Speedup of Kernel Eigenvoice Speaker Adaptation by Embedded Kernel PCA

Brian Mak, Simon Ho, and *James T. Kwok*

Department of Computer Science
Hong Kong University of Science & Technology, Hong Kong
{mak,csho,james}@cs.ust.hk

Abstract

Recently, we proposed an improvement to the eigenvoice (EV) speaker adaptation called *kernel eigenvoice (KEV) speaker adaptation*. In KEV adaptation, eigenvoices are computed using kernel PCA, and a new speaker's adapted model is implicitly computed in the kernel-induced feature space. Due to many online kernel evaluations, both adaptation and subsequent recognition of KEV adaptation are slower than EV adaptation. In this paper, we eliminate all online kernel computations by finding an approximate *pre-image* of the implicit adapted model found by KEV adaptation. Furthermore, the two steps of finding the implicit adapted model and its approximate pre-image are integrated by embedding the kernel PCA procedure in our new *embedded kernel eigenvoice (eKEV) speaker adaptation* method. When tested in an TIDIGITS task with less than 10s of adaptation speech, eKEV adaptation obtained a speedup of 6–14 times in adaptation and 136 times in recognition over KEV adaptation with 12–13% relative improvement in recognition accuracy.

1. Introduction

Adaptation methods like the Bayesian-based MAP adaptation [1] and the transformation-based MLLR adaptation [2] have been popular for many years. Nevertheless, when the amount of available adaptation speech is really small — say, only a few seconds — the more recent eigenvoice-based adaptation method is found particularly more effective. The basic idea of the eigenvoice (EV) adaptation method [3] is to derive from a diverse set of speakers a small set of basis vectors called *eigenvoices* that are believed to represent different voice characteristics (e.g. gender, age, accent, etc.); any training or new speaker is then a point in the eigenvoice subspace. In practice, since the number of estimation parameters is greatly reduced, fast speaker adaptation using EV adaptation is possible with a few seconds of speech.

Recently, we proposed an improvement to the EV adaptation called *kernel eigenvoice (KEV) speaker adaptation* [4, 5] by exploiting possible nonlinearity in the speaker supervector space using kernel methods [6]. The basic idea is to map speaker supervectors to a high dimensional feature space[1] via some nonlinear map, and then apply principal component analysis (PCA) there to derive the eigenvoices in the feature space. During the actual computation, the exact nonlinear map need not be known, and the kernel eigenvoices are obtained by *kernel PCA*. In an TIDIGITS adaptation task, KEV adaptation was shown to outperform an speaker-independent model by about 30% using less than 10s of adaptation speech [4], and was better than EV, MAP, and MLLR adaptation [5].

However, adaptation and subsequent recognition using KEV adaptation method can be substantially slower than EV adaptation due to many online kernel evaluations in the computation of observation likelihoods. The problem is that since the eigenvoices found by KEV adaptation reside in the kernel-induced feature space, a new speaker adapted (SA) model is represented *implicitly* as a linear combination of these kernel eigenvoices only in the feature space. As there is no *explicit* SA model in the input speaker supervector space, any computation of its observation likelihoods has to be done online using the *implicit* SA model in the feature space via expensive kernel evaluations. Finding an object in the input space from its image in the feature space is known as the *pre-image* problem in kernel methods. In this paper, we attempt to solve the efficiency problem by integrating the finding of the implicit SA model in the feature space and its pre-image together so as to arrive at an *explicit* SA model in the input speaker supervector space. The novelty of our method is that there are no kernel evaluations during adaptation involving the adaptation speech, and there are no kernel evaluations at all during recognition. Consequently, adaptation is faster and subsequent recognition runs as fast as conventional EV adaptation. Our new method will be called *embedded kernel eigenvoice (eKEV) speaker adaptation*.

2. Embedded Kernel Eigenvoice Speaker Adaptation (eKEV)

Suppose there is a set of N speaker-dependent (SD) acoustic models which are hidden Markov models (HMMs) of the same topology with mixture Gaussian states. For each speaker, a *speaker supervector* is constructed by concatenating all his HMM Gaussian mean vectors. That is, for the ith speaker, if there are R Gaussians in his HMMs, each having a mean vector $\mathbf{x}_{ir} \in \mathbb{R}^{n_1}, r = 1, \ldots, R$, then his speaker supervector is $\mathbf{x}_i = [\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iR}]' \in \mathbb{R}^{n_2}$ and $n_2 = n_1 R$.

