

Kernel Eigenvoices (Revisited) for Large-Vocabulary Speech Recognition

Zoi Roupakia* and Mark Gales, *Fellow, IEEE*

Abstract—Kernelised eigenvoice methods, which apply a nonlinear transform in speaker space, have previously been proposed for rapid adaptation. This paper examines and addresses a number of limitations and issues with the current schemes. First, the requirements for valid probability functions using kernel representations are discussed. Second, rapid speaker adaptation using these forms of representations is analysed and the general update formulae for kernelised eigenvoice adaptation derived. The existing kernelised eigenvoice methods are then described within this formulation. This allows an EM-based, rather than gradient-descent-based, parameter estimation. To enable these approaches to be applied to large-vocabulary speech recognition tasks, eigen-bases using transformations of an underlying canonical model are described and related to existing adaptation methods. Preliminary experiments on a large-vocabulary conversational telephone speech task are finally detailed.

Index Terms—kernel adaptation, eigenvoices, adaptive training

I. INTRODUCTION

State-of-the-art speech recognition systems achieve high accuracy when they are evaluated on speakers and in environments similar to those seen in the training data. However, the performance of these systems degrades as the mismatch between the training speakers and environments and those in test increases. To address this problem, speaker adaptation approaches have been proposed [1], [2], [3]. These approaches normally apply transformations to reduce the mismatch between the training and testing conditions given some adaptation data. When there is limited training data, linear transformations are often used, either in the form of interpolating a set of bases, as in Eigenvoices [2] and Cluster Adaptive Training (CAT) [3], or as linear transform of a canonical acoustic model, as in maximum likelihood linear regression (MLLR)-based schemes [1], [4], [5].

Though linear transform-based schemes have proved to yield performance gains, the intra- and inter-speaker variability is not expected to be linear. To address this, kernelised versions of eigenvoices and MLLR have been introduced to enable nonlinear transformations using kernel functions [6], [7]. This paper revisits these kernelised approaches. The requirements for valid probability functions using kernel representations are discussed, along with approaches for estimating CAT-like speaker-dependent interpolation weights using EM-based schemes with these forms of representations. Existing forms

of kernel adaptation are then discussed in the same framework [6], [7]. In these current schemes, the final distributions are not guaranteed to be valid, when the most general form is used. When the form is restricted to ensure valid distributions, the standard gradient-descent-based optimisation schemes are unnecessary. This paper shows that EM-based schemes can be used ensuring that the final distribution is a valid probability density function.

Another issue is how to scale these kernel adaptation methods to large vocabulary speech recognition systems. In [6], [7] speaker dependent models or transforms are built to train the eigenvoices and eigenmatrices respectively. For thousands of speakers, however, this is impractical. This paper addresses this problem with the use of compact representations of parameters.

II. KERNEL REPRESENTATIONS OF DISTRIBUTIONS

In this section, probability distributions are described in terms of a general kernel representation. This will enable nonlinear adaptation using kernelised versions of linear transforms. Consider a mapping Φ (linear or nonlinear) from an original vector space \mathcal{X} to a real inner product feature space \mathcal{F} [8]

$$\Phi : \mathbf{x}_t \in \mathbb{R}_{\mathcal{X}}^d \rightarrow \Phi(\mathbf{x}_t) \in \mathbb{R}_{\mathcal{F}}^f. \quad (1)$$

In this work, the general kernel representation of probabilistic functions is expressed by the inner product of the mapped observation vector $\Phi(\mathbf{x}_t)$ and the distributional parameters $\boldsymbol{\theta}_m^\Phi \in \mathbb{R}_{\mathcal{F}}^f$ (of the component m) in the embedding space \mathcal{F}

$$p(\mathbf{x}_t|m) = \frac{1}{Z_m} \langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_m^\Phi \rangle, \quad (2)$$

where Z_m is the normalisation constant and the operator $\langle \cdot, \cdot \rangle$ denotes the inner product in a real vector space.

In order to be a valid distribution, the standard stochastic constraints must be satisfied

$$\int_{-\infty}^{+\infty} \frac{1}{Z_m} \langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_m^\Phi \rangle d\mathbf{x}_t = 1, \quad (3)$$

$$\frac{1}{Z_m} \langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_m^\Phi \rangle \geq 0. \quad (4)$$

The normalisation constant $Z_m = \int_{-\infty}^{+\infty} \langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_m^\Phi \rangle d\mathbf{x}_t$ must, hence, be bounded. Both the inner product and the normalisation constant of (4) must be either negative or positive. Without loss of generality, this paper will assume: $\langle \Phi(\mathbf{x}_t), \boldsymbol{\theta}_m^\Phi \rangle \geq 0$ and $Z_m > 0$. Kernel representations of distributions defined as in (2), thus, establish the use of real inner products and, consequently, the use of kernel functions in probability density functions in a mathematical framework.

