per state, the density function of state j in the adapted system is:

$$b_j(o) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{o}-W_j\nu_j)'\Sigma_j^{-1}(\mathbf{o}-W_j\nu_j)} \quad (2)$$

Given a set of adaptation speech data O consisting of T observation vectors ($O = \mathbf{o}_1 \ldots \mathbf{o}_T$) a maximum likelihood estimate of W_j can be found by iteratively maximising an auxiliary function $Q(\lambda, \bar{\lambda})$ [3].

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} f(O, \theta|\lambda) \log(f(O, \theta|\bar{\lambda})) \quad (3)$$

where λ represents the current parameter set, $\bar{\lambda}$ is the adapted parameter set, and $f(O, \theta|\lambda)$ is the likelihood of generating O using the state sequence $\theta(\theta = \theta_1 \ldots \theta_T)$ and the parameters λ. Θ is the set of all possible state sequences of length T.

Defining $\gamma_j(t)$ as the probability of occupying state j at time t while generating O using the current parameter set

$$\gamma_j(t) = \frac{\sum_{\theta \in \Theta} f(O, \theta_t = j|\lambda)}{\sum_{\theta \in \Theta} f(O, \theta|\lambda)} \quad (4)$$

maximising $Q(\lambda, \bar{\lambda})$ leads to the condition:

$$\sum_{t=1}^{T} \gamma_j(t)\Sigma_j \mathbf{o}_t \nu_j' = \sum_{t=1}^{T} \gamma_j(t)\Sigma_j W_j \nu_j \nu_j' \quad (5)$$

If the transform W_j is shared by R states $\{j_1, j_2 \ldots j_R\}$ the data for these states is combined and the condition becomes:

$$\sum_{t=1}^{T}\sum_{r=1}^{R} \gamma_{j_r}(t)\Sigma_{j_r}^{-1} \mathbf{o}_t \nu_{j_r}' = \sum_{t=1}^{T}\sum_{r=1}^{R} \gamma_{j_r}(t)\Sigma_{j_r}^{-1} W_j \nu_{j_r} \nu_{j_r}' \quad (6)$$

Assuming that all covariances are diagonal leads to a column by column estimation of W_j:

$$\mathbf{w}_i = G_i^{-1}\mathbf{z}_i \quad (7)$$

where $\mathbf{z}_i$ is the i^{th} column of the matrix Z produced by the left hand side of equation (6):

$$Z = \sum_{t=1}^{T}\sum_{r=1}^{R} \gamma_{j_r}(t)\Sigma_{j_r}^{-1} \mathbf{o}_t \nu_{j_r}' \quad (8)$$

and G_i is the sum over all tied mixture components of the outer product of the mean vectors scaled by the variance:

$$G_i = \sum_{r=1}^{R} c_{ii}^{(r)} \nu_{j_r} \nu_{j_r}' \quad (9)$$

with $c_{ii}^{(r)}$ being the i^{th} diagonal element of the r^{th} tied state mixture component covariance scaled by the total state occupation probability:

$$C^{(r)} = \sum_{t=1}^{T} \gamma_{j_r}(t)\Sigma_{j_r}^{-1} \quad (10)$$

A full derivation of this result is given in [5].

If exactly one state is occupied at each time frame and the covariances of all states tied to the same regression matrix are equal the regression matrix can be shown to be the least squares estimate [2]. If each state has a separate regression matrix the adaptation results in a Baum-Welch re-estimation of the means using the adaptation data as training data [1].

For the extension to mixture densities γ is the mixture component occupation probability.

Assuming the model sequence is known (i.e. the adaptation speech is labelled) the probabilities $\gamma_{j_r}(t)$ can be computed using a forward-backward alignment of the speech data. Thus all the necessary statistics can be gathered and the transform calculated.

4. EXPERIMENTAL SETUP

The ARPA Resource Management RM1 database was used to evaluate the speaker adaptation method. All speech was coded into frames consisting of a 39 component vector containing 12 MFCCs and normalised energy, plus first and second time derivatives.

