Rapid Nonlinear Speaker Adaptation for Large-Vocabulary Continuous Speech Recognition

Zoi Roupakia, Anton Ragni and Mark Gales

Engineering Department, University of Cambridge, Cambridge, U.K.

zr216@cam.ac.uk, ar527@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Recently, kernel eigenvoices were revisited using kernel representations of distributions for rapid nonlinear speaker adaptation. These representations reassure the validity of the adapted distribution functions and enable expectation-maximisation training. Though gains have been shown in terms of word error rate for rapid speaker adaptation, this approach leads to an increase in decoding cost as the number of likelihood evaluations is amplified. The present paper addresses this issue by providing a coherent framework for systematic probabilistic approaches aimed at reducing the recognition cost and yet yielding equally powerful adapted models. The common denominator of such approaches is the use of probabilistic criteria, such as Kullback-Leibler divergence. However, in the general case, the resulting adapted models have full covariance matrices. In order to overcome this issue, the use of predictive semi-tied transforms to yield diagonal covariances for decoding is investigated in this paper. Experimental results are presented on a large-vocabulary conversational telephone task.

Index Terms: kernel eigenvoices, compact nonlinear adaptation, Kullback Leibler divergence

1. Introduction

The accuracy of state-of-the-art speech recognition systems is often degraded when there is a mismatch between the speakers and environments in training and those in test. Speaker adaptation approaches have been proposed to address the problem of speaker variability. As there is commonly not enough adaptation data to train speaker dependent models, most adaptation techniques, such as cluster adaptive training (CAT) [1], eigenvoices [2] and maximum likelihood stochastic transformations (MLST) [3], have focused on applying linear transformations to canonical model parameters. It is, however, widely known, that intra- and inter-speaker variability are nonlinear in nature. This has led to an interest in nonlinear adaptation.

Several nonlinear approaches have been developed to enhance the performance of speaker adaptation. Some, like neural networks [4], apply multilayer perceptrons to nonlinearly transform Gaussian means, while other nonlinear approaches can be shown to be kernelised versions of the standard linear techniques such as kernel eigenvoices [5]. The standard kernel eigenvoices, however, use an approximation of the required normalisation constant which can impact both estimation and recognition. Moreover, its application in large-vocabulary continuous speech recognition (LVCSR) is computationally and memory expensive due to the use of gradient-descent optimisation and large number of parameters to obtain eigenvoices.

Recently, kernel eigenvoice adaptation (KEA) has been revisited using kernel representations of distributions [6]. The latter provide valid distribution functions and expectation-maximisation training to be applied. Compact representations of the eigenvoices have been used to scale up its application to LVCSR. As these kernel-based approaches operate in high dimensional feature spaces, the decoding cost increases due to the number of likelihood, or kernel function, evaluations. Though several schemes are available to KEA to reduce decoding cost [5, 3], none of them take into account the probabilistic aspect of current speech recognisers. In contrast, model selection schemes employing probabilistic criteria have been used to reduce the size of statistical models in other areas [7], but have never been considered for nonlinear speaker adaptation. Furthermore, no methodical ways are available to address the issue of full-covariances in the adapted models.

This paper proposes a probabilistic framework to reduce the recognition cost without deteriorating the performance of the adapted models. Kullback-Leibler (KL) divergence-based schemes are exploited to reduce the size of these models, as this criterion is strongly related to maximum likelihood criterion. A clustering algorithm with a guaranteed increase in the objective function between two successive iterations is detailed, which yields compact nonlinearly adapted models. The algorithm reduces the computational load of minimising KL cost function from otherwise factorial to at most cubic in the number of components complexity. Given compact nonlinearly adapted models, this paper provides update equations to further refine model parameters. In order to address the generally full-covariance nature of the adapted models, schemes such as predictive semi-tied transforms are investigated [8].

