ADAPTIVE TRAINING USING STRUCTURED TRANSFORMS

K. Yu and M.J.F. Gales

Engineering Department, Cambridge University
Trumpington St. Cambridge, CB2 1PZ, U.K.
{ky219,mjfg}@eng.cam.ac.uk

ABSTRACT

Adaptive training is an important approach to train speech recognition systems on *found*, non-homogeneous, data. Standard approach employs a single transform to represent unwanted acoustic variability. However, for found data there are commonly multiple acoustic factors affecting the speech signal. This paper investigates the use of multiple forms of transformations, structured transforms (ST), to represent the complex non-speech variabilities in an adaptive training framework. Two forms of transformations are considered, cluster mean interpolation and constrained MLLR, consequently, the canonical model here is a multi-cluster HMM model. Both ML and minimum phone error (MPE) re-estimation formulae for the canonical model, are presented. This multi-cluster MPE training is also applicable to eigenvoice systems. Experiments to compare ST to standard adaptive training schemes were performed on a conversational telephone speech task. ST were found to significantly reduce the word error rate.

1. INTRODUCTION

The majority of state-of-the-art speech recognition systems are trained on *found* data, for example broadcast news and telephone conversations. This data is typically highly non-homogeneous, there are multiple factors that vary across the corpus that alter the speech signal. For example, the speaker or the background acoustic noise condition changes across training utterances. *Adaptive* training techniques [1, 2] aim to overcome this problem by using transformations to represent the unwanted acoustic variability. A *canonical* model can then be trained, given these transforms, which should only represent the desired variability of a particular phone without the effects of the unwanted acoustic factors. Commonly used approaches are based on maximum likelihood linear regression (MLLR) and constrained MLLR (CMLLR) transformations, referred to as speaker adaptive training (SAT) [1], or cluster mean interpolation, referred to as cluster adaptive training (CAT) [2] or eigenvoices [3]. These adaptive training schemes use a single form of transform during training. Multiple forms of transform may then be used during test set adaptation [2]. More recently schemes using multiple forms of transformation during the adaptive training process have been examined [4, 5].

This paper considers the use of multiple transformations, referred to as *structured transforms*, for adaptive training. The use of multiple types of transformation should help in removing the effects of unwanted acoustic factors. The use of these structured transforms may be viewed as an initial step to constructing large systems using acoustic factorisation [5]. In acoustic factorisation multiple transformations are used. However, each transform is constrained to relate to a specific acoustic factor. This gives additional flexibility in how the model may be used [5]. This restriction is not applied to the structured transformations used here.

In this work two transformations are used for the structured transforms. The transformations selected are cluster mean interpolation [2] and CMLLR [6]. For these forms of transformation both maximum likelihood (ML) and discriminative training, in this case minimum phone error (MPE) [10], are considered. By defining appropriate speaker-level smoothing function and prior distribution, standard MPE training can be extended for training multi-cluster HMM model. A simplified form of the discriminative training of the canonical model may also be used to discriminatively train an eigenvoice system [3]. In common with other combinations of adaptive training with discriminative training, the transformation parameters are estimated using ML. These are then fixed and used during discriminative training. This dramatically simplifies the estimation of test set transforms and has been found to show the same performance as combined discriminative schemes for the unsupervised adaptation tasks considered here [7]. In contrast to previous work on multiple transform schemes, a state-of-the-art task, conversational telephone speech, is examined.

The paper is arranged as follows. In section 2, the theory of adaptive training using structured transforms with both the ML and MPE criteria is introduced. Section 3 details experimental results on an English conversational telephone speech task. Conclusions are then given in section 4.

2. ADAPTIVE TRAINING USING STRUCTURED TRANSFORMS

The structured transforms considered in this work comprise a combination of CMLLR and CAT. Though both are linear, the representation of non-speech variability is very different. CMLLR is a linear transform of the features, CAT is a linear interpolation of a set of cluster means.

2.1. Maximum Likelihood Training

Maximum likelihood training of the model parameters uses expectation maximisation in the same fashion as CAT [2] and SAT [1]. An iterative approach is adopted where first the transform parameters are estimated, then the canonical model parameters. The whole process is then repeated. Note, the canonical model to be trained has multiple cluster, in contrast to the standard SAT set-up.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

The auxiliary function of the adaptive training with ST is[1]

$$\mathcal{Q}(\mathcal{M}) = -\frac{1}{2}\sum_{m,t,s}\gamma_m(t)\Big(\log|\boldsymbol{\Sigma}^{(m)}| - \log(|\mathbf{A}^{(s)}|^2) \qquad (1)$$
$$+ (\mathbf{o}^{(s)}(t)-\boldsymbol{\mu}^{(sm)})^T\boldsymbol{\Sigma}^{(m)-1}(\mathbf{o}^{(s)}(t)-\boldsymbol{\mu}^{(sm)})\Big)$$

where $\gamma_m(t)$ is the posterior probability of component m generating the observation $\mathbf{o}(t)$ given the current model parameters $\hat{\mathcal{M}}$, the CMLLR feature transform gives

$$\mathbf{o}^{(s)}(t) = \mathbf{A}^{(s)}\mathbf{o}(t) + \mathbf{b}^{(\mathbf{s})} \qquad (2)$$

the CAT interpolation of the means gives

$$\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)} \qquad (3)$$

where $\boldsymbol{\lambda}^{(s)}$ are the set of interpolation weights for speaker s and $\mathbf{M}^{(m)}$ consists of P cluster mean vectors.

$$\mathbf{M}^{(m)} = \left[\boldsymbol{\mu}_1^{(m)}, \cdots, \boldsymbol{\mu}_P^{(m)}\right] \qquad (4)$$

The model parameters may be split into two distinct parts. The first are the parameters of the canonical model[2] $\{\mathbf{M}^{(m)}, \boldsymbol{\Sigma}^{(m)}\}$ for each component m. Second are the parameters associated with the transform for speaker s, $\{\mathbf{A}^{(s)}, \mathbf{b}^{(s)}, \boldsymbol{\lambda}^{(s)}\}$.

Canonical model estimation. This is a simple extension to the model-based CAT estimation approach [2]. In addition to considering the speaker specific interpolation weights, the features are transformed using the associated CMLLR transform. The sufficient statistics required to estimate the model parameters are

$$\mathbf{G}^{(m)} = \sum_{s,t}\gamma_m(t)\hat{\boldsymbol{\lambda}}^{(s)}\hat{\boldsymbol{\lambda}}^{(s)T} \qquad (5)$$
$$\mathbf{K}^{(m)} = \sum_{s,t}\gamma_m(t)\hat{\boldsymbol{\lambda}}^{(s)}\mathbf{o}^{(s)}(t)^T \qquad (6)$$
$$\mathbf{L}^{(m)} = \sum_{s,t}\gamma_m(t)\mathbf{o}^{(s)}(t)\mathbf{o}^{(s)}(t)^T \qquad (7)$$

Where the transformed features are determined by the speaker specific CMLLR transform in equation 2. For all the systems in this paper, diagonal covariance matrices are used. The ML-estimates of the model parameters are then given by

$$\mathbf{M}^{(m)T} = \mathbf{G}^{(m)-1}\mathbf{K}^{(m)} \qquad (8)$$
$$\boldsymbol{\Sigma}^{(m)} = \text{diag}\left(\frac{\mathbf{L}^{(m)} - \mathbf{M}^{(m)}\mathbf{K}^{(m)}}{\sum_{s,t}\gamma_m(t)}\right) \qquad (9)$$

Note the ML-based eigenvoices formulae are a simplified form of this where the covariance matrices are not updated [8].

Transform estimation. This is a simple iterative process, where given the interpolation weights, the adapted mean, $\boldsymbol{\mu}^{(sm)}$ is used to estimate the CMLLR transform as described in [6]. Then the interpolation weights, $\boldsymbol{\lambda}^{(s)}$ are estimated using the transformed features $\mathbf{o}^{(s)}(t)$ as described in [2]. The initialisation of interpolation weights can be found in [2]. As the standard ML estimates of these transforms are used in this paper, and are not involved in the discriminative training, they are not described in more detail.

