VARIOUS REFERENCE SPEAKERS DETERMINATION METHODS FOR EMBEDDED KERNEL EIGENVOICE SPEAKER ADAPTATION

Brian Mak and *Simon Ho*

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

ABSTRACT

Recently, we proposed two improvements to the eigenvoice (EV) speaker adaptation using kernel methods: *kernel eigenvoice (KEV) speaker adaptation*, and *embedded kernel eigenvoice (eKEV) speaker adaptation*. In both KEV and eKEV adaptation methods, kernel eigenvoices are computed using kernel PCA, and an implicit speaker adapted model is defined as a linear combination of the leading kernel eigenvoices in the kernel-induced feature space. eKEV adaptation further finds an approximate *pre-image* of the implicit speaker adapted model so that all online kernel evaluations involving any acoustic vectors are eliminated during adaptation and subsequent recognition. The pre-image finding algorithm is cast as a constrained optimization problem using the distances between the expected pre-image and a set of pre-determined *reference speakers* as constraints. In this paper, we investigate two different ways to determine the reference speakers and the effect of their numbers on the eKEV adaptation performance.

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 two improvements to the EV adaptation called *kernel eigenvoice (KEV) speaker adaptation* [4, 5] and *embedded kernel eigenvoice (eKEV) speaker adaptation* [6], which exploit possible nonlinearity in the speaker supervector space using kernel methods [7]. The basic idea is to map speaker supervectors to a high dimensional feature space 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*. Then a new speaker's adapted model is constructed *implicitly* as a linear combination of the leading kernel eigenvoices in the *feature space*. eKEV adaptation further projects the implicit speaker adapted (SA) model back to the input speaker supervector space by finding an approximate *pre-image* of the SA model. Thus, unlike KEV adaptation which only produces an implicit SA model in the kernel-induced feature space, eKEV adaptation produces an *explicit* speaker supervector for the adapting speaker so that all online kernel evaluations involving adaptation or testing speech are eliminated. As a consequence, both adaptation and recognition speeds of eKEV adaptation are faster than that of KEV adaptation. In fact, the recognition speed of eKEV adaptation is as fast as normal HMM decoding. In an TIDIGITS adaptation task, eKEV adaptation was shown to outperform a speaker-independent model by about 40% using less than 10s of adaptation speech, and was better than EV, KEV, MAP, and MLLR adaptation [6].

In eKEV adaptation, the finding of the pre-image of the speaker adapted model in the kernel-induced feature space is cast as a constrained optimization problem using the distances between the expected pre-image and a set of pre-determined *reference speakers* as constraints, and the optimal pre-image is solved in the least-square sense[1]. This paper investigates two ways to define the set of reference speakers as well as the effect of their numbers on the performance of eKEV adaptation.

2. REVIEW OF THE EMBEDDED KERNEL EIGENVOICE SPEAKER ADAPTATION (EKEV)

The eKEV adaptation method is illustrated in Fig. 1. In the figure, all the five training speakers are used to derive eigenvoices in the kernel-induced feature space by kernel PCA. The implicit speaker-adapted model $\varphi(\mathbf{s}_x^{(ekev)})$ is restricted to the kernel eigenvoice subspace of the feature space, and its (approximate) pre-image $\mathbf{s}_x^{(ekev)}$ is found by using its distance constraints from a set of three reference speakers $\mathbf{x}_1 - \mathbf{x}_3$.

The procedure is outlined briefly step-by-step as follows; the details can be found in [6].

STEP 1: Construction of Speaker Supervectors

Suppose there are N speaker-dependent (SD) hidden Markov models (HMMs) of the same topology with R mixture Gaussians. For each speaker, say, the ith speaker, a *speaker supervector* $\mathbf{x}_i$ is constructed by concatenating all his HMM Gaussian mean vectors

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

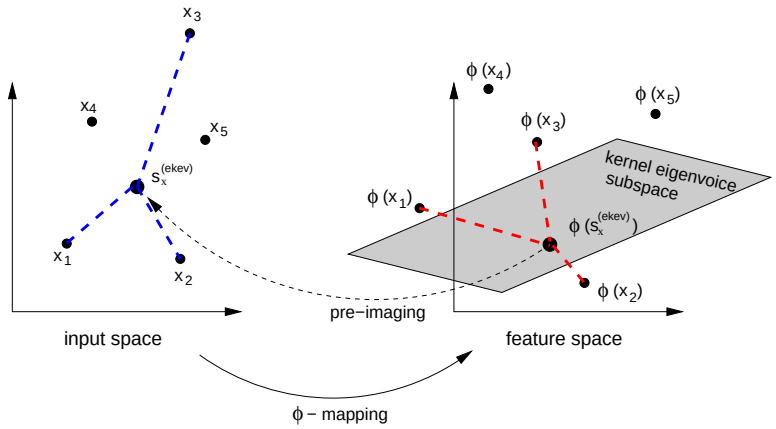

Fig. 1. The eKEV adaptation method. (Without the pre-imaging step, it is the KEV adaptation method.)