Zoi Roupakia and Mark Gales are with Cambridge University Engineering Department, Trumpington St., Cambridge CB2 1PZ. U.K.
E-mail: zr216@cam.ac.uk and mjfg@eng.cam.ac.uk

EDICS: SPE-RECO Speech and speaker recognition. Manuscript received August, 2011. Manuscript revised September, 2011.

III. KERNEL SPEAKER ADAPTATION

In speaker adaptation, the model parameters of a speaker s , θ_{sm}^Φ , are obtained by applying a transformation of the model parameters θ_m^Φ given the adaptation data. Thus, the adapted observation likelihood has the form

$$p(\mathbf{x}_t|m, s) = \frac{1}{Z_m^{(s)}} \langle \Phi(\mathbf{x}_t), \theta_{sm}^\Phi \rangle. \quad (5)$$

The form of speaker adaptation approaches may be required to operate in a high-dimensional space. Linear-transform-based approaches, such as MLLR [1], are impractical as the size of the parameter space becomes very large.

In this paper, the following form of adaptation is used

$$\theta_{sm}^\Phi = \sum_{e=1}^E \lambda_e^{(s)} \theta_{em}^\Phi, \quad (6)$$

where $\lambda_e^{(s)}$ are the speaker-specific adaptation parameters. The parameters for the new speaker are then given by the linear combination of some basis model parameters θ_{em}^Φ , similar to CAT [3] or eigenvoices [2]. The bilinear property of inner products [8] implies that for real Hilbert spaces and for this particular form of transformation, the observation likelihood, as defined by (5), is itself an interpolation of inner products

$$p(\mathbf{x}_t|m, s) = \frac{1}{Z_m^{(s)}} \sum_{e=1}^E \lambda_e^{(s)} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle. \quad (7)$$

This form of kernel adaptation is required to satisfy the constraints from the previous section. Thus, from (3),

$$\sum_{e=1}^E \frac{\lambda_e^{(s)}}{Z_m^{(s)}} \int_{-\infty}^{+\infty} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle d\mathbf{x}_t = 1. \quad (8)$$

Considering a normalisation constant $Z_m^{(e)}$ for each of the inner products $\langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle$ yields the following constraint

$$\sum_{e=1}^E \frac{\lambda_e^{(s)} Z_m^{(e)}}{Z_m^{(s)}} \int_{-\infty}^{+\infty} \frac{1}{Z_m^{(e)}} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle d\mathbf{x}_t = 1. \quad (9)$$

The integral in (9) with appropriate normalisation $Z_m^{(e)}$ is equal to one and the constraints are then

$$\sum_{e=1}^E \frac{\lambda_e^{(s)} Z_m^{(e)}}{Z_m^{(s)}} = 1 \Rightarrow \sum_{e=1}^E \tilde{\lambda}_e^{(s)} = 1 \quad (10)$$

$$\frac{\lambda_e^{(s)} Z_m^{(e)}}{Z_m^{(s)}} \geq 0 \Rightarrow \tilde{\lambda}_e^{(s)} \geq 0 \quad (11)$$

From the previous section, $Z_m^{(e)}$ and $Z_m^{(s)}$ must be positive and $\langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle$ non-negative. Thus, both $\lambda_e^{(s)}$ and $\tilde{\lambda}_e^{(s)}$ must also be non-negative. It should be noted that estimating $\tilde{\lambda}_e^{(s)}$ instead of $\lambda_e^{(s)}$ implies no loss of generality.

A. Parameter Estimation

During adaptation, the new speaker is defined as a ‘‘point’’ in speaker space by estimating E speaker dependent weights $\lambda_e^{(s)}$, or equivalently $\tilde{\lambda}_e^{(s)}$. As only E weights are estimated, the scheme is suitable for rapid adaptation when little data is

available. The likelihood of HMM at state q at time t for the components m , which belong to state q , is

$$p(\mathbf{x}_t|q, s) = \sum_{m \in q} c_m \sum_{e=1}^E \frac{\tilde{\lambda}_e^{(s)}}{Z_m^{(e)}} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle. \quad (12)$$

Using EM, the associated auxiliary function is

$$Q(\hat{\Lambda}; \tilde{\Lambda}) = \sum_{t=1}^T \sum_{m=1}^M \sum_{e=1}^E \gamma_{em}^{(s)}(t) \log \frac{c_m \tilde{\lambda}_e^{(s)}}{Z_m^{(e)}} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle, \quad (13)$$

where $\hat{\Lambda}$ is the set of new parameters, $\tilde{\lambda}_e^{(s)}$ the new weights and

$$\gamma_{em}^{(s)}(t) = \frac{c_m \tilde{\lambda}_e^{(s)}}{Z_m^{(e)}} \langle \Phi(\mathbf{x}_t), \theta_{em}^\Phi \rangle. \quad (14)$$

By taking the derivative and imposing a sum-to-one constraint with Lagrