MULTIPLE-CLUSTER ADAPTIVE TRAINING SCHEMES

M.J.F. Gales

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
mjfg@eng.cam.ac.uk

ABSTRACT

This paper examines the training of multiple-cluster systems using adaptive training schemes. Various forms of transformation and canonical model are described in a consistent framework allowing re-estimation formulae for all cases to be simply derived. Initial experiments using these various schemes on a large vocabulary speech recognition task are presented. The initial experiments indicate that to achieve best performance when adapting these multiple-cluster systems requires the use of adaptive training schemes rather than using simpler cluster initialisation schemes.

1. INTRODUCTION

Adaptive training is a powerful training technique for building speech recognition systems on non-homogeneous data. This variability in the training data may result from the speaker changing, differing acoustic environments or varying channel conditions. The basic concept of adaptive training is to use one or more transformation to represent these speaker and environment differences. A canonical model is then trained, given the set of speaker/environment transforms. This canonical model should be more compact and amenable to being transformed to a new speaker, or acoustic condition, than standard speaker independent (SI) systems. Adaptive training schemes may be split into three broad classes. These are:

1. **Model independent**: these schemes do not make explicit use of any model information. The two most common forms are cepstral mean normalisation and variance normalisation [4]. These transforms are directly applied to the features.
2. **Feature transformation**: these transforms also act on the features but are derived, normally in an maximum likelihood (ML) fashion, using the current estimate of the model. Linear feature transformations may more generally be viewed as a constrained transformation of the model parameters [2]. Common versions of these feature transforms are vocal tract normalisation [7] and constrained MLLR [2]. Incorporating these schemes into adaptive training is simple as they require minimal changes to the standard re-estimation formulae [2].
3. **Model transformation**: the model parameters, means and possibly variances, are transformed. Common schemes are the original speaker adaptive training (SAT) [1], and cluster adaptive training (CAT) [3].

Normally the form of the canonical model for these schemes is the same as the SI model set.

This work was partially funded by the European Commission under the Language project Le-5 Coretex.

This paper examines the use of linear adaptive training schemes for general multiple cluster systems. All the schemes are model transformation adaptive training. The multiple cluster canonical model consists of C, the number of clusters, sets of means one for each component for each cluster, a single set of covariance matrices and a single set of weights and transition matrices. This paper describes ML re-estimation formulae for these multiple cluster systems, and the related transforms, within an adaptive training framework. In addition two possible initialisation schemes are described.

2. ADAPTIVE TRAINING SCHEMES

The general form of the adaptive training transformation in this paper is[1]

$$\hat{\boldsymbol{\mu}}^{(sm)} = \mathbf{W}^{(s)}\boldsymbol{\xi}^{(m)} \tag{1}$$

where $\mathbf{W}^{(s)}$ is a $n \times p$ transformation matrix for speaker s, $\boldsymbol{\xi}^{(m)}$ is $p \times 1$ vector of cluster means (n-dimensional data is assumed) and $\hat{\boldsymbol{\mu}}^{(sm)}$ is the adapted mean of component m for speaker s. The value of p will depend on the number of clusters, whether a bias cluster is being used, and whether a bias is used in the transform. The differences between the schemes are restrictions on the transformation matrix or number of clusters.

2.1. Standard SAT

In standard SAT the transformation of the mean is given by

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(sm)} &= \mathbf{A}^{(s)}\boldsymbol{\mu}^{(m)} + \mathbf{b}^{(s)} \\ &= \begin{bmatrix} \mathbf{A}^{(s)} & \mathbf{b}^{(s)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}^{(m)} \\ 1 \end{bmatrix} \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}^{(m)}$ is the unadapted mean of component m. This is a single cluster system ($C = 1$), so there are no additional model parameters than those in the SI system.

2.2. Cluster Adaptive Training

Though initially derived as an extension to speaker clustering, CAT, fits within the general adaptive training framework[2].

$$\hat{\boldsymbol{\mu}}^{(sm)} = \sum_c \lambda_c^{(s)} \boldsymbol{\mu}_{\mathbf{c}}^{(mc)} + \boldsymbol{\mu}_{\mathbf{b}}^{(m)}$$

[1] For the theoretical discussions in this paper only a single transform will be used. The extensions to multiple transforms are trivial and are described, along with schemes for determining assignment of components to classes, in previous papers (e.g. see [2]).