2.1. Variance Normalization

Because the pre-image finding algorithm uses Euclidean distance constraints, the constituent means of a speaker supervector $\mathbf{x}$ is first normalized by its own covariance to the normalized supervector $\mathbf{y} = \mathbf{C}^{-\frac{1}{2}}\mathbf{x}$ where

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_R \end{bmatrix}. \quad (1)$$

[1] In the kernel methods terminology, the original space where raw data reside is called the *input space* and the space to which raw data are mapped is called the *feature space*. In order not to confuse this with the acoustic feature space in speech, the feature space in kernel methods will be simply called the "feature space" but may be sometimes called the "*kernel-induced feature space*" for additional clarity.

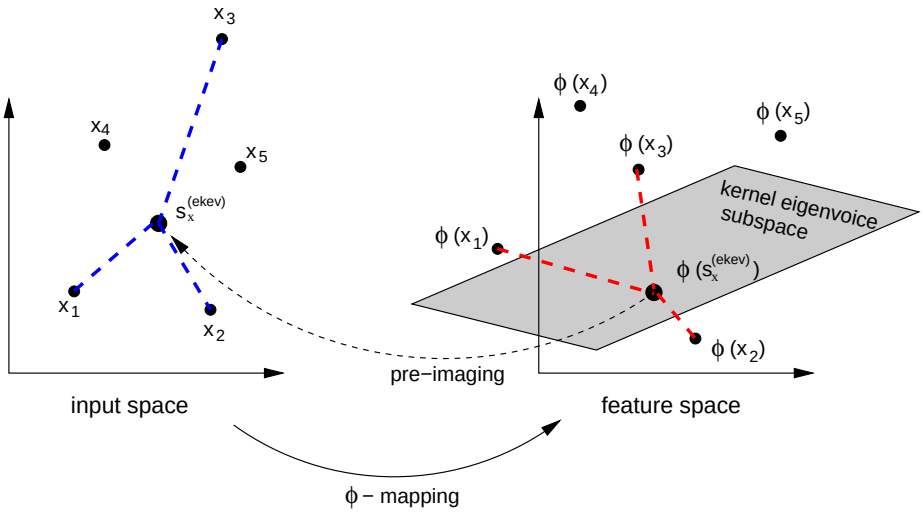

Figure 1: The eKEV adaptation method. (Without the pre-imaging step, it is the KEV adaptation method.)

Similarly, the new speaker's adapted model will be represented by $\mathbf{s}_x^{(ekev)}$ and $\mathbf{s}_y^{(ekev)}$ in the original and normalized input supervector space respectively.

2.2. Kernel Eigenvoices

Let $k(\cdot,\cdot)$ be the kernel with an associated mapping φ which maps a speaker $\mathbf{y} \in \mathbb{R}^{n_2}$ in the normalized input speaker supervector space $\mathcal{Y}$ to $\varphi(\mathbf{y}) \in \mathbb{R}^{n_3}$ in the kernel-induced high dimensional feature space $\mathcal{F}$. Given the set of N speaker supervectors $\{\mathbf{y}_1,\ldots,\mathbf{y}_N\} \in \mathcal{Y}$, their φ-mapped feature vectors are $\{\varphi(\mathbf{y}_1),\ldots,\varphi(\mathbf{y}_N)\} \in \mathcal{F}$. Let the "centered" map be $\tilde{\varphi}$ so that $\tilde{\varphi}(\mathbf{y}) = \varphi(\mathbf{y}) - \bar{\varphi}$ where $\bar{\varphi} = \frac{1}{N}\sum_{i=1}^{N}\varphi(\mathbf{y}_i)$. In addition, let $\mathbf{K}$ be the kernel matrix with $\mathbf{K}_{ij} \equiv k(\mathbf{y}_i,\mathbf{y}_j) \equiv \varphi(\mathbf{y}_i)'\varphi(\mathbf{y}_j)$, and $\tilde{\mathbf{K}}$ be its centered version with $\tilde{\mathbf{K}}_{ij} = \tilde{\varphi}(\mathbf{y}_i)'\tilde{\varphi}(\mathbf{y}_j)$.

Kernel PCA may be performed by eigendecomposition on $\tilde{\mathbf{K}}$ as $\tilde{\mathbf{K}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, where $\mathbf{U} = [\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_N]$ with $\boldsymbol{\alpha}_i = [\alpha_{i1},\ldots,\alpha_{iN}]'$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_N)$. The mth orthonormal eigenvector of the covariance matrix in the feature space $\mathcal{F}$ is given by [7] as

$$\mathbf{v}_m = \sum_{i=1}^{N} \frac{\alpha_{mi}}{\sqrt{\lambda_m}} \tilde{\varphi}(\mathbf{y}_i)\ , \tag{2}$$

which are our kernel eigenvoices in the feature space $\mathcal{F}$.

2.3. Composite Kernel

In our KEV adaptation, composite kernels are used to preserve HMM state information which is generally lost during the φ-mapping of speaker supervectors from the input space $\mathcal{Y}$ to the feature space $\mathcal{F}$. That is, a possibly different mapping is used for each of the R constituent Gaussian means, and then a composite function is applied. For example, the following direct sum composite kernel was used in [4, 5] with good results:

$$k(\mathbf{y}_i,\mathbf{y}_j) = \sum_{r=1}^{R} \varphi_r(\mathbf{y}_{ir})'\varphi_r(\mathbf{y}_{jr}) = \sum_{r=1}^{R} k_r(\mathbf{y}_{ir},\mathbf{y}_{jr})\ . \tag{3}$$

The direct sum composite kernel again will be used in the following discussion of eKEV adaptation though its use is not necessary for the eKEV adaptation method. They are retained so that we may compare KEV and eKEV adaptation using the same kind of kernels; another reason is that they give good adaptation performance in our previous works [4, 5].

2.4. eKEV Adaptation Procedure

In eKEV adaptation, all online kernel evaluations with speech observations are eliminated by using the analytical solution in [8] to find an approximate pre-image[2] of the adapted model found by KEV adaptation which resides in the feature space. The method uses the distances between the expected pre-image and a set of *"reference speakers"* as constraints and solves for the optimal pre-image in the least-square sense[3].

The eKEV adaptation method is illustrated pictorially in Figure 1 where there are only 5 training speakers $\mathbf{x}_1 - \mathbf{x}_5$ and $\mathbf{x}_1 - \mathbf{x}_3$ are chosen as the reference speakers.

STEP 1: Similarity between the New Speaker and the Reference Speakers in the Feature Space

As in EV adaptation, the adapted speaker model $\tilde{\varphi}(\mathbf{s}_y^{(ekev)})$ is a linear combination of the M leading kernel eigenvoices in the feature space. Using Eqn. (2), we have

$$\tilde{\varphi}(\mathbf{s}_y^{(ekev)}) = \sum_{m=1}^{M} w_m\mathbf{v}_m = \sum_{m=1}^{M}\sum_{i=1}^{N} \frac{w_m\alpha_{mi}}{\sqrt{\lambda_m}}\tilde{\varphi}(\mathbf{y}_i)\ . \tag{4}$$

And its rth constituent is given by

$$\tilde{\varphi}_r(\mathbf{s}_{yr}^{(ekev)}) = \sum_{m=1}^{M}\sum_{i=1}^{N} \frac{w_m\alpha_{mi}}{\sqrt{\lambda_m}}\tilde{\varphi}_r(\mathbf{y}_{ir})\ . \tag{5}$$

Hence, the similarity between the rth constituent $\mathbf{s}_{yr}^{(ekev)}$ of the new speaker model $\mathbf{s}_y^{(ekev)}$ and the rth constituent $\mathbf{y}_{jr}$ of the jth reference speaker $\mathbf{y}_j$ in $\mathcal{F}$ can be computed by

$$\begin{aligned} k_r(\mathbf{s}_{yr}^{(ekev)},\mathbf{y}_{jr}) &\equiv \varphi_r(\mathbf{s}_{yr}^{(ekev)})'\varphi_r(\mathbf{y}_{jr}) \\ &= A_r(j) + \sum_{m=1}^{M}\frac{w_m}{\sqrt{\lambda_m}}B_r(m,j)\ , \end{aligned} \tag{6}$$

[2] Generally speaking, an exact pre-image may not exist.