A set of speaker independent models was trained on the speaker independent SI-109 portion of the database, using standard Baum-Welch maximum likelihood estimation. The model set consisted of tied state cross-word triphone models using a total of 1778 states, with 6 mixture components per state, plus a phrase initial silence model with 12 mixtures for each of 3 states. This system gives a 2.5% word error rate on the RM Feb'91 SI test set [8].

The speaker dependent portion of the database, consisting of data from 12 speakers, was used to evaluate the adaptation method. Adaptation data for each speaker was drawn from the training portion, and the adapted models tested on the 100 utterances in the test set using the standard word-pair grammar (perplexity 60). The average length of an utterance is 3.4 seconds.

The adaptation statistics were gathered using a forward-backward algorithm and computed as described in section 3. Only one iteration of the adaptation procedure was performed in all cases since the preliminary experiments showed that the state alignment in successive iterations was very similar.

5. EQUIVALENCE CLASSES

The regression equivalence classes for tying transforms were defined by using a between mixture component distance measure to place similar components into the same regression class. The assumption is that all components representing similar acoustic characteristics in the SI models should be adapted in the same manner for the new speaker.

The number of regression transforms is small in comparison to the large number of mixture components making the class allocations very broad. The experiments reported here investigate the variation of performance while changing the number of classes.

In the case of a global class all mixture components are tied to a single regression matrix. Further classes are obtained by clustering mixture components using a distance measure based on the overall class covariance when all mixture components are combined into a single distribution.

In the equivalence class definitions all components relating to states in the silence model were omitted. This is so that any phrase initial/phrase final silence does not dominate the transform calculations.

6. BASELINE SI/SD RESULTS

To give an idea of the comparative performance of SI model performance, and that achievable using a full speaker dependent (SD) system, a set of pseudo-SD models was trained for each speaker using all 600 SD training utterances in a Baum-Welch re-estimation using the SI models as seed models. The models are pseudo SD-models since the initial set-up and the state-clustering were based on the SI models. The small amount of training data available for each speaker leads to data insufficiency problems and tailoring the clustering to the available data may improve recognition rates.

The average word error rate over all speakers using the SI models was 4.3% while the SD models gave an average of 1.8% word error. Thus the error rate of the SI system is on average about 2.4 times that of the SD system. The speakers which perform particularly poorly using the SI models show a dramatic improvement using the SD models (e.g. for speaker rkm0_5 the SI error rate of 8.3% is reduced to 2.8% with the SD models). These results are used as a guide to judge the effectiveness of the adaptation method.

7. SUPERVISED ADAPTATION

A series of experiments first investigated adaptation using different amounts of data. The adaptation was performed in a static supervised mode using correctly labelled data with adaptation completed before any recognition tests were performed.

Initial tests used a global transform, where all mixture components (except those for silence) were tied to a single regression matrix.

Figure 1 shows the effect of adaptation on a small amount of data using the global transform. It can be seen that using as few as 3 utterances for adaptation results in an improvement over the SI models. Once this improvement is obtained, adding more adaptation data has a limited effect. With fewer than 3 utterances only a few mixture components are seen in the data and the estimation of the transform is weighted towards these components, resulting in poor transformation of the large number of unseen mixture components. As more data is added more varied mixture components are seen and the effect of dominant mixture components is reduced.

The effect of adaptation on the speakers recognised poorly by the SI system is the most significant (see Table 1). Adaptation on only 1 or 2 utterances has a detrimental effect, but using 15 adaptation utterances the error rate is significantly reduced over that of the SI system. On the better speakers the adaptation has a much smaller effect.

