A STUDY OF VARIOUS COMPOSITE KERNELS FOR KERNEL EIGENVOICE SPEAKER ADAPTATION

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

Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{mak,jamesk,csho}@cs.ust.hk

ABSTRACT

Eigenvoice-based methods have been shown to be effective for fast speaker adaptation when the amount of adaptation data is small, say, less than 10 seconds. In traditional eigenvoice (EV) speaker adaptation, *linear* principal component analysis (PCA) is used to derive the eigenvoices. Recently, we proposed that eigenvoices found by *nonlinear kernel PCA* could be more effective, and the eigenvoices thus derived were called *kernel eigenvoices* (KEV). One of our novelties is the use of *composite kernel* that makes it possible to compute state observation likelihoods via kernel functions. In this paper, we investigate two different composite kernels: direct sum kernel and tensor product kernel for KEV adaptation. In an evaluation on the TIDIGITS task, it is found that KEV speaker adaptation using both forms of composite kernel are equally effective, and they outperform a speaker-independent model and the adapted models from EV, MAP, or MLLR adaptation using 2.1s and 4.1s of speech. For example, with 2.1s of adaptation data, KEV adaptation outperforms the speaker-independent model by 27.5%, whereas EV, MAP, or MLLR adaptation are not effective at all.

1. INTRODUCTION

It is commonly known that a well-trained speaker-dependent (SD) model generally achieves a significantly lower word error rate than a speaker-independent (SI) model on recognizing speech from the specific speaker. For many applications such as phone services, it is hard to acquire a large amount of data from a user to train his/her SD model. A common technique to approach the SD performance is to adapt the SI model with a relatively small amount of SD speech using speaker adaptation methods. Adaptation methods like the Bayesian-based *maximum a posteriori* (MAP) adaptation [1] and the transformation-based *maximum likelihood linear regression* (MLLR) adaptation [2] have been popular for many years. Nevertheless, when the amount of available adaptation speech is really small — for example, only a few seconds, the more recent eigenvoice-based adaptation method is found particularly more effective. The (original) eigenvoice (EV) adaptation method [3] was motivated by the eigenface approach in face recognition [4]. The idea is to derive a small set of basis vectors called *eigenvoices* that are believed to represent different voice characteristics (e.g. gender, age, accent, etc.), and each individual speaker is then a point in the eigenspace. The simple algorithm was later extended to work for large-vocabulary continuous speech recognition [5, 6, 7]. In practice, a few to a few tens of eigenvoices are found adequate for fast speaker adaptation. Since the number of estimation parameters is greatly reduced, fast adaptation using EV is possible with a few seconds of speech.

At the heart of eigenvoice-based adaptation methods is the principal component analysis (PCA) employed to find the eigenvoices. Then a new speaker is represented as a linear combination of a few (most important) eigenvoices. Traditionally, these eigenvoices are found by linear PCA. Recently, we investigated the use of nonlinear *kernel PCA* [8] to find the eigenvoices using a *composite kernel*, and the eigenvoices thus derived were called "*kernel eigenvoices*" [9]. In a pilot study on the TIDIGITS task [10], compared with an SI model, our kernel eigenvoice method reduced the word error rate WER) by 27.5% using 2.1 seconds of adaptation speech while conventional eigenvoice approach could only match the performance of the SI model.

In this paper, we generalize the definition of composite kernels and investigate KEV adaptation with two different composite kernels: direct sum kernel and tensor product kernel. In additional, we also compare the performance of KEV adaptation with that of EV, MAP, and MLLR adaptation methods.

2. KERNEL EIGENVOICE ADAPTATION (KEV)

Let's first review the conventional eigenvoice (EV) adaptation procedure, and then point out the differences between EV and KEV adaptation.

2.1. Eigenvoice (EV) Adaptation

The conventional EV adaptation is computed as follows:

STEP 1. Train a set of speaker-dependent (SD) models.