[2] This is the same expression as used for Eigenvoices [6].

$$= \begin{bmatrix} \lambda_1^{(s)}\mathbf{I} & \dots & \lambda_C^{(s)}\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{c}}^{(m1)} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{c}}^{(mC)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \end{bmatrix} \quad (3)$$

where $\boldsymbol{\mu}_{\mathbf{b}}^{(m)}$ and $\boldsymbol{\mu}_{\mathbf{c}}^{(mc)}$ are the bias cluster mean and mean for cluster c, respectively, of component m and $\boldsymbol{\lambda}^{(s)}$ is the vector of interpolation weights (or point in "eigenspace"). The complexity of the canonical model has increased. There are now $(C+1)$ sets of cluster means.

If additional restrictions are placed on the nature of the weights, $\lambda_i^{(s)} \in \{0,1\}$ and $\sum_i \lambda_i^{(s)} = 1$, CAT becomes a restricted form of speaker clustering where the variances of the clusters for the same component are constrained to be the same. For the specific case where $C = 2$ this is the equivalent of restricted gender dependent (GD) models. This is the form of GD models used in this paper.

2.3. Bias SAT

In [3] the concept of bias clusters with linear transformations was described. This may be expressed as

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(sm)} &= \mathbf{A}^{(s)}\boldsymbol{\mu}^{(m)} + \boldsymbol{\mu}_{\mathbf{b}}^{(m)} + \mathbf{b}^{(s)} \\ &= \begin{bmatrix} \mathbf{A}^{(s)} & \mathbf{I} & \mathbf{b}^{(s)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}^{(m)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \\ 1 \end{bmatrix} \quad (4) \end{aligned}$$

The canonical model is again a multiple cluster system as there is an additional set of component means, $\boldsymbol{\mu}_{\mathbf{b}}^{(m)}$. However the number of transform parameters to be estimated for a test speaker is identical to the standard SAT system. This scheme will be referred to as bias SAT (BSAT). BSAT has a simple intuitive interpretation as it allows the linear regressions to be based around component-specific points in space, rather than global points as defined by the transform bias, $\mathbf{b}^{(s)}$.

2.4. Extended SAT

Extended SAT (ESAT) may be seen as a generalisation of CAT. Each cluster is now transformed by an MLLR-like transform rather than a single weight.

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(sm)} &= \sum_c \mathbf{A}_c^{(s)}\boldsymbol{\mu}_{\mathbf{c}}^{(mc)} + \boldsymbol{\mu}_{\mathbf{b}}^{(m)} + \mathbf{b}^{(s)} \\ &= \begin{bmatrix} \mathbf{A}_1^{(s)} & \dots & \mathbf{A}_C^{(s)} & \mathbf{I} & \mathbf{b}^{(s)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{c}}^{(m1)} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{c}}^{(mC)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \\ 1 \end{bmatrix} \quad (5) \end{aligned}$$

Without the bias cluster this is the same as the cluster transformation in [5]. However in [5] the clusters were not embedded into an adaptive training scheme, nor was a bias cluster considered. This general form subsumes all the other transformation schemes.

3. PARAMETER ESTIMATION

As with the standard adaptive training scheme the parameters are estimated in an iterative fashion.

1. Initialise the multiple-cluster means (see section 3.4).
2. Estimate a transform, $\mathbf{W}^{(s)}$, for each of the training speakers, s.
3. Re-estimate the model parameters, $\mathcal{M}$, given the current estimate of the speaker transforms. For SAT, BSAT and ESAT the standard adaptive training formulae result in either the means or the variances being updated in a single iteration. To update both requires two iterations. In this work the means are first updated then the variances.
4. Repeat from 2 unless the convergence criterion is satisfied.

Since SAT systems are being used, during recognition an initial set of alignments are required from some standard (non-adaptively trained) model. For all systems the following procedure was used

1. Use an SI model set to obtain frame/component posteriors on the adaptation data ($\gamma_m(\tau)$). Using these alignments, estimate the transform(s).
2. Using the transform estimates, obtain new alignments and re-estimate the transform(s). This stage is them repeated as required.