The parameters associated with the structured transforms are not considered in the next section. Thus the parameters associated with $\mathcal{M}$ will simply be the canonical model parameters.

2.2. Discriminative Training

For state-of-the-art speech recognition systems, discriminative training is becoming increasingly popular [9]. Various criteria are possible, for example MMI training, however minimum phone error (MPE) training [10] has been found to yield good performance. The criterion may be expressed as

$$\mathcal{F}(\mathcal{M}) = \frac{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)\text{RawAccuracy}(w)}{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)} \qquad (10)$$

where $\text{RawAccuracy}(w)$ is a measure of the number of phones accurately transcribed, $\mathcal{M}_w$ is the composite model for word sequence w and κ is an acoustic deweighting factor commonly used in discriminative training.

To optimise the MPE criterion, a *weak-sense* auxiliary function is used to derive close-form re-estimation formulae [10]. The weak sense auxiliary function for MPE can be expressed as

$$\mathcal{Q}(\mathcal{M}) = \mathcal{Q}^n(\mathcal{M}) - \mathcal{Q}^d(\mathcal{M}) + \mathcal{G}(\mathcal{M}) + \log p(\mathcal{M}) \qquad (11)$$

where $\mathcal{Q}^n(\mathcal{M})$ and $\mathcal{Q}^d(\mathcal{M})$ are standard auxiliary functions with a similar form to equation 1 for numerator and denominator respectively. The only difference is that the equivalent of the posterior for components, $\gamma_m^n(t)$ in $\mathcal{Q}^n(\mathcal{M})$ and $\gamma_m^d(t)$ in $\mathcal{Q}^d(\mathcal{M})$, are calculated in a different way [10]. A smoothing function, $\mathcal{G}(\mathcal{M})$ is added to improve stability of the optimisation. This must satisfy the following equation to ensure that it is still a valid weak sense function

$$\left.\frac{\partial}{\partial\mathcal{M}}\mathcal{G}(\mathcal{M})\right|_{\hat{\mathcal{M}}} = 0 \qquad (12)$$

Finally, a prior , $p(\mathcal{M})$, may also be introduced, either based on the ML statistics, which is called I-smoothing [9], or on the maximum a posteriori (MAP) estimates, which is called MPE-MAP [11]. By definition, a log-prior is a weak-sense function of itself, so that equation 11 is a valid weak-sense auxiliary function.

In common with many implementations of discriminative adaptive training schemes, this work will only consider discriminatively training the model parameters given ML estimates of all transform parameters. This yields about the same performance as discriminatively training all the parameters, but is simpler and more consistent when dealing with unsupervised adaptation task [7]. The combination of CMLLR with discriminatively training the model parameters is simple as it is a transformation of the features. The rest of this section concentrates on the discriminative training of the CAT model parameters. It is worth noting that this also allows discriminative training of eigenvoices when a maximum likelihood eigenspace is used, as the schemes are the same other than the initialisation [2].

The smoothing auxiliary function $\mathcal{G}(\mathcal{M})$ is different from the standard form given in [10] as it must yield the current CAT parameters as the ML estimate. One suitable smoothing function is

[1] The dependence on the current model parameters$\hat{\mathcal{M}}$ will be assumed in the following expressions.

[2] For this paper the estimation of the component priors and transition matrices are not considered. The formulae are identical to the standard CAT updates given in [2].

given by[3]

$$\mathcal{G}(\mathcal{M}) = -\sum_{m,s} \frac{D_m \nu_m^{(s)}}{2} \Big(\log |\boldsymbol{\Sigma}^{(m)}| \tag{13}$$
$$+\mathrm{tr}((\hat{\boldsymbol{\Sigma}}^{(m)} + \hat{\boldsymbol{\mu}}^{(sm)} \hat{\boldsymbol{\mu}}^{(sm)T}) \boldsymbol{\Sigma}^{(m)-1})$$
$$+\hat{\boldsymbol{\lambda}}^{(s)T} \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} (\mathbf{M}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)} - 2\hat{\boldsymbol{\mu}}^{(sm)}) \Big)$$