STEP 2. For each SD model, concatenate all its mean vectors into a speaker supervector.

STEP 3. Perform linear PCA on the supervectors using their correlation matrix.

STEP 4. Arrange the eigenvectors in descending order of their eigenvalues and pick the top M eigenvectors; they are the required eigenvoices, $\boldsymbol{e}_j, j = 1, \ldots, M$.

0-7803-8484-9/04/$20.00 ©2004 IEEE

STEP 5. A new speaker's supervector $\boldsymbol{s}$ is represented by a linear combination of the M chosen eigenvoices:

$$\boldsymbol{s} = \sum_{j=1}^{M} w_j \cdot \boldsymbol{e_j} \; .$$

STEP 6. Estimate the eigenvoice weights by maximizing the likelihood of the adaptation data. Mathematically, one finds the eigenvoice weight vector $\mathbf{w}$ by *maximizing* the following Q_b function:

$$\begin{aligned} Q_b(\mathbf{w}) \;\; &= \;\; -\frac{1}{2}\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_t(r)\left[d_1 \log(2\pi)\right. \\ & \qquad \left. + \log|\mathbf{C}_r| + \|\mathbf{o}_t - \mathbf{s}_r(\mathbf{w})\|^2_{\mathbf{C}_r}\right] , \end{aligned} \qquad (1)$$

where $\gamma_t(r)$ is the posterior probability of observation $\mathbf{o}$ being at state r at time t; d_1 is the dimension of acoustic vectors; $\mathbf{s}_r(\mathbf{w})$ is the Gaussian mean vector of state r of the speaker-adapted (SA) model; $\mathbf{C}_r$ is the covariance matrix of the Gaussian at state r; and $\|\mathbf{o}_t - \mathbf{s}_r(\mathbf{w})\|^2_{\mathbf{C}_r} = (\mathbf{o}_t - \mathbf{s}_r(\mathbf{w}))'\mathbf{C}_r^{-1}(\mathbf{o}_t - \mathbf{s}_r(\mathbf{w}))$.

KEV uses nonlinear kernel PCA in Step 3. Subsequently, the formulation in Step 5 and Step 6 have to be modified.

2.2. Kernel Principal Component Analysis (KPCA)

The basic idea of kernel PCA [8] is to map data $\mathbf{x}$ in an *input space* $\mathcal{X}$ to a high-dimensional *kernel-induced feature space*[1] $\mathcal{F}$ via some nonlinear map φ, and apply linear PCA in the feature space. The computational procedure depends only on the inner products $\varphi(\mathbf{x}_i)'\varphi(\mathbf{x}_j), \forall i, j$ which are obtained from a suitable kernel function $k(\cdot,\cdot)$ as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)'\varphi(\mathbf{x}_j) \; . \qquad (2)$$

Notice that the input space $\mathcal{X}$ consists of speaker supervectors in our application. Given a set of N input speaker supervectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, let us denote the mean of the φ-mapped feature vectors by $\bar{\varphi} = \frac{1}{N}\sum_{i=1}^{N}\varphi(\mathbf{x}_i)$, and the "centered" map by $\tilde{\varphi}$ (with $\tilde{\varphi}(\mathbf{x}) = \varphi(\mathbf{x}) - \bar{\varphi}$). Eigendecomposition is performed on $\tilde{\mathbf{K}}$, the centered version of the kernel matrix $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{ij}$, as $\tilde{\mathbf{K}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$, where $\mathbf{U} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N]$ with $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iN}]'$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_N)$. Notice that $\tilde{\mathbf{K}}$ is related to $\mathbf{K}$ by $\tilde{\mathbf{K}} = \mathbf{HKH}$, where $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'$ is the centering matrix.