2. Kernel Eigenvoice Adaptation

In [6], kernel eigenvoice adaptation (KEA) was revisited with the use of kernel representations of distributions guaranteeing that *valid* probabilistic functions are used in both adaptation and recognition stage. During adaptation, the new speaker s is defined as a "point" in speaker space by estimating E speaker dependent parameters $\lambda_e^{(s)}$, which weigh speaker independent eigenvoices $\boldsymbol{\theta}_{em}^{\Phi}$ in a high dimensional space $\mathbb{R}_{\boldsymbol{\Phi}}$ defined by a mapping function Φ. The eigenvoices are obtained from training data. As only E *scalar* weights are estimated from adaptation data, this scheme is suitable for rapid adaptation, when little data is available. The likelihood of HMM at state q at time t for speaker s is a M-component mixture model of the form

$$p(\mathbf{x}_t|q,s) = \sum_{m=1}^{M} c_m \sum_{e=1}^{E} \frac{\lambda_e^{(s)}}{Z_m^{(e)}} \langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_{em}^{\Phi} \rangle, \tag{1}$$

where $\langle .,. \rangle$ denotes the inner product (kernel function) between the mapped observation vector and the eigenvoices.

Applying the standard KEA to LVCSR systems is computationally prohibitive as speaker dependent models or transforms for thousands of speakers are built to obtain eigenvoices [5] or eigenmatrices [9]. This problem is addressed in [6] with the use of compact parametric approaches for each of the eigenvoices $\boldsymbol{\theta}^{\Phi}_{em}$ based on the constrained MLLR (CMLLR) technique [10]. Instead of training $M \times E$ distinct eigenvoices, only E CMLLR transforms $\mathbf{A}_e$ and biases $\mathbf{b}_e$ are trained. The form of the inner product used in this representation is based on a Gaussian kernel in the following form

$$\langle \Phi(\mathbf{x}_t), \Phi(\boldsymbol{\mu}^{(e)}_m) \rangle = \exp(-\frac{1}{2} \|\boldsymbol{\Sigma}^{(e)}_m{}^{-1/2}(\mathbf{x}_t - \boldsymbol{\mu}^{(e)}_m)\|^2_2), \quad (2)$$

where $\boldsymbol{\mu}^{(e)}_m$ is the basis mean obtained from the canonical model mean vector $\boldsymbol{\mu}_m$ as

$$\boldsymbol{\mu}^{(e)}_m = \mathbf{A}_e^{-1}\boldsymbol{\mu}_m - \mathbf{A}_e^{-1}\mathbf{b}_e \quad (3)$$

and $\boldsymbol{\Sigma}^{(e)}_m$ is

$$\boldsymbol{\Sigma}^{(e)}_m = \mathbf{A}_e^{-1}\boldsymbol{\Sigma}_m\mathbf{A}_e^{-1^{\mathsf{T}}}. \quad (4)$$

The local normalisation constant $Z^{(e)}_m = |\mathbf{A}_e|/(2\pi^d|\boldsymbol{\Sigma}_m|)^{1/2}$ ensures the validity of the adapted distribution. This allows an EM, rather than gradient-descent, based parameter estimation used in the standard KEA. It can be shown [6] that given the inner product in equation (2) the likelihood in equation (1) is equivalent to the following Gaussian mixture model (GMM) called in this paper as the *original GMM*

$$\begin{aligned} p(\mathbf{x}_t|q,s) &= \sum_{m=1}^{M} c_m \sum_{e=1}^{E} \lambda^{(s)}_e |\mathbf{A}_e| \mathcal{N}(\mathbf{A}_e\mathbf{x}_t + \mathbf{b}_e; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \\ &= \sum_{m=1}^{M}\sum_{e=1}^{E} c_m \lambda^{(s)}_e \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}^{(e)}_m, \boldsymbol{\Sigma}^{(e)}_m). \end{aligned} \quad (5)$$

With the use of Gaussian kernel, the resulting model is thus an expanded speaker dependent GMM with $M_s = M \times E$ components. For the rest of the paper (m, e) denotes the mixture component of the original GMM with mean $\boldsymbol{\mu}^{(e)}_m$, covariance $\boldsymbol{\Sigma}^{(e)}_m$ and weight $c_m\lambda^{(s)}_e$. Though WER gains were reported given little adaptation data, the proposed scheme increases the decoding cost linearly as the number of likelihood evaluations has increased by E times. For the experimental setup of this work, this results in 32 times more expensive recognition cost.