where $\hat{\boldsymbol{\mu}}^{(sm)} = \hat{\mathbf{M}}^{(m)} \hat{\boldsymbol{\lambda}}^{(m)}$, $\hat{\boldsymbol{\Sigma}}^{(m)}$ are current parameters and "tr()" is trace of matrix. The constant D_m is a positive smoothing constant for component m to ensure convergence. This expression satisfies the smoothing constraint of equation 12, for all values of $\nu_m^{(s)}$. However it is sensible to use this value to reflect the proportions of data for that particular components of a speaker, so in this work

$$\nu_m^{(s)} = \frac{\sum_t \gamma_m^n(t)}{\sum_{s,t} \gamma_m^n(t)} \tag{14}$$

where the summation in the numerator only involves data associated with speaker s.

For MPE training it is essential to perform some additional smoothing to improve the generalisation of the resultant model. This is normally achieved by incorporating a prior into the estimation scheme. For this work the prior distribution, $p(\mathcal{M})$ will be based on the ML estimates of the the cluster means, $\tilde{\mathbf{M}}^{(m)}$, and covariance matrix, $\tilde{\boldsymbol{\Sigma}}^{(m)}$. This is an I-smoothing version of discriminative CAT training. By taking the Normal-Wishart distribution [12] as the prior for model parameters $\{\mathbf{M}^{(m)}, \boldsymbol{\Sigma}^{(m)}\}$ at speaker level and assuming appropriate Normal-Wishart parameters, the log prior for model parameters may be written as a weighted sum of speaker-level priors:

$$\log p(\mathcal{M}) = K - \frac{\tau^I}{2} \sum_{s,m} \tilde{\nu}_m^{(s)} \Big(\log |\boldsymbol{\Sigma}^{(m)}| + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^{(m)} \boldsymbol{\Sigma}^{(m)-1}) +$$
$$(\mathbf{M}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)} - \tilde{\mathbf{M}}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)})^T \boldsymbol{\Sigma}^{(m)-1} (\mathbf{M}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)} - \tilde{\mathbf{M}}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)}) \Big)$$

where τ^I is the specified parameter of the Normal-Wishart distribution. $\tilde{\nu}_m^{(s)}$ is a slightly modified version of equation 14, instead of using the MPE numerator values $\gamma_m^n(t)$, the standard ML parameters, $\gamma_m(t)$, are used. K is the appropriate normalisation term. The ML-estimates[4] are derived from the ML statistics, $\tilde{\mathbf{G}}^{(m)}$, $\tilde{\mathbf{K}}^{(m)}$ and $\tilde{\mathbf{L}}^{(m)}$. However in contrast to equations 5 to 7, the statistics are all normalised by $\sum_{s,t} \gamma_m(t)$ to yield "unit" counts.

Differentiating the whole auxiliary function with respect to the canonical model parameters and setting it to zero leads to model parameters re-estimation formulae. These updates may be expressed in terms of modified sufficient statistics.

$$\mathbf{G}^{(m)} = \sum_{s,t} \gamma_m^{mpe}(t) \hat{\boldsymbol{\lambda}}^{(s)} \hat{\boldsymbol{\lambda}}^{(s)T} + D_m \mathbf{G}_D^{(m)} + \tau^I \tilde{\mathbf{G}}^{(m)}$$
$$\mathbf{K}^{(m)} = \sum_{s,t} \gamma_m^{mpe}(t) \hat{\boldsymbol{\lambda}}^{(s)} \mathbf{o}^{(s)}(t)^T + D_m \mathbf{K}_D^{(m)} + \tau^I \tilde{\mathbf{K}}^{(m)}$$
$$\mathbf{L}^{(m)} = \sum_{s,t} \gamma_m^{mpe}(t) \mathbf{o}^{(s)}(t) \mathbf{o}^{(s)}(t)^T + D_m \mathbf{L}_D^{(m)} + \tau^I \tilde{\mathbf{L}}^{(m)}$$