It is shown that the mth orthonormal eigenvector $\mathbf{v}_m$ of the covariance matrix in $\mathcal{F}$ is a linear combination of the $\tilde{\varphi}$-mapped feature vectors, and is given by [8] as

$$\mathbf{v}_m = \sum_{i=1}^{N}\frac{\alpha_{mi}}{\sqrt{\lambda_m}}\tilde{\varphi}(\mathbf{x}_i) \; . \qquad (3)$$

2.3. Composite Kernel

One of the major challenges in KEV adaptation is to compute the state observation likelihoods of the speaker-adapted HMMs during the estimation of the kernel eigenvoice weights and subsequent decoding of test speech. The reason is that unlike the conventional EV approach, the SA model found by KEV adaptation does not exist in the input supervector space $\mathcal{X}$ but in the kernel-induced feature space $\mathcal{F}$. Thus, in general, one cannot break up the SA model found by KEV adaptation into its constituent HMM Gaussians as in the EV approach. Our solution is the use of a composite kernel.

Firstly, since each speaker supervector is the result of concatenation of R mean vectors, one from each Gaussian, the ith speaker supervector will be denoted as $\mathbf{x}_i = [\mathbf{x}'_{i1} \ldots \mathbf{x}'_{iR}]' \in \mathbb{R}^{d_2}$, and $d_2 = Rd_1$. Then we map each constituent $\mathbf{x}_{ir}$ via a separate kernel $k_r(\cdot,\cdot)$ to $\varphi_r(\mathbf{x}_{ir})$, and construct $\varphi(\mathbf{x}_i)$ as $\varphi(\mathbf{x}_i) = [\varphi_1(\mathbf{x}_{i1})', \ldots, \varphi_R(\mathbf{x}_{iR})']'$. The similarity between two speaker supervectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the composite kernel-induced feature space $\mathcal{F}$ is measured by

$$k(\mathbf{x}_i, \mathbf{x}_j) = G(k_r(\mathbf{x}_{ir}, \mathbf{x}_{jr}), r = 1, \ldots, R) \qquad (4)$$

where G is some function that combines the constituent kernels $k_r(\cdot,\cdot)$ into a valid composite kernel $k(\cdot,\cdot)$. Using this composite kernel, we can then proceed with the usual kernel PCA on the set of N training speaker supervectors and obtain the set of eigenvoices in the feature space $\mathcal{F}$ as given by Eqn (3) in Section 2.2.

2.3.1. *Two Different Composite Kernels*

In this paper, we investigate two different forms of G for the composite kernel:

Direct sum kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{R} k_r(\mathbf{x}_{ir}, \mathbf{x}_{jr}) \; . \qquad (5)$$

Tensor product kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \prod_{r=1}^{R} k_r(\mathbf{x}_{ir}, \mathbf{x}_{jr}). \qquad (6)$$

Furthermore, if the constituent kernels are Gaussian kernels

$$k_r(\mathbf{x}_{ir}, \mathbf{x}_{jr}) = \exp(-\beta\|\mathbf{x}_{ir} - \mathbf{x}_{jr}\|^2_{\mathbf{C}_r}) \; , \qquad (7)$$

then the tensor product kernel is equivalent to a single Gaussian kernel with a block-diagonal covariance composed of the covariances from each $k_r(\mathbf{x}_{ir}, \mathbf{x}_{jr})$.

In both cases, if $k_r(\cdot,\cdot)$'s are valid Mercer kernels, so is $k(\cdot,\cdot)$ [12].

2.4. New Speaker in the Feature Space

If the supervector of a new speaker in the input space $\mathcal{X}$ is $\mathbf{s}$, then its centered image in the kernel-induced feature space $\mathcal{F}$ is $\tilde{\varphi}(\mathbf{s})$, which is assumed to be a linear combination of the first M eigenvectors found by KPCA in $\mathcal{F}$. i.e.

$$\tilde{\varphi}(\mathbf{s}) = \sum_{m=1}^{M} w_m \mathbf{v}_m = \sum_{m=1}^{M}\sum_{i=1}^{N}\frac{w_m \alpha_{mi}}{\sqrt{\lambda_m}}\tilde{\varphi}(\mathbf{x}_i). \qquad (8)$$

[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 latter will always be called "acoustic feature space", while the feature space in kernel methods will be simply called the 'feature space" but may be sometimes called the *kernel-induced feature space* if additional clarity is necessary.