3. Compact Nonlinear Adaptation

Kernel eigenvoice adaptation (KEA) provides the framework for rapid nonlinear adaptation when little data is available. However, as described in section 2, it significantly slows down the decoding process. Using KEA to yield the original GMM in equation (5), this section aims at finding a *compact* GMM

$$\hat{p}(\mathbf{x}_t|q,s) = \sum_{m=1}^{\hat{M}} \hat{c}^{(s)}_m \mathcal{N}(\mathbf{x}_t; \hat{\boldsymbol{\mu}}^{(s)}_m, \hat{\boldsymbol{\Sigma}}^{(s)}_m). \quad (6)$$

If $\hat{M}{=}M$ and $\hat{\boldsymbol{\Sigma}}^{(s)}_m$ is constrained to be diagonal, the compact system has similar decoding cost to the original system. This GMM will be referred as the *target GMM*. The latter, like the original GMM, is supposed to maximise the likelihood of the adaptation data. The cost function should therefore minimise the expected loss in likelihood by using the target GMM $\hat{p}$ instead of the original GMM p. One suitable function is the Kullback-Leibler divergence between two GMMs. The parameters of the target GMM $\{\hat{c}^{(s)}_m, \hat{\boldsymbol{\mu}}^{(s)}_m, \hat{\boldsymbol{\Sigma}}^{(s)}_m\}$, thus, should be estimated so that

$$\underset{\{\hat{c}^{(s)}_m, \hat{\boldsymbol{\mu}}^{(s)}_m, \hat{\boldsymbol{\Sigma}}^{(s)}_m\}}{\operatorname{argmin}} \{\mathcal{KL}(p||\hat{p})\}. \quad (7)$$

The estimation of the model parameters in equation (7) depends on the nonlinearly adapted model, so that the target GMM constitutes a more compact form for nonlinear adaptation.

Unfortunately, Kullback-Leibler (KL) divergence between two GMMs has no analytic closed form solution. One well known option to address this issue is to adopt the matched pair bound approximation (MPBA)[11]. The latter is an upper bound of the true KL divergence between two GMMs. The form used in this work is given by

$$\mathcal{KL}(p||\hat{p}) \leq \sum_{m=1}^{M}\sum_{e=1}^{E} c_m\lambda^{(s)}_e \mathcal{KL}(\mathcal{N}^e_m||\hat{\mathcal{N}}^{(s)}_{\hat{m}}), \quad (8)$$

where $\mathcal{N}^e_m$ and $\hat{\mathcal{N}}^{(s)}_{\hat{m}}$ denote the mixture components of the original $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}^{(e)}_m, \boldsymbol{\Sigma}^{(e)}_m)$ and the target $\mathcal{N}(\mathbf{x}_t; \hat{\boldsymbol{\mu}}^{(s)}_{\hat{m}}, \hat{\boldsymbol{\Sigma}}^{(s)}_{\hat{m}})$ GMM respectively. The component $\hat{m}$ is obtained by a mapping function f, which assigns each component of the original GMM $(m, e) \in M_s$ to one component of the target GMM $\hat{m} \in \hat{M}$. The natural choice for mapping function is given by

$$f(m,e) = \underset{\hat{m}\in\hat{M}}{\operatorname{argmin}}\{\mathcal{KL}(\mathcal{N}^e_m||\hat{\mathcal{N}}^{(s)}_{\hat{m}})\}, \quad (9)$$

which determines the optimal component $\hat{m}$ by minimising the KL divergence from the original GMM component (m, e). Other popular methods for estimating KL divergence between GMMs are based on Monte Carlo sampling and variational approximations [11]. Though known for being more accurate, they are computationally more expensive and hence, in this work, the simpler approach based on MPBA is investigated.