where $\gamma_m^{mpe}(t) = \gamma_m^n(t) - \gamma_m^d(t)$, and

$$\mathbf{G}_D^{(m)} = \sum_s \nu_m^{(s)} \hat{\boldsymbol{\lambda}}^{(s)} \hat{\boldsymbol{\lambda}}^{(s)T} \tag{15}$$
$$\mathbf{K}_D^{(m)} = \mathbf{G}_D^{(m)} \hat{\mathbf{M}}^{(m)T} \tag{16}$$
$$\mathbf{L}_D^{(m)} = \hat{\boldsymbol{\Sigma}}^{(m)} + \hat{\mathbf{M}}^{(m)} \mathbf{G}_D^{(m)} \hat{\mathbf{M}}^{(m)T} \tag{17}$$

The canonical model parameters are then given by

$$\mathbf{M}^{(m)T} = \mathbf{G}^{(m)-1} \mathbf{K}^{(m)} \tag{18}$$
$$\boldsymbol{\Sigma}^{(m)} = \mathrm{diag}\left(\frac{\mathbf{L}^{(m)} - \mathbf{M}^{(m)} \mathbf{K}^{(m)}}{\sum_{s,t} \gamma_m^{mpe}(t) + D_m + \tau^I} \right) \tag{19}$$

In common with standard I-smoothed MPE training, there are two constants that must be specified. The first is D_m. In this paper it is set to be $E\gamma_m^{den}$, where E is a constant [10]. The second constant is τ^I which determines the smoothing with the ML estimates of the model parameters.

The MPE criterion has been investigated for traditional adaptive training[7]. In this paper, we investigated a simplified discriminative adaptive training with structured transforms. First, multi-cluster canonical model and structured transforms are estimated using ML criterion as described in section 2.1. Then both CAT weights and CMLLR transforms are fixed and only the canonical model is updated with the MPE criterion. This update employs the formulae above. The component posterior probability for numerator, denominator and ML estimates are obtained by using the adapted model and the accumulation of sufficient statistics uses the transformed feature vectors.

3. RESULTS

The performance of structured adaptive training was evaluated on a state-of-the-art large vocabulary speech recognition system, conversational telephone speech (Switchboard). The training corpus consisted of 5446 speakers (2747 female, 2699 male), giving a total of about 295 hours of data. This is referred to as the `h5train03` training data. The test corpus was a subset, half, of the `dev01` test data consisting of 59 speakers (30 female, 29 male), about 3 hours. This is the `dev01sub` test data. All systems had 16 Gaussians per state, and use PLP front-end with C0 and first, second derivatives, HLDA and VTLN were also applied. The use of VTLN decreased the possible gains that could be obtained using adaptive training, but gave a more realistic baseline. A tri-gram language model was used in decoding.

For each form of system both MLE and MPE training were used. The simplified form of adaptive training with MPE was used where the structured transforms were estimated using MLE and only the canonical model parameters estimated using MPE. The hypothesis for adaptation for all the adaptively trained systems was taken from the associated MPE or MLE gender independent (GI) system. For the gender dependent (GD) systems, the test set per-side "gender" was assumed to be known, i.e., no gender classification error. During training and test adaptation, a global interpolation weight was estimated and separate speech and silence transforms were used for CMLLR. Two multiple cluster systems were built, both used two clusters. The first was a standard CAT system. The second used structured transforms (STs), mean interpolation and CMLLR, for adaptive training. These systems were

[3] The interpolation weights are not updated so are simply indicated as using the current parameters throughout.

[4] Note for MPE training these statistics differ from the numerator statistics. For MMI estimation they would be the same and I-smoothing can be implemented using count scaling [10].

initialised using gender information during training[5], however, in test adaptation, no prior gender information was used.

System	Training	Test	Estimation	
	Adaptation	Adaptation	MLE	MPE
GI	—	—	33.4	30.4
		CMLLR	31.5	28.3
GD	gender info	—	32.7	30.3
		CMLLR	30.9	28.4
GD (MPE-MAP)	gender info	—	—	29.7
		CMLLR	—	27.8
SAT	CMLLR	CMLLR	31.0	27.8
CAT	CAT	CAT	32.6	29.6
		ST	30.8	28.0
ST	ST	ST	30.6	27.5

Table 1. WER on `dev01sub` comparing different adaptive training with both MLE and MPE training. ST refers to the structured transform of CAT plus CMLLR.