Its rth constituent is then given by

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

Hence, the similarity between $\varphi_r(\mathbf{s}_r)$ and $\varphi_r(\mathbf{o}_t)$ is given by

$$\begin{aligned} k_r(\mathbf{s}_r, \mathbf{o}_t) &\equiv \varphi_r(\mathbf{s}_r)' \varphi_r(\mathbf{o}_t) \\ &= A(r,t) + \sum_{m=1}^{M} \frac{w_m}{\sqrt{\lambda_m}} B(m,r,t) \; , \end{aligned} \tag{9}$$

where

$$A(r,t) = \bar{\varphi}_r' \varphi_r(\mathbf{o}_t) = \frac{1}{N} \sum_{j=1}^{N} k_r(\mathbf{x}_{jr}, \mathbf{o}_t) \; , \tag{10}$$

and

$$B(m,r,t) = \left(\sum_{i=1}^{N} \alpha_{mi} k_r(\mathbf{x}_{ir}, \mathbf{o}_t) \right) - A(r,t) \sum_{i=1}^{N} \alpha_{mi} \; . \tag{11}$$

2.5. Maximum Likelihood Adaptation Using Gaussian Kernel

By using a composite kernel and Gaussian kernels of Eqn (7) for the constituent kernels, $\|\mathbf{o}_t - \mathbf{s}_r\|^2_{\mathbf{C}_r}$ of Eqn (1) can be expressed as a function of $\mathbf{w}$ as follows:

$$\begin{aligned} \|\mathbf{o}_t - \mathbf{s}_r\|^2_{\mathbf{C}_r} &= -\frac{1}{\beta} \log k_r(\mathbf{s}_r, \mathbf{o}_t) \\ &= -\frac{1}{\beta} \log \left[A(r,t) + \sum_{m=1}^{M} \frac{w_m}{\sqrt{\lambda_m}} B(m,r,t) \right] . \end{aligned} \tag{12}$$

Substituting Eqn (12) for the Q_b function in Eqn (1), and differentiating the result with respect to each eigenvoice weight, $w_j, j = 1, \ldots, M$, we obtain

$$\frac{\partial Q_b}{\partial w_j} = \frac{1}{2\beta\sqrt{\lambda_j}} \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) \cdot \frac{B(j,r,t)}{k_r(\mathbf{s}_r, \mathbf{o}_t)} . \tag{13}$$

Because of the nonlinear nature of kernel PCA, Eqn (13) is nonlinear in $\mathbf{w}$ and there is no closed form solution for the optimal $\mathbf{w}$. We instead apply the generalized EM algorithm (GEM) [11] to find the optimal weights.

2.6. Robust KEV Adaptation

When the amount of adaptation data is really small, the SA model $\tilde{\varphi}(\mathbf{s})^{(kev)}$ found by KEV adaptation may not be reliable. In robust KEV adaptation, the model $\tilde{\varphi}(\mathbf{s})^{(kev)}$ is interpolated with the SI model $\tilde{\varphi}(\mathbf{x}^{(si)})$ in the kernel-induced feature space to compute a more robust estimate of the final SA model as follows:

$$\tilde{\varphi}(\mathbf{s}) = w_0 \tilde{\varphi}(\mathbf{x}^{(si)}) + (1 - w_0) \tilde{\varphi}(\mathbf{s})^{(kev)} \; , \; 0 \le w_0 \le 1 \; . \tag{14}$$

The interpolation weight w_0 is estimated jointly by GEM together with the other eigenvoice weights.

3. EXPERIMENTAL EVALUATION

The kernel eigenvoice adaptation method was evaluated on the TIDIGITS speech corpus [10]. There are 163 speakers (of both genders) in each of its standard training set and test set. The speaker characteristics is quite diverse with speakers coming from 22 dialect regions of USA and their ages ranging from 6 to 70 years old.