$\mathbf{x}_{ir} \in \mathbb{R}^{n_1}, r = 1, \ldots, R$. That is, $\mathbf{x}_i = [\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iR}]' \in \mathbb{R}^{n_2}$ and $n_2 = n_1 R$.

STEP 2: Variance Normalization

Normalize each constituent of any speaker supervector $\mathbf{x}$ by its own covariance. The normalized model of $\mathbf{x}$ is represented by $\mathbf{y} = \mathbf{C}^{-\frac{1}{2}}\mathbf{x}$ where

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

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

STEP 3: Computation of Kernel Eigenvoices

Let's use the following direct sum composite kernel

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

which is associated with a mapping φ that maps $\mathbf{y}$ in the normalized input speaker supervector space $\mathcal{Y}$ to $\varphi(\mathbf{y})$ in the kernel-induced feature space $\mathcal{F}$. Compute the centered kernel matrix $\tilde{\mathbf{K}}$ with $\tilde{\mathbf{K}}_{ij} = \tilde{\varphi}(\mathbf{y}_i)'\tilde{\varphi}(\mathbf{y}_j)$ (where $\tilde{}$ is used to represent a quantity centered around its centroid throughout this paper). Kernel PCA is 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 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$. Then the mth orthonormal kernel eigenvoice in $\mathcal{F}$ is given by [8] as $\mathbf{v}_m = \sum_{i=1}^{N} \frac{\alpha_{mi}}{\sqrt{\lambda_m}} \tilde{\varphi}(\mathbf{y}_i) , \quad m = 1, \ldots, M.$

STEP 4: Similarity between the New Speaker and a Training Speakers in the Feature Space

The adapted speaker model $\tilde{\varphi}(\mathbf{s}_y^{(ekev)})$ is a linear combination of the M leading kernel eigenvoices in $\mathcal{F}$. That is,

$$\tilde{\varphi}(\mathbf{s}_y^{(ekev)}) = \sum_{m=1}^{M} w_m \mathbf{v}_m = \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{w_m \alpha_{mi}}{\sqrt{\lambda_m}} \tilde{\varphi}(\mathbf{y}_i) . \qquad (2)$$

And its rth constituent is given by

$$\tilde{\varphi}_r(\mathbf{s}_{yr}^{(ekev)}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{w_m \alpha_{mi}}{\sqrt{\lambda_m}} \tilde{\varphi}_r(\mathbf{y}_{ir}) . \qquad (3)$$

Hence, we get

$$k_r(\mathbf{s}_{yr}^{(ekev)}, \mathbf{y}_{jr}) = A_r(j) + \sum_{m=1}^{M} \frac{w_m}{\sqrt{\lambda_m}} B_r(m, j) , \qquad (4)$$

where

$$A_r(j) = \frac{1}{N} \sum_{i=1}^{N} k_r(\mathbf{y}_{ir}, \mathbf{y}_{jr}) , \qquad (5)$$

and

$$B_r(m, j) = \sum_{i=1}^{N} \alpha_{mi} \left(k_r(\mathbf{y}_{ir}, \mathbf{y}_{jr}) - A_r(j) \right) . \qquad (6)$$

STEP 5: 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*. Singular value decomposition (SVD) of the centered $\mathbf{Y}$ gives

$$\tilde{\mathbf{Y}} = \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{V}' = \mathbf{U}_2 \mathbf{Z} , \qquad (7)$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ and its columns, say, $\mathbf{z}_i$, are the projections of $\mathbf{y}_i$ onto the eigenvectors of $\mathbf{U}_2$. Collect $\|\mathbf{z}_i\|^2, i = 1, \ldots, n$, into an n-dimensional vector,

$$\mathbf{d}_0 = [\|\mathbf{z}_1\|^2, \|\mathbf{z}_2\|^2, \ldots, \|\mathbf{z}_n\|^2]' \in \mathbb{R}^n . \qquad (8)$$