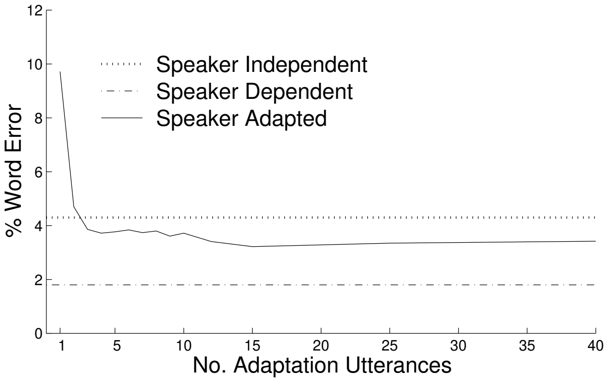

Figure 1. Adaptation on small amounts of data - global transform (average word error rate for all 12 speakers)

Speaker	SI	Num. Adapt. Utts			SD
		1	5	15	
bef0_3	3.2	8.3	2.9	2.7	2.3
cmr0_2	7.4	19.7	6.2	4.8	1.6
das1_2	1.8	2.5	1.6	1.8	0.9
dms0_4	3.2	4.0	2.7	2.6	1.0
dtb0_3	3.3	7.4	2.8	2.4	1.2
dtd0_5	4.6	5.8	4.3	4.0	2.3
ers0_7	3.5	13.7	3.3	3.5	2.6
hxs0_6	6.5	12.8	4.3	3.2	1.5
jws0_4	4.5	7.9	4.1	3.3	1.8
pgh0_1	2.6	5.7	2.6	2.5	2.1
rkm0_5	8.3	24.8	7.6	5.5	2.6
tab0_7	2.2	3.1	2.6	2.2	1.8
Average	4.3	9.2	3.8	3.2	1.8

Table 1. Supervised adaptation performance for individual speakers - global transform (% word error rate)

Using more data clearly requires more specific transforms to gain maximum information from the data. Figure 2 shows the adaptation performance using different numbers of classes when using 40 utterances for adaptation. The performance using a global matrix is similar to that obtained using only a small number of utterances. Using more classes results in reduced error rates until a performance threshold is reached when 15 classes are used. After this threshold it appears that the amount of data assigned to

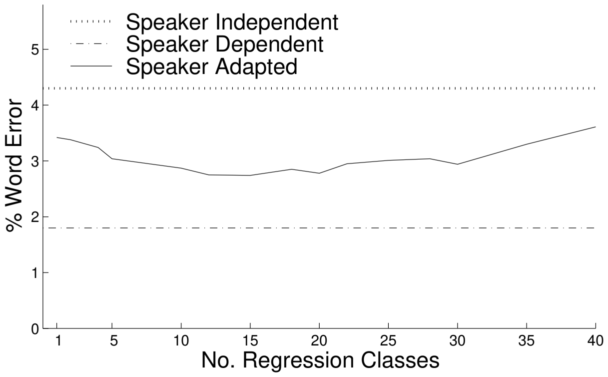

Figure 2. Effect of Number of Regression Classes using 40 utterances for adaptation(average % word error)

each regression matrix becomes too small and dominant mixture components again become a problem so performance degrades. This indicates that the number of classes should be tuned to the amount of data available.

8. UNSUPERVISED ADAPTATION

Unsupervised adaptation was implemented using a similar approach to the supervised adaptation, but used the initial SI models to perform recognition on the adaptation data to generate the speech labels. The forward-backward algorithm used the recognised labels to generate the adaptation statistics.

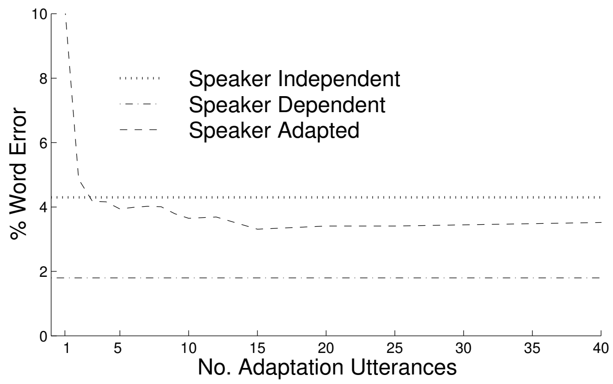

Figure 3. Unsupervised adaptation using a global transform (% Word error averaged over all speakers)

The results obtained using a small amount of adaptation data (Figure 3) show that unsupervised adaptation is almost as effective as the supervised implementation. This is partly due to the good performance of the SI models in labelling the adaptation data correctly. However, it is noticeable that even for those speakers which are poorly recognised by the SI system adaptation results in a significant error reduction.