For the experiments in this paper stage (2) was performed twice.

The next sections briefly describe the ML estimation of the transform parameters. Other than for CAT, a modification to the standard MLLR training described in [8] is used. The training of the canonical models is a modification to the standard adaptive training scheme described in [1]. For details of the CAT canonical model and transform estimation schemes, see [3].

3.1. Transform Estimation

The estimation of the transform parameters is an extension of the standard MLLR parameter estimation. The schemes may be split into two sets, those that use a bias cluster and those that don't. Row i of the "bias-free" system transform, $\mathbf{W}^{(s)}$, $\mathbf{w}_i^{(s)}$, is estimated using

$$\mathbf{w}_i^{(s)T} = \mathbf{G}^{(i)-1}\mathbf{k}^{(i)T} \quad (6)$$

where

$$\mathbf{W}^{(s)} = \begin{bmatrix} \mathbf{A}_1^{(s)} & \dots & \mathbf{A}_C^{(s)} & \mathbf{b}^{(s)} \end{bmatrix} \quad (7)$$

$$\mathbf{G}^{(i)} = \sum_m \frac{1}{\sigma_i^{(m)2}} \boldsymbol{\xi}^{(m)}\boldsymbol{\xi}^{(m)T} \sum_\tau \gamma_m(\tau) \quad (8)$$

$$\mathbf{k}^{(i)} = \sum_m \sum_\tau \gamma_m(\tau) \frac{1}{\sigma_i^{(m)2}} o_i(\tau)\boldsymbol{\xi}^{(m)T} \quad (9)$$

$$\boldsymbol{\xi}^{(m)} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{c}}^{(m1)} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{c}}^{(mC)} \\ 1 \end{bmatrix} \quad (10)$$

$\sigma_i^{(m)2}$ is the variance of element i of component m and $\gamma_m(\tau)$ is the posterior probability of component m generating the observation at time τ. For the schemes where a bias cluster is used the observation, $\mathbf{o}(\tau)$ in equation 9 is replaced by $(\mathbf{o}(\tau) - \boldsymbol{\mu}_{\mathbf{b}}^{(m)})$.

Irrespective of the form of the transform the same sufficient statistics are required. These are simply the occupancy counts and observation vector sum for each component.

3.2. Model Estimation

The model parameter estimation is performed in two stages. For the case where there the transform has no bias, $\mathbf{b}^{(s)}$,

$$\boldsymbol{\xi}^{(m)\prime} = \left(\sum_{s,\tau} \gamma_m(\tau) \mathbf{W}^{(s)\prime T} \boldsymbol{\Sigma}^{(m)-1} \mathbf{W}^{(s)\prime} \right)^{-1} \sum_{s,\tau} \gamma_m(\tau) \mathbf{W}^{(s)\prime T} \boldsymbol{\Sigma}^{(m)-1} \left(\mathbf{o}(\tau) - \mathbf{b}^{(s)} \right) \quad (11)$$

where

$$\mathbf{W}^{(s)\prime} = \left[\begin{array}{cccc} \mathbf{A}_1^{(s)} & \dots & \mathbf{A}_C^{(s)} & \mathbf{I} \end{array} \right] \quad (12)$$

and

$$\boldsymbol{\xi}^{(m)\prime} = \left[\begin{array}{c} \boldsymbol{\mu}_{\mathbf{c}}^{(m1)} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{c}}^{(mC)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \end{array} \right] \quad (13)$$

For the case of no bias transform $\mathbf{b}^{(s)}$ may simply be set to $\mathbf{0}$. The estimation of the variances is identical to standard SAT [1].

3.3. Number of Parameters

Parms.	GD	CAT	SAT	BSAT	ESAT
Mean (p)	$2n$	$Cn+n$	n	$2n$	$Cn+n$
Trans.	1	C	n^2+n	n^2+n	Cn^2+n

Table 1: Number of mean parameters per cluster and number of transform parameters per speaker.