3.1. Acoustic Models

All training data were processed to extract 12 MFCCs and the normalized frame energy from each speech frame of 25 ms at every 10 ms. Each of the eleven 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 observation space d_1 is 13 and that of the speaker supervector space d_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. All HMMs were trained by the EM algorithm. Furthermore, the SD HMMs shared the transition probabilities and Gaussian variances learned in the SI HMMs.

3.2. Experiments

The following models/systems are compared:

SI: speaker-independent model.

EV: speaker-adapted model found by the conventional eigenvoice adaptation method.

Robust-EV: speaker-adapted models found by our robust version of EV, which is the interpolation between the SI supervector and the supervector found by EV.

KEV: speaker-adapted model found by our new kernel eigenvoice adaptation method as described in Section 2.

Robust-KEV: speaker-adapted model found by our robust KEV as described in Section 2.6.

MAP: speaker-adapted model found by MAP adaptation.

MLLR: speaker-adapted model found by MLLR adaptation.

Five, ten, and twenty digits were used for adaptation, which correspond to an average of 2.1s, 4.1s, and 9.6s of adaptation speech (or 3s, 5.5s, and 13s of speech if the leading and ending silences are counted). To improve the statistical reliability of the results, all results were the average of 5-fold cross-validation over all 163 test speakers. Moreover, all adaptation experiments were performed in supervised mode.

The best results from each of the adaptation methods are compared. For EV or KEV adaptation, the best results were obtained with the optimal number of eigenvoices; for MAP adaptation, the best results were achieved with the optimal scaling factors; for MLLR adaptation, only global MLLR was tried, and the better results from using block-diagonal or full transformation matrices were used for comparison. The word accuracy of the baseline SI model on the test data is 96.25%[2].

3.2.1. Experiment I: Direct Sum Kernel vs. Tensor Product Kernel

We first compare the two types of composite kernels, direct sum kernel and tensor product kernel, using the robust KEV adaptation. The results are shown in Table 1. There is no significant difference between their performance.

[2] Notice that the word accuracy of our SI model is lower than the best reported result on TIDIGITS which is about 99.7%. The main reasons are that we used only 13-dimensional static cepstra and energy, and each state was modeled by a single Gaussian with diagonal covariance. The use of this simple model allowed us to run experiments with 5-fold cross-validation using very short adaptation speech. Right now our approach requires online computation of many kernel function values and is very computationally expensive. As a first attempt on the approach, we feel that the use of this simple model is justified. We are now working on its speed-up and its extension to HMM states of Gaussian mixtures.

Table 1. Performance of direct sum kernel and tensor product kernel in robust KEV adaptation. Results are word accuracies.

Type of Composite Kernel	2.1s	4.1s	9.6s
direct sum kernel	97.28%	97.44%	97.50%
tensor product kernel	97.33%	97.42%	97.43%

3.2.2. Experiment II: KEV vs. EV, MAP, MLLR

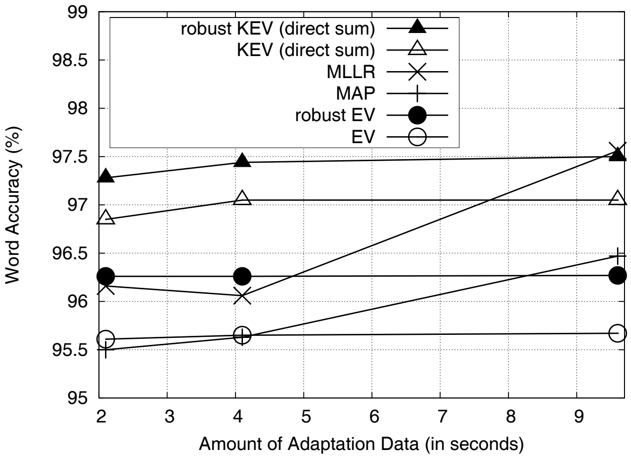

Fig. 1. Comparison among EV, KEV, MAP, and MLLR adaptation methods. All results are word accuracies, and the accuracy of the baseline SI model is 96.25%.