Table 1 shows the baseline results for both GI and GD models. For the MLE trained systems, the GD model yielded significant gains over the GI model both with and without CMLLR adaptation. However, for MPE training, the performance of the two systems was approximately the same. Using MPE-MAP, significant performance gains over the MPE GD system were obtained, about 0.6% absolute. This is consistent with the gains that were obtained on the Broadcast News task [11].

Three adaptively trained systems were generated. The first used CMLLR during the training process, consistent with the forms used in the CUED-HTK evaluation systems[6]. This is referred as speaker adaptive training (SAT). For the MLE system, gains of about 0.5% absolute were obtained over the GI system and about the same performance as the GD system. For the MPE trained systems, the performance was comparable with the MPE-MAP system. The CAT system using just interpolation weights for test adaptation (CAT) gave about the same performance as the GD system in MLE training. Using ST for test adaptation again yielded about the same as a GD system with CMLLR. In MPE training, it is interesting to note that the CAT system performed well without the need to use MPE-MAP. This is due to the use of "soft" interpolation weights allowing clusters to make use of training data from all speakers. Using ST in training obtained statistically significant[7] improvements compared to SAT in both MLE and MPE training. Compared to the other adaptive training or adaptation techniques in both ML and MPE, ST gave the lowest error rates.

4. CONCLUSION

This paper has described adaptive training using structured transforms for separately removing complex non-speech variabilities. Cluster mean interpolation and CMLLR transforms were used as the structured transforms. ML adaptive training with structured transforms was presented as a simple extension to standard adaptive training. A discriminative criterion based on the MPE objective function was used to estimate multi-cluster model parameters, which finally leads to a simplified discriminative version of adaptive training with structured transforms. Experiments showed that the ST-based adaptive training significantly outperformed the traditional adaptive training techniques. These gains were observed for both MLE and MPE training. In future, more forms of prior distribution for multi-cluster HMM model will be investigated for improving performance of MPE training. The nature of structured transforms will be studied. Acoustic constrain will also be added to model different non-speech variabilities explicitly, which is closer to acoustic factorisation.

5. REFERENCES

[1] T.Anastasakos, J.Mcdonough, R.Schwartz, and J.Makhoul, "A compact model for speaker adaptive training," *Proc. ICSLP'96, Philadelphia*, pp. 1137–1140, 1996.

[2] M.J.F.Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.

[3] R.Kuhn, J.C.Junqua, P.Nguyen, and N.Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 695–707, 2000.

[4] M.J.F.Gales, "Multiple-cluster adaptive training schemes," *Proceedings ICASSP*, 2001.

[5] M.J.F.Gales, "Acoustic factorization," *ASRU 2001*, 2001.

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

[7] L.Wang and P.C.Woodland, "Discriminative Adaptive Training Using The MPE Criterion," *ASRU*, 2003, to appear.

[8] P.Nuguyen, C.Wellekens, and J.C.Junqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments," *Proc. Eurospeech'99*, pp. 2519–2522, 1999.

[9] P.C.Woodland and D. Povey, "Large scale discriminative training of hidden markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–48, 2002.

[10] Daniel Povey, "Discriminative training for large vocabulary speech recognition," *PhD Dissertation*, 2003.

[11] D Povey, P C Woodland, and M J F Gales, "Discriminative MAP for acoustic model adaptation," in *Proceedings ICASSP*, 2003.

[12] J.L.Gauvain and C.H.Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on SAP*, vol. 2, pp. 291–298, 1994.

[5] This was found to yield slightly better performance than eigenvoices initialisation which has a bias cluster with fixed weight associated though.

[6] For details see the presentations on the HTK web-site `http://htk.eng.cam.ac.uk/docs/cuhtk.shtml`.

[7] Statistical significance testing used NIST provided software `sctk-1.2`