Table 1 shows the number of mean parameters per component[3] for each of the systems. This determine the memory requirements to store the models and whether robust estimates of the model parameters may be obtained[4]. An important issue in training the models is the number of parameters required to be stored (and accumulated) at each time instance. The primary cost is the estimation of the mean parameter, which for a $p \times n$ transform is dominated by $p(p+1)/2$ parameters to be stored per component.

Table 1 also shows the number of transform parameters for each of the systems. This is important as the fewer the transform parameters to be estimated, the more rapid the adaptation scheme. Thus GD and CAT are the most appropriate for small amounts of adaptation data, ESAT the least appropriate.

3.4. Cluster Initialisation

Since the training scheme described is an iterative one it is necessary to initialise the multiple clusters in some fashion. There are two basic initialisation schemes that may be used to initialise the clusters.

1. **Eigenvoices**: this is based on the eigenspace approach described in [6]. However in order to build the complex model sets required for state-of-the-art systems, the modified scheme described in [3] is required. This form of initialisation generates a bias cluster.
2. **Speaker clustering**: here each speaker is assigned in a hard fashion to one of the available speaker clusters. The simplest approach to this is to initialise with gender dependent models. The weights are either zero or one ($\lambda_i^{(s)} \in \{0,1\}$ and $\sum_i \lambda_i^{(s)} = 1$).

For the work presented here where a bias cluster is used the models were initialised using the Eigenvoices scheme. For those with no bias cluster, gender dependent clusters were used.

4. RESULTS

The baseline system used for the recognition task was a gender-independent cross-word-triphone mixture-Gaussian tied-state HMM system. This was the same as the "HMM-1" model set used in the HTK 1994 ARPA evaluation system [9]. The speech was parameterised into 12 MFCCs, C_1 to C_{12}, along with normalised log-energy and the first and second differentials of these parameters. This yielded a 39-dimensional feature vector, to which cepstral mean normalisation was applied. The acoustic training data consisted of 36493 sentences from the SI-284 WSJ0 and WSJ1 sets, and the LIMSI 1993 WSJ lexicon and phone set were used. The standard HTK system was trained using decision-tree-based state clustering to define 6399 speech states. The number of components per state was 12 for the speech state and 24 for the "silence" states. For the H1 task a 65k word list and dictionary was used with the trigram language model described in [9]. All decoding used a dynamic-network decoder.

For the adaptive training schemes two transform classes were used during training. One was used for the speech components, the second for the "silence" components. In all cases four complete iterations of adaptive training were used[5]. During testing, unless otherwise stated, two transform classes (as previously described) were used. For all the multi-cluster systems a simple 2 cluster system was used. The ESAT system did not have a bias cluster associated with it and $C = 2$ (since $C = 1$ with a bias cluster is BSAT), and CAT used a bias cluster ($C = 1$). All the test speaker adaptation was performed using supervised adaptation with 40 sentences.

System	Adapt	Error Rate (%)		
		H1 Dev	H1 Eval	Average
SI	—	9.38	10.01	9.71
SI	MLLR	9.04	10.05	9.57
SAT	MLLR	8.99	9.95	9.49
BSAT	BSAT	8.89	9.58	9.26

Table 2: Baseline and diagonal transforms

First some experiments on the possible advantages of using BSAT were performed. The results using diagonal transforms are

[3] This excludes the transition probabilities, component weights and variances, as they will contribute the same number of parameters for each of the schemes.

[4] It is possible to have more compact representations of the clusters using schemes like transform-based CAT [3].

[5] This results in a slight inconsistency between SAT, BSAT and ESAT and the CAT and GD models because the later two can update all the model parameters in a single iteration.

shown in table 2[6]. From the results there is a small gain in using a BSAT system over the SAT system with simple transforms, though the gain is not statistically significant at the 95% confidence level.