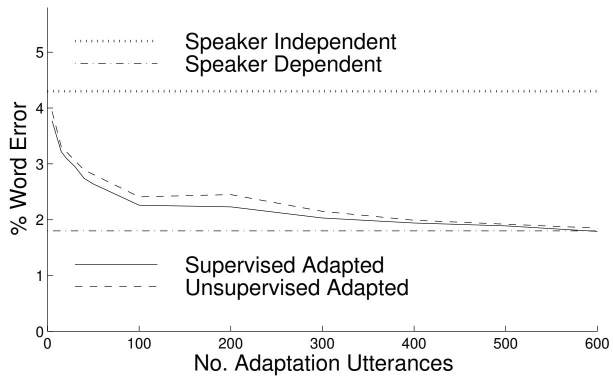

Figure 4. Best supervised/unsupervised adaptation results for available adaptation data (% word error)

It is clear from Figure 2 that choosing an appropriate number of regression classes for the available data has a significant effect on performance. Figure 4 compares supervised and unsupervised adaptation modes using different amounts of adaptation data and an appropriate number of classes (determined by experimentation). Performance improves as more data is used, and there is only a small difference between supervised and unsupervised adaptation. Table 2 shows the performance and number of classes used for different amounts of data.

Adaptation using all 600 utterances is comparable to the SD models indicating that the assumption of using a linear transform for generating mean adjustments is reasonable, and that tying similar mixture components to one transform is justified.

No. Utts	15	40	100	600
No. Classes	1	15	40	200
Supervised	3.2%	2.7%	2.3%	1.8%
Unsupervised	3.3%	2.9%	2.4%	1.9%

Table 2. Best performance for different amounts of adaptation data (% word error)

9. CONCLUSION

A method for adapting a speaker independent model system has been described and implemented. The method uses linear regression approach with a maximum likelihood objective function.

The approach is ideal for adaptation on small amounts of adaptation data and can be scaled to larger amounts of data with a corresponding improvement in performance. Once adaptation has been performed there is no extra computational requirement during recognition.

Results on the ARPA RM database have shown that a reduction in error over the initial SI system can be achieved using as few as 3 adaptation utterances (less than 11 seconds of speech), and with more adaptation data performance approaching speaker dependent systems is achieved.

Although only the mixture component means are adapted the method is still effective. Adaptation of the covariances or mixture component weights may give further error reductions.

Acknowledgement

C.J. Leggetter is funded by an EPSRC studentship.

REFERENCES

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

[2] F. Class, A. Kaltenmeier, et al. Fast Speaker Adaptation for Speech Recognition Systems. *Proc. ICASSP*, Vol. 1, pp. 133–136, 1990.

[3] B-H. Juang. Maximum Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains. *A.T. & T. Technical Journal*, Vol. 64, No. 6, pp. 1235–1249, July-August 1985.

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

[5] C.J. Leggetter and P.C. Woodland. Speaker Adaptation Using Linear Regression. *Technical Report* CUED/F-INFENG/TR.181, Cambridge University Engineering Department, June 1994.

[6] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. Large Vocabulary Continuous Speech Recognition Using HTK. *Proc. ICASSP*, Vol. 2, pp. 125–128, 1994.

[7] P.C. Woodland and S.J. Young. The HTK Tied-State Continuous Speech Recogniser. *Proc. EuroSpeech*, Vol. 3, pp. 2207–2210, Berlin, 1993.

[8] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-Based State Tying for High Accuracy Acoustic Modelling. *Proc. ARPA Human Language Technology Workshop*, Princeton, March 1994.