3.1. Model Selection Process

In order to minimise the KL divergence between the two GMMs, a mapping function $f(m, e)$ needs to be obtained. As $f(m, e)$ minimises the KL divergences between components of the original and target GMM, an initial model for the latter is required. In this work, the target GMM is initialised by a subset of components from the original GMM

$$\{\hat{c}^{(s)}_m, \hat{\boldsymbol{\mu}}^{(s)}_m, \hat{\boldsymbol{\Sigma}}^{(s)}_m\} \subset \{c_m\lambda^{(s)}_e, \boldsymbol{\mu}^{(e)}_m, \boldsymbol{\Sigma}^{(e)}_m\}. \quad (10)$$

This subset consists of $\hat{M}$ elements named as the set of *active components* and the remained $M_d = M_s - \hat{M}$ components of the original GMM belong to the set of *deactivated components*. The latter consists of components, which if deactivated, yield the minimal expected loss in likelihood

$$\underset{M_d}{\operatorname{argmin}}\{ \sum_{(m,e)\in M_d} c_m\lambda^{(s)}_e \mathcal{KL}(\mathcal{N}^e_m||\hat{\mathcal{N}}^{(s)}_{\hat{m}})\}, \quad (11)$$

where the component $\hat{m}$ is obtained by the mapping function that assigns each deactivated component of the original GMM to one active component according to equation (9). Unfortunately, finding the globally optimal set of deactivated components is computationally intractable, as the cost of different combinations of components to form that set is $\frac{M_s!}{M_d!\hat{M}!}$.

Instead of examining all possible combinations of components in a brute force manner, it is possible to minimise the KL distance between two GMMs in equation (8) iteratively through a *clustering scheme* providing a less tighter bound. At the beginning, each component $(m, e) \in M_s$ forms a cluster with centroid $(\tilde{m}, \tilde{e}) = (m, e)$. The cost of each cluster is defined by

$$\mathcal{C}(\tilde{m}, \tilde{e}) = \sum_{\substack{(m,e)\in\tilde{M} \\ (m,e)\neq(\tilde{m},\tilde{e})}} c_m \lambda_e^{(s)} \mathcal{KL}\big(\mathcal{N}_m^e || \mathcal{N}_{\tilde{m}}^{\tilde{e}}\big), \tag{12}$$

where $\tilde{M}$ are the cluster members. Therefore, in the first iteration, the total cost of clusters is zero. In each iteration i, one cluster of the original GMM with centroid $(\tilde{m}, \tilde{e})$ and $\tilde{M}$ members is deactivated, so that the KL divergence between the original p and the current target $\hat{p}^{(i)}$ GMM is minimised. Then, the components of the deactivated cluster are re-assigned to other clusters based on the KL divergence between them and the centroid of current clusters using equation (9), and the cost of each cluster is updated.

One attribute of this clustering scheme is that the KL divergence between the original and the current target GMM is guaranteed to be increased in each iteration

$$\mathcal{KL}\big(p||\hat{p}^{(i)}\big) = \mathcal{KL}\big(p||\hat{p}^{(i-1)}\big) + \Delta\mathcal{KL}\big(p||\hat{p}^{(i)}\big), \tag{13}$$

where $\Delta\mathcal{KL}(p||\hat{p}^{(i)}) \geq 0$ is the increase in KL divergence. It is therefore sufficient to minimise only

$$\underset{(\tilde{m},\tilde{e})}{\operatorname{argmin}}\{\Delta\mathcal{KL}(p||\hat{p}^{(i)})\}. \tag{14}$$

This increase consists of the difference between two costs, the cost of re-assigning the members of the cluster to new ones (first term in (15)) and the cost of the cluster in the previous iteration

$$\Delta\mathcal{KL}(p||\hat{p}^{(i)}) = \sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)} \mathcal{KL}\big(\mathcal{N}_m^e || \hat{\mathcal{N}}_{f(m,e)}^{(s)}\big) - \mathcal{C}(\tilde{m}, \tilde{e}). \tag{15}$$

It is important to note that the components of the deactivated cluster are not constrained to follow the centroid and can be re-assigned to other components that minimise their corresponding KL divergences given the mapping function (9). This property prevents the development of very large clusters. Fig. 1 illustrates the clustering process in two successive iterations. In iteration 1, four clusters exist (A,B,C,D). Deactivating cluster C with centroid $(3, 1)$ and cluster member $(2, 1)$ is assumed to yield the minimum increase in KL divergence between the original and the current target GMM. Thus, it is eliminated in iteration 2. The cluster components are then re-assigned to other clusters (A,B,D). Arrows show the KL distances between component (3,1) and the centroids of clusters A, B, D. In this case component $(3, 1)$ is assigned to $(4, 1)$ and $(2, 1)$ to $(1, 2)$.

Clustering is terminated when $\hat{M}$ clusters are formulated and the centroids of the clusters are used to initialise the target GMM. To ensure the validity of that distribution, it is necessary to update the weights using the following updating equation

$$\hat{c}_{\hat{m}}^{(s)} = \sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)}. \tag{16}$$

In this case, only the weights are speaker dependent, while the means and the covariances are speaker independent. As the latter are originated from CMLLR transformed covariances, the decoding cost is $\mathcal{O}(\hat{M}d)$.

Variations of this implementation can yield greedier clustering schemes, such as the one used in [7]. There, the components assigned to centroids are not allowed to be remapped to different clusters as soon as this cluster is deactivated and the scheme operates in a merging-down fashion.

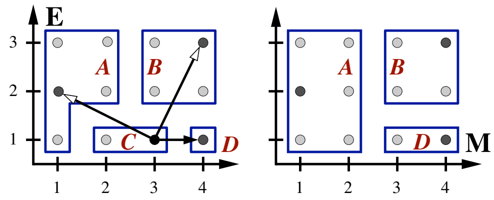

Figure 1: *Illustration of clustering between two iterations.*

3.2. Refining the Model Parameters

Rather than estimating only the component weights, the means and the covariances of the target GMM may also be updated by minimising the KL divergence between the original GMM and the initialised target GMM. The updated mean is

$$\hat{\boldsymbol{\mu}}_{\hat{m}}^{(s)} = \frac{\sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)} \boldsymbol{\mu}_m^{(e)}}{\sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)}}, \tag{17}$$

where $\tilde{M}$ are the members of the cluster with centroid $\hat{m}$, while the updated covariance is

$$\hat{\boldsymbol{\Sigma}}_{\hat{m}}^{(s)} = \frac{\sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)} \Big(\boldsymbol{\Sigma}_m^{(e)} + (\boldsymbol{\mu}_m^{(e)} - \hat{\boldsymbol{\mu}}_{\hat{m}}^{(s)})(\boldsymbol{\mu}_m^{(e)} - \hat{\boldsymbol{\mu}}_{\hat{m}}^{(s)})^{\mathsf{T}}\Big)}{\sum_{(m,e)\in\tilde{M}} c_m \lambda_e^{(s)}}. \tag{18}$$

In this case, it is apparent that the means and the covariances are speaker dependent and the covariance matrix is full. If only the means are updated, the decoding cost is $\mathcal{O}(\hat{M}d)$, but updating full covariance matrices yields cost of $\mathcal{O}(\hat{M}d^2)$.