System	Adapt	Error Rate (%)		
		H1 Dev	H1 Eval	Average
SI	—	9.38	10.01	9.71
	MLLR	8.24	9.24	8.77
SAT	MLLR	7.62	8.36	8.01
GD	GD	8.99	9.39	9.20
	ESAT	8.12	8.61	8.28
CAT	CAT	8.84	9.13	8.99
	ESAT	7.77	8.60	8.20
BSAT	BSAT	7.63	8.23	7.95
	ESAT	7.51	8.33	7.94
ESAT	ESAT	7.16	8.25	7.74

Table 3: Block-diagonal transforms

Table 3 shows the performance of a series of adaptively trained systems using block diagonal transforms n(static, delta and delta-delta blocks). As expected, given the amount of adaptation data, the simpler transformations, CAT and GD, performed worse than the more complex ones, BSAT and ESAT. The performance of the CAT system was not significantly, at the 95% level, better than the GD system's performance (though significant gains have been observed on other tasks and with more clusters). Comparing the GD and ESAT systems, which were initialised in the same fashion using the more complex ESAT transform, indicates that the appropriate training of the clusters is important. The ESAT system was significantly, at the 95% level, better, than the GD system reducing the word error rate by around 7% relative. Again the performance difference between the SAT and BSAT systems was not statistically significant.

System	Adapt	Error Rate (%)		
		H1 Dev	H1 Eval	Average
SI	MLLR	7.96	8.62	8.31
SAT	MLLR	7.28	8.18	7.75
BSAT	BSAT	7.53	8.17	7.86
ESAT	ESAT	7.02	8.13	7.61

Table 4: Multiple block-diagonal transforms

Rather than using two block-diagonal transforms, the number of transforms was determined using a regression class tree. The SAT system performs marginally (though not significantly) better than the BSAT system[7]. The difference was reduced when the same BSAT transform as used in training was used to obtain a set of means, and then standard MLLR was performed on those adapted, single set, means[8]. In this case the average error rate was 7.66%. Again the FSAT gave slightly (though not significantly) better performance than any other adaptive training scheme. Both adaptively trained systems were significantly ,at the 95% level, better than the SI adapted system.

[6] The type of adaptation scheme will be labelled according to the adaptive training scheme as described in section 2 if it differs from standard MLLR.

[7] The minimum occupancy counts for the regression class trees was set independently for each transform type. As a result ESAT had far fewer transforms than the standard MLLR, or BSAT, systems

[8] This increases the number of updates for the transform to 4. Increasing SAT and ESAT transform iterations to 4 yielded slight gains in both cases to 7.67% and 7.53% respectively.

5. CONCLUSIONS

This paper has described various multiple-cluster adaptive training schemes. Re-estimation formulae for both the canonical models and the multiple cluster transforms have been detailed. The performance on a large vocabulary speaker independent task of some two cluster systems were compared. For cases where complex transforms, such as ESAT are to be used then it is important to use adaptive training schemes to obtain the clusters, rather than using, for example, gender dependent models.

Currently these schemes have not been combined with more complex model independent adaptive training schemes, such as variance normalisation, and other non-linear feature transformation schemes such as VTN. However, both these schemes are used in current state-of-the-art systems for the switchboard task. Further work will examine the interaction of these multiple-cluster adaptive training schemes with other adaptive training schemes.

6. REFERENCES

[1] T Anastasakos, J McDonough, R Schwartz, and J Makhoul. A compact model for speaker-adaptive training. In *Proceedings ICSLP*, pages 1137–1140, 1996.

[2] M J F Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[3] M J F Gales. Cluster adaptive training of hidden markov models. *IEEE Transactions Speech and Audio Processing*, 8:417–428, 2000.

[4] T Hain, P C Woodland, T R Niesler, and E W D Whittaker. The 1998 HTK system for transcription of conversational telephone speech. In *Proceedings ICASSP*, pages 57–60, 1999.

[5] J Huang and M Padmanabhan. A study of adaptation techiques on a voicemail transcription task. In *Proceedings Eurospeech*, pages 13–16, 1999.

[6] R Kuhn, P Nguyen, J-C Junqua, L Goldwasser, N Niedzielski, S Fincke, K Field, and M Contolini. Eigenvoices for speaker adaptation. In *Proceedings ICSLP*, pages 1771–1774, 1998.

[7] L Lee and R C Rose. Speaker normalisation using efficient frequency warping procedures. In *Proceedings ICASSP*, volume 1, pages 353–356, 1996.

[8] C J Leggetter and P C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language*, 9:171–186, 1995.

[9] P C Woodland, J J Odell, V Valtchev, and S J Young. The development of the 1994 HTK large vocabulary speech recognition system. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 104–109, 1995.