[3] It is analogous to finding the location of an object using a set of global positioning system satellites.

where

$$A_r(j) = \frac{1}{N}\sum_{i=1}^{N} k_r(\mathbf{y}_{ir}, \mathbf{y}_{jr}) \,, \tag{7}$$

and

$$B_r(m,j) = \sum_{i=1}^{N} \alpha_{mi}\left(k_r(\mathbf{y}_{ir}, \mathbf{y}_{jr}) - A_r(j)\right) \,. \tag{8}$$

STEP 2: Finding the Distances of all Reference Speakers from Their Centroid in the Input Space

Without loss of generality, let the column vectors of $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ be the n *reference speakers*. Assuming that the rank of $\mathbf{Y}$ is q, singular value decomposition (SVD) of the centered $\mathbf{Y}$ gives

$$\tilde{\mathbf{Y}} = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{V}' = \mathbf{U}_2\mathbf{Z} \,, \tag{9}$$

where $\mathbf{U}_2 = [\mathbf{e}_1, \ldots, \mathbf{e}_q]$ is an $n_2 \times q$ matrix with orthonormal columns $\mathbf{e}_i$; $\mathbf{\Lambda}_2 = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ is a $q \times q$ diagonal matrix containing the eigenvalues; $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ is a $q \times n$ matrix with columns $\mathbf{z}_i$ being the projections of $\mathbf{y}_i$ onto the $\mathbf{e}_j$'s. Thus, $\|\mathbf{z}_i\|^2$ gives the squared Euclidean distance of each $\mathbf{y}_i$ from the centroid $\bar{\mathbf{y}}$. They are collected into a vector,

$$\mathbf{d}_0 = [\|\mathbf{z}_1\|^2, \|\mathbf{z}_2\|^2, \ldots, \|\mathbf{z}_n\|^2]' \in \mathbb{R}^n \,. \tag{10}$$

STEP 3: Finding the Distance Constraints between the New Speaker and the Reference Speakers in the Input Space

It is further assumed that the required pre-image $\mathbf{s}_y^{(ekev)}$ lies in the span of the n reference speakers. If the direct sum composite kernel of Eqn. (3) is used, and each constituent kernel is an isotropic Gaussian kernel of the form: $k_r(\mathbf{u}, \mathbf{v}) = \exp(-\beta_r\|\mathbf{u} - \mathbf{v}\|^2)$, then we have

$$k_r(\mathbf{s}_{yr}^{(ekev)}, \mathbf{y}_{jr}) = e^{-\beta_r\|\mathbf{s}_{yr}^{(ekev)} - \mathbf{y}_{jr}\|^2} = e^{-\beta_r d_{jr}} \,.$$

Thus, the squared Euclidean distance d_j between $\mathbf{s}_y^{(ekev)}$ and $\mathbf{y}_j$ in the *input space* can be computed from their similarity in the *feature space* as follows:

$$d_j \equiv \sum_{r=1}^{R} d_{jr} = -\sum_{r=1}^{R} \frac{1}{\beta_r} \log k_r(\mathbf{s}_{yr}^{(ekev)}, \mathbf{y}_{jr}) \,. \tag{11}$$

These distances can be computed using Eqns. (6,7,8), and are collected into the vector

$$\mathbf{d}(\mathbf{w}) = [d_1, d_2, \ldots, d_n]' \in \mathbb{R}^n \,. \tag{12}$$

STEP 4: Finding the Distance Gradients

Differentiating $\mathbf{d}$ of Eqn. (11) w.r.t. the mth eigenvoice weight w_m, we get

$$\frac{\partial d_j}{\partial w_m} = -\frac{1}{\sqrt{\lambda_m}} \sum_{r=1}^{R} \frac{B_r(m,j)}{\beta_r k_r(\mathbf{s}_{yr}^{(ekev)}(\mathbf{w}), \mathbf{y}_{jr})} \,, j = 1, \ldots, n \,. \tag{13}$$

STEP 5: Finding the Pre-image

From [8], an approximate pre-image that optimally satisfies the distance constraints in $\mathbf{d}$ in the least-square sense is given by the following equation:

$$\mathbf{s}_x^{(ekev)}(\mathbf{w}) = \mathbf{C}^{\frac{1}{2}}\mathbf{s}_y^{(ekev)} = \mathbf{C}^{\frac{1}{2}}(\mathbf{P}\mathbf{d}(\mathbf{w}) + \mathbf{q}) \,, \tag{14}$$

where

$$\mathbf{P} = -\frac{1}{2}\mathbf{U}_2\mathbf{\Lambda}_2^{-1}\mathbf{V}' \text{ and } \mathbf{q} = -\mathbf{P}\mathbf{d}_0 + \bar{\mathbf{y}} \,, \tag{15}$$

and $\mathbf{U}_2$, $\mathbf{\Lambda}_2$, and $\mathbf{V}$ are given by Eqn. (9).