Updating the covariance matrices provides the target GMM with full covariance matrices. Though the number of components in the target GMM has been reduced E times and less likelihood evaluations are needed, the recognition cost increases from diagonal- to full-covariance decoding. One option to overcome this problem is to diagonalise the full covariance matrices by keeping only the diagonal elements. Another way to achieve that is through the use of semi-tied transforms. Training a semi-tied transform from little adaptation data may not yield robust estimates. As an alternative, predictive semi-tied transforms can be considered [8]. In this paper, the semi-tied transforms are estimated by minimising the KL divergence between the *original GMM* and a *semi-tied target GMM* distribution having the following form given the formulated clusters

$$\hat{p}(\mathbf{x}_t|q, s) = \sum_{m=1}^{\hat{M}} \hat{c}_m^{(s)} |\mathbf{A}| \mathcal{N}(\mathbf{A}\mathbf{x}_t; \mathbf{A}\hat{\boldsymbol{\mu}}_m^{(s)}, \hat{\boldsymbol{\Sigma}}_{m(diag)}^{(s)}), \tag{19}$$

where $\hat{\boldsymbol{\Sigma}}_m^{(s)} = \mathbf{A}^{-1}\hat{\boldsymbol{\Sigma}}_{m(diag)}^{(s)}\mathbf{A}^{-1^{\mathsf{T}}}$ and $\mathbf{A}$ is the global semi-tied transform. The statistics needed to obtain the transform are based on the component occupancies taken from the training data [8]. This results in a *diagonal target GMM* and the decoding cost for $\hat{M}$ components reduces from $\mathcal{O}(\hat{M}d^2)$ to $\mathcal{O}(\hat{M}d)$.

4. Experiments

Compact nonlinear adaptation was evaluated on a large-vocabulary conversational telephone speech (CTS) task. The 76-hour training dataset (h5etrainsub) consists of 1118 speakers and 77201 utterances in three corpora. A 3-hour subset of the 2001 development data for CTS (dev01sub) was used for evaluation. This subset has 59 speakers (30 female, 29 male) and 2663 utterances. The feature vectors were 12-D MF-PLP with c_o energy, first, second and third derivatives. Cepstral mean and variance normalisation (CMN, CVN) was applied. The data was projected to a 39-D space by an HLDA transform. Vocal tract length normalisation (VTLN) was used. For consistency with previous systems, VTLN, CMN and CVN were estimated per-speaker. It should be emphasised that all these decorrelating and adaptation approaches are expected to limit the pontential gains from the application of additional adaptation methods including KEA. State-clustered triphone HMMs with 12 components per state were maximum-likelihood trained. The supervision hypotheses for estimating the adaptation parameters were generated using the speaker independent (SI) models. Adaptation was performed at the utterance or the speaker level. The approaches to reduce the size of the original KEA system were evaluated at the speaker level, as it is computationally expensive to apply them for all the 2663 separate utterances. The gains, however, are expected to be representative.

Table 1: *WER of Baseline and Adaptation Approaches.*

System	Update			Decoding	WER	
	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\mathbf{A}$	Cost	Spkr	Utter
SI	-			$\mathcal{O}(Md)$	34.6	
CMLLR	-			$\mathcal{O}(Md)$	33.0	34.4
KEA	-			$\mathcal{O}(MEd)$	33.3	33.2
KEA+CLST	✗	✗	✗	$\mathcal{O}(Md)$	34.8	-
	✓	✗	✗	$\mathcal{O}(Md)$	34.7	-
KEA+fullC	✓	✓	✗	$\mathcal{O}(Md^2)$	33.8	-
KEA+diagC	✓	✓	✗	$\mathcal{O}(Md)$	34.1	-
KEA+pSEMIT	✓	✓	✓	$\mathcal{O}(Md)$	34.3	-