Fig. 1 compares the performance of other adaptation methods with our KEV adaptation using the direct sum kernel. We find that when only 2.1s or 4.1s of adaptation data are available, only our new KEV and robust KEV work better than the SI model; EV, MAP, and MLLR all perform worse than the SI model, and robust EV can only match the SI performance in this task. Only for the case with 10 seconds of adaptation data, then MLLR works marginally better than the robust KEV method by an absolute 0.06%.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate the use of two types of composite kernels — direct sum kernel and tensor product kernel — to improve the conventional eigenvoice speaker adaptation method using nonlinear kernel PCA. In the TIDIGITS task, it is found that both composite kernels work similarly well, and while the conventional eigenvoice approach does not help, our robust kernel eigenvoice method outperforms the speaker-independent model by (in terms of error rate reduction) 27.5%, 31.7%, and 33.3% with 2.1s, 4.1s, and 9.6s of adaptation speech respectively.

Right now, our KEV adaptation method results in slower recognition. The reason is that any state observation likelihoods cannot be directly computed but through evaluating the kernel values with all training supervectors. We are pursuing two possible solutions: (1) reduce the number of kernel functions to compute by the application of sparse kernel PCA [13], or the use of compactly supported kernels [14]; (2) compute an approximate pre-image (which will be a speaker supervector residing in the input space) of the speaker-adapted model found in the kernel-induced feature space [15].

5. ACKNOWLEDGEMENTS

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST6201/02E, CA02/03.EG04, HKUST2033/00E, and HKUST6195/02E.

6. REFERENCES

[1] J.L. Gauvain and C.H. Lee, ‘Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Transactions on Speech and Audio Processing*, 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 Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, ‘Rapid Speaker Adaptation in Eigenvoice Space,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 695–707, Nov 2000.

[4] M. Turk and A. Pentland, ‘Face recognition using eigenface,” in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1991, pp. 586–591.

[5] R. Kuhn, F. Perronnin, P. Nguyen, J. C. Junqua, and L. Rigazio, ‘Very Fast Adaptation with a Compact Context-Dependent Eigenvoice Model,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[6] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, ‘Fast Speaker Adaptation using Eigenspace-based Maximum Likelihood Linear Regression,” in *Proceedings of the International Conference on Spoken Language Processing*, 2000, vol. 3, pp. 742–745.

[7] B. Zhou and J. Hansen, “A Novel Algorithm for Rapid Speaker Adaptation Based on Structural Maximum Likelihood Eigenspace Mapping,” in *Proceedings of the European Conference on Speech Communication and Technology*, Aalborg, Denmark, Sept 2001, vol. 2, pp. 1215–1218.

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

[9] J. T. Kwok, B. Mak, and S. Ho, ‘Eigenvoice Speaker Adaptation via Composite Kernel PCA,” in *Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec 2003.

[10] R. G. Leonard, “A Database for Speaker-Independent Digit Recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1984, vol. 3, pp. 4211–4214.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, ‘Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

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

[13] A.J. Smola, O.L. Mangasarian, and B. Schölkopf, ‘Sparse kernel feature analysis,” Tech. Rep. 99-03, Data Mining Institute, University of Wisconsin, Madison, 1999.

[14] M.G. Genton, ‘Classes of kernels for machine learning: A statistics perspective,” *Journal of Machine Learning Research*, vol. 2, pp. 299–312, 2001.

[15] J. T. Kwok and I. W. Tsang, “The Pre-Image Problem in Kernel Methods,” in *Proceedings of the 20th International Conference on Machine Learning*, Washington, D.C., USA, August 2003, pp. 408–415.