STEP 6: ML Estimation of Kernel Eigenvoice Weights

A maximum likelihood estimation of $\mathbf{w}$ may be found by maximizing the following $Q(\mathbf{w})$ function:

$$Q(\mathbf{w}) = -\sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t(r)\|\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w})\|^2_{\mathbf{C}_r} \,, \tag{16}$$

where $\gamma_t(r)$ is the posterior probability of the observation sequence $\mathbf{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ being at the rth Gaussian at time t; $\mathbf{s}_{xr}^{(ekev)}$ is the rth constituent of the new speaker's model which is also the mean vector of his rth HMM Gaussian; $\|\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w})\|^2_{\mathbf{C}_r} \equiv (\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w}))'\mathbf{C}_r^{-1}(\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w}))$ and $\mathbf{C}_r$ is the covariance matrix of the rth Gaussian.

Now, from Eqns. (1,14), we have

$$\mathbf{s}_{xr}^{(ekev)}(\mathbf{w}) = \mathbf{C}_r^{\frac{1}{2}}(\mathbf{P}_r\mathbf{d}(\mathbf{w}) + \mathbf{q}_r) \,, \tag{17}$$

where $\mathbf{P}_r \in \mathbb{R}^{n_1 \times n}$ consists of the $((r-1)n_1+1)$th to (rn_1)th rows of $\mathbf{P}$, and $\mathbf{q}_r = -\mathbf{P}_r\mathbf{d}_0 + \bar{\mathbf{y}}_r$. Substituting Eqn. (17) into $Q(\mathbf{w})$ and differentiating the result w.r.t. w_m, we have

$$\frac{\partial Q}{\partial w_m} = \sum_{r=1}^{R}\sum_{t=1}^{T} \gamma_t(r)(\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w}))'\mathbf{C}_r^{-1}\frac{\partial \mathbf{s}_{xr}^{(ekev)}(\mathbf{w})}{\partial w_m} \,. \tag{18}$$

From Eqn. (17), we get

$$\frac{\partial \mathbf{s}_{xr}^{(ekev)}(\mathbf{w})}{\partial w_m} = \mathbf{C}_r^{\frac{1}{2}}\left(\mathbf{P}_r\frac{\partial \mathbf{d}(\mathbf{w})}{\partial w_m} + \mathbf{q}_r\right) \,. \tag{19}$$

Finally, substituting the results of Eqns. (19, 13) into Eqn. (18), the derivative of $Q(\mathbf{w})$ w.r.t. each eigenvoice weight w_m can be readily obtained. The gradient of Eqn (18) is nonlinear in $\mathbf{w}$ and there is no closed form solution for the optimal $\hat{\mathbf{w}}$. As in KEV adaptation, the Gradient Ascent algorithm is used to search for the optimal eigenvoice weights.

3. Experimental Evaluation

The proposed embedded kernel eigenvoice (eKEV) adaptation method was evaluated on the TIDIGITS speech corpus [9]. There are 163 speakers (of both genders) in each of its standard training set and test set.

3.1. Acoustic Models

Twelve MFCCs and the normalized energy were extracted from each speech frame of 25 ms at every 10 ms. Each of the 11 digit models was a strictly left-to-right HMM comprising 16 states and one Gaussian with diagonal covariance per state. Thus, the dimension of the acoustic vectors is $n_1 = 13$ and that of the speaker supervector space n_2 is $11 \times 16 \times 13 = 2288$. In addition, there were a 3-state "sil" model and a 1-state "sp" model to capture silence speech and pauses between digits respectively. Furthermore, the SD HMMs shared the transition probabilities and Gaussian variances learned in the SI HMMs.

3.2. Experiments

Supervised adaptation was carried out using 5, 10, and 20 digits, which correspond to an average of 2.1s, 4.1s, and 9.6s of adaptation speech. To improve the statistical reliability of the results, all results were the average of 5-fold cross-validation over all 163 test speakers.

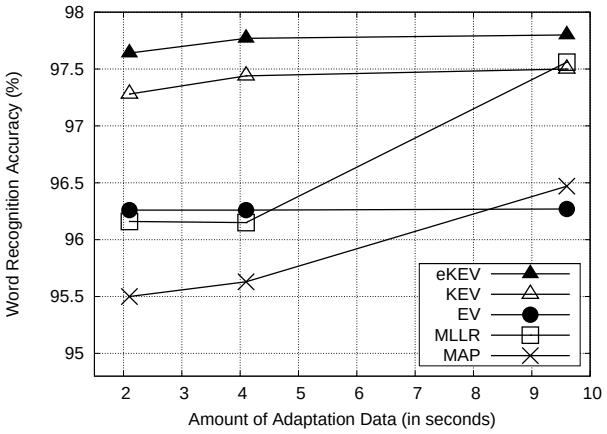

Figure 2: Performance comparison among MLLR, MAP, EV, KEV, and eKEV adaptation methods using diagonal covariance Gaussians. (Accuracy of the baseline SI model is 96.25%.)