STEP 6: Finding the Distance Constraints between the New Speaker and the Reference Speakers in the Input Space

Let d_j be the squared Euclidean distance between $\mathbf{s}_y^{(ekev)}$ and $\mathbf{y}_j$ in the input space, and each constituent kernel be a Gaussian kernel, then we have

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

$$\Rightarrow \quad d_j = -\sum_{r=1}^{R} \frac{1}{\beta_r} \log k_r(\mathbf{s}_{yr}^{(ekev)}, \mathbf{y}_{jr}) . \qquad (9)$$

These distances are collected into the vector

$$\mathbf{d}(\mathbf{w}) = [d_1, d_2, \ldots, d_n]' \in \mathbb{R}^n \ . \quad (10)$$

STEP 7: Finding the Distance Gradients

Differentiating $\mathbf{d}$ of Eqn. (9) w.r.t. w_m, we get

$$\frac{\partial d_j}{\partial w_m} = -\frac{1}{\sqrt{\lambda_m}} \sum_{r=1}^{R} \frac{B_r(m,j)}{\beta_r k_r(\mathbf{s}_{yr}^{(ekev)}(\mathbf{w}), \mathbf{y}_{jr})} \ , j = 1, \ldots, n \ . \quad (11)$$

STEP 8: Finding the Pre-image

From [9], the optimal pre-image that satisfies the distance constraints in $\mathbf{d}$ in the least-square sense is given by

$$\mathbf{s}_x^{(ekev)}(\mathbf{w}) = \mathbf{C}^{\frac{1}{2}} \mathbf{s}_y^{(ekev)} = \mathbf{C}^{\frac{1}{2}} (\mathbf{P}\mathbf{d}(\mathbf{w}) + \mathbf{q}) \ , \quad (12)$$

where

$$\mathbf{P} = -\frac{1}{2} \mathbf{U}_2 \mathbf{\Lambda}_2^{-1} \mathbf{V}' \text{ and } \mathbf{q} = -\mathbf{P}\mathbf{d}_0 + \bar{\mathbf{y}} \ . \quad (13)$$

STEP 9: ML Estimation of Kernel Eigenvoice Weights

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

$$Q(\mathbf{w}) = -\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) \|\mathbf{o}_t - \mathbf{s}_{xr}^{(ekev)}(\mathbf{w})\|_{\mathbf{C}_r}^2 \ , \quad (14)$$

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.

Differentiating $Q(\mathbf{w})$ w.r.t. each eigenvoice weight and using the distance gradients of Eqn. (11), the derivatives of $Q(\mathbf{w})$ can be readily obtained. These derivatives are nonlinear in $\mathbf{w}$ and there is no closed form solution for the optimal $\hat{\mathbf{w}}$. the Gradient Ascent algorithm is used to search for the optimal eigenvoice weights instead.

3. DIFFERENT METHODS TO DETERMINE THE REFERENCE SPEAKERS

The computation of the pre-image relies on its distances to a set of reference speakers. In the reference paper of the pre-image finding method [9], the neighbors of a de-noised image in the feature space are used as the reference speakers. However, in our problem, the whereabout of the speaker-adapted (SA) model is not known beforehand — neither in the kernel-induced feature space, nor in the input supervector space — and so are the locations of its neighbors. Here, we investigate two ways to determine the initial set of reference speakers of the SA model to be found.

3.1. SI Neighbors

If there is no additional information, it is reasonable to start with the neighbors of the speaker-independent (SI) model since the adaptation starts its search from the SI model. The neighbors can be computed using the Euclidean or Mahalanobis distances. One advantage of using SI neighbors is that since they are fixed, they can be computed once beforehand and applied to all adapting speakers.

3.2. Maximum Likelihood (ML) Neighbors

Conceptually, since we are using the ML for determining the SA model, it should be close to those training speakers that also have high likelihood of the adaptation data. Since the ML neighbors are dependent on the adaptation speech, they vary across adaptation sessions and must be computed online.

3.3. Number of Reference Speakers

Another issue about the reference speakers is how many of them are adequate. The current pre-image finding algorithm uses the distances from the reference speakers of a neighborhood to exploit localized information to constrain the solution space. If there are too few reference speakers, the distance constraints may be too weak to lead to a good pre-image solution. However, if too many reference speakers are included, those that are far away will dominate the distance constraints, and weaken the localized information for the determination of the pre-image.