The original KEA system was built using an iterative transform splitting approach, where the number of CMLLR transforms (eigen-bases) was increased to 32. Hence, the original GMM has 384 components. The original KEA gave gains of 1.4% and 1.3% absolute over the SI system at the utterance and speaker level, as it is shown in Table 1. CMLLR performed well at the speaker level expectedly, but it was outperformed by KEA when little data is available. Using the clustering approach, the size of the target GMM was reduced from 384 to 12 components and the resulting target system (KEA+CLST) was evaluated in each step of refining its parameters (weights, means, covariances). Predictive semi-tied transforms were applied using component occupancies from training data (KEA+pSEMIT). Table 1 shows the decoding cost and the performance of the compact nonlinear adaptation approaches for the target GMM with 12 components (equivalent number to the size of SI system). The target KEA system with updated weights, or weights and means deteriorated the recognition accuracy. Updating the full covariance matrices (KEA+fullC) improved the WER over the SI system by 0.8% absolute, but increased the decoding cost to $\mathcal{O}(Md^2)$. Global KEA+pSEMIT did not yield the expected gains in WER, though it improved the auxiliary function, while zeroing the off-diagonal elements (KEA+diagC) had better WER. Table 2 shows that the WER of KEA+pSEMIT is bounded by the performance of KEA+fullC and global KEA+pSEMIT, while state dependent transforms increased the WER over KEA+fullC by just 0.2% absolute.

Table 2: *WER of KEA+pSEMIT vs Regression Classes.*

# Classes	1	4	256	1024	5923(state)
WER	34.3	34.4	34.3	34.2	34.0

Table 3 shows the performance of KEA+CLST systems as the number of components decrease from 384 to 96, 48, 24 and 12. Gains of 1% and 1.4% absolute over the SI system were obtained by KEA+CLST with four and eight times less components than KEA respectively. The 48-component KEA+CLST had almost nine times less decoding cost than the 12-component full-covariance KEA+CLST and yielded better performance.

Table 3: *WER of KEA+CLST vs Number of Components.*

# Components	12	24	48	96	384
WER	34.8	34.1	33.6	33.2	33.3

5. Conclusion

This paper has examined kernel eigenvoice adaptation for rapid nonlinear speaker adaptation in large-vocabulary continuous speech recognition. Kernel eigenvoice adaptation was presented using kernel representations of distributions and the issues related to the expensive decoding cost were analysed. The general framework for nonlinear adaptation was described using Kullback Leibler divergence as criterion to obtain a more compact nonlinearly adapted model. An optimal clustering scheme was detailed in order to obtain this model. Predictive semi-tied transforms were proposed to speed up the decoding of the resulting full-covariance system. Results on a conversational telephone speech task showed relative gains in the case of the full- (and diagonalised-) covariance model when the number of components is equivalent to the size of speaker independent system.

6. References

[1] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," in *IEEE TSAP*, 2000.

[2] R. Kuhn, P. Nguyen, J.-C. Jungua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *ICSLP*, 1998.

[3] V. D. Diakoloukas and V. V. Digalakis, "Maximum-likelihood stochastic transformation adaptation of hidden Markov models," in *IEEE TSAP*, 1999.

[4] V. Abrash, A. Sankar, H. Franco, and M. Cohen, "Acoustic adaptation using nonlinear transformations of HMM parameters," in *ICASSP*, 1996.

[5] S. H. B. Mak and J. T. Kwok, "Speedup of kernel eigenvoice speaker adaptation by embedded kernel PCA," in *ICSLP*, 2004.

[6] Z. Roupakia and M. Gales, "Kernel eigenvoices (revisited) for large-vocabulary speech recognition," *IEEE SPL*, 2011.

[7] P. L. Dognin, J. R. Hershey, V. Goel, and P. A. Olsen, "Refactoring acoustic models using variational density approximation," in *ICASSP*, 2009.

[8] M. J. F. Gales and R. C. V. Dalen, "Predictive linear transforms for noise robust speech recognition," in *ASRU*, 2007.

[9] B. Mak and R. Hsiao, "Kernel eigenspace-based MLLR adaptation," in *IEEE TSAP*, 2007.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *CSL*, 1998.

[11] J. Hershey and P. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *ICASSP*, 2007.