3.2.1. Parameters Initialization

Since numerical method was used to find the optimal eigenvoice weights, proper initialization of various system parameters can be important for its success. Various system parameters are initialized as follows; details can be found in [10].

- The initial eigenvoice weights were the projections of the SI model onto the corresponding kernel eigenvoices.
- $\beta_r = \beta = 0.0005$ for $r = 1, \ldots, R$.
- The learning rate for Gradient Ascent was 0.0001.
- The number of kernel eigenvoices M was fixed to 7.
- The 5 training speakers that give the highest likelihood of the adaptation speech were chosen as the reference speakers.

Experiment 1: Comparison with other adaptation methods

The following models/systems are compared: the baseline speaker-independent (SI) model, eKEV, KEV, EV, MAP, and MLLR adaptation methods. The speaker adapted (SA) models of the three EV-based methods were interpolated with the SI model as described in [4]. The best setting for each adaptation method was empirically found to give the best results for comparison.

The word accuracy of the baseline SI model on the test data is 96.25%, and the comparison results are plotted in Figure 2. We have the following observations:

- KEV and eKEV adaptation outperform all other methods when the amount of adaptation data is less than 10s.
- Both KEV and eKEV adaptation outperform the standard EV adaptation. It shows that nonlinear kernel PCA using composite kernels finds more effective eigenvoices.
- eKEV adaptation is consistently better than KEV adaptation by (relatively) 12–13%.
- eKEV adaptation reduces the word error rate of the SI model by 37.0%, 40.5%, and 41.3% respectively for 2.1s, 4.1s, and 9.6s of adaptation speech.

Experiment 2: Speed Comparison

The main objective of eKEV adaptation is to improve the speed of adaptation and recognition of KEV adaptation. Thus, all KEV and eKEV adaptation experiments were run on a Pentium III 1GHz machine with 512MB RAM, and their adaptation and decoding times were recorded. It was observed that the adaptation speed of eKEV adaptation is indeed an order of magnitude faster than that of KEV adaptation. (The exact speedup factors by eKEV adaptation over KEV adaptation are 6.24, 8.75, and 14.5 for 2.1s, 4.1s, and 9.6s of adaptation speech respectively.) Furthermore, on average, KEV adapted models took 227s to recognize one second of test speech, while eKEV adapted models — regular HMMs — only took 1.67s; that is, a speed up of 136 times.

4. Conclusions

In this paper, we propose an improvement to our previous KEV speaker adaptation called *embedded kernel eigenvoice (eKEV) speaker adaptation*. We show that by embedding the kernel PCA procedure in KEV adaptation and finding an approximate pre-image of the speaker-adapted model in the kernel-induced feature space, we may eliminate all online kernel evaluations involving the adaptation or testing observations. In the TIDIGITS adaptation task, we obtained a speedup of 6–14 times in adaptation and 136 times in recognition for eKEV adaptation over KEV adaptation with a performance improvement of (relatively) 12–13%.

5. Acknowledgements

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST6195/02E, HKUST6201/02E, and CA02/03.EG04.

6. References

[1] 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, no. 2, pp. 291–298, April 1994.

[2] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Journal of CSL*, vol. 9, pp. 171–185, 1995.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on SAP*, vol. 8, no. 4, pp. 695–707, Nov 2000.

[4] J. T. Kwok, B. Mak, and S. Ho, "Eigenvoice speaker adaptation via composite kernel PCA," in *NIPS 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds., Cambridge, MA, 2004, MIT Press.

[5] B. Mak, J. T. James, and S. Ho, "A study of various composite kernels for kernel eigenvoice speaker adaptation," in *Proc. of ICASSP*, Montreal, Canada, 2004.

[6] B. Schölkopf and A.J. Smola, *Learning with Kernels*, MIT, 2002.

[7] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[8] J. T. Kwok and I. W. Tsang, "The pre-image problem in kernel methods," in *Proc. of the 20th ICML*, Washington, D.C., USA, August 2003, pp. 408–415.

[9] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. of ICASSP*, 1984, vol. 3, pp. 4211–4214.

[10] B. Mak and S. Ho, "Various reference speakers determination methods for embedded kernel eigenvoice speaker adaptation." (Submitted to ICSLP 2004)