4. EXPERIMENTAL EVALUATION

The proposed two reference speakers determination methods for eKEV adaptation method were evaluated on the TIDIGITS speech corpus [10]. There are 163 speakers in each of its standard training set and test set.

4.1. 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 with a single diagonal covariance Gaussian 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.

The word accuracy of the baseline SI model on the test data is 96.25%[2].

Table 1. Effect of different reference speakers (#neighbors = 10).

Amount of Data	ML Neighbors	SI Neighbors	
		Euclidean	Mahalanobis
2.1s	**97.41**	96.33	96.52
4.1s	**97.53**	96.43	96.60
9.6s	**97.58**	96.50	96.68

[2]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 many experiments with very short adaptation speech. We are now working on its extension to HMM states of Gaussian mixtures.

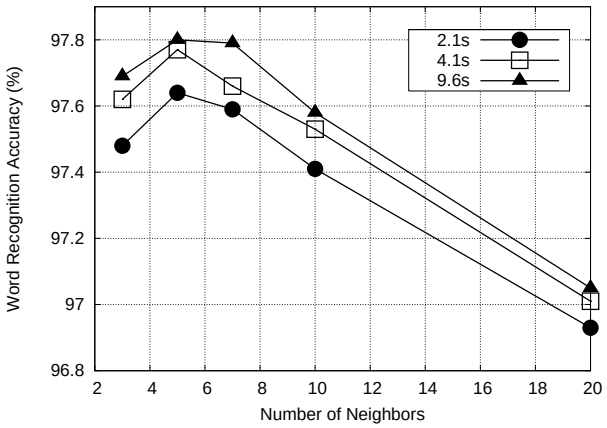

Fig. 2. Effect of the number of reference speakers (ML neighbors)

4.2. 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 statistical reliability of the results, all results were the average of 5-fold cross-validation over all 163 test speakers.

Parameters Initialization

The eKEV adaptation method employs the iterative Gradient Ascent algorithm to compute the (locally) optimal eigenvoice weights in each maximization step of the GEM algorithm. Some system parameters were initialized as follows:

- 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.

Experiment 1: Reference Speakers Determination Methods

The results of different types of reference speakers on the performance of eKEV adaptation method is shown in Table 1. The number of neighbors were fixed to 10 for this investigation. From the results, we find that the use of ML neighbors outperforms the use of SI neighbors by about 1% absolute or 25% relative. It indeed seems that the final SA model is closer to its ML neighbors than the SI neighbors. Since there can be many local maxima in the solution of the gradient method, we hypothesize that a good initialization of its neighborhood to the ML neighbors may have avoided the method from being trapped in a poorer local maximum around the SI neighbors. We had run additional iterations and updated the ML neighbors of the SA model found to its actual neighbors (as determined by the Mahalanobis distances), but it was found that most of the neighbors remained unchanged, and the final model had very similar performance as the SA model obtained without the neighbor updates.

Experiment 2: Number of Reference Speakers

Fig. 2 shows the performance of eKEV adaptation using different numbers of ML neighbors as the reference speakers. It is observed that for this problem, 3 to 8 ML neighbors give good results, and the best performance is obtained with 5 ML neighbors. In practice, the optimal number of reference speakers may be determined by cross-validation. It is also good to know that a small number of reference speakers is adequate as this will mean less computation during eKEV adaptation as well.

5. CONCLUSIONS

In this paper, we investigate the use of SI neighbors and ML neighbors to determine the set of reference speakers needed in the pre-image finding algorithm of our new eKEV speaker adaptation method. The former are neighbors of the speaker-independent (SI) model, and the latter are those training speakers that have the highest likelihoods of the adaptation data. In a TIDIGITS adaptation task, eKEV adaptation using ML neighbors outperform eKEV adaptation using SI neighbors by about 1% absolute or 25% relative. It is also found that the number of reference speakers affects the results. Too many reference speakers are not recommended; instead it is better to use a few speakers of good localized constraints (e.g. ML neighborhood).

6. 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.

7. 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. MIT Press, Cambridge, MA, 2004.

[5] B. Mak, J. T. Kwok, and S. Ho, "A study of various composite kernels for kernel eigenvoice speaker adaptation," in *Proc. of ICASSP*, Montreal, Canada, 2004, vol. I, pp. 325–328.

[6] B. Mak, S. Ho, and J. T. Kwok, "Speedup of kernel eigenvoice speaker adaptation by embedded kernel PCA," in *Proc. of ICSLP*, Jeju Island, South Korea, 2004.

[7] B. Schölkopf and A.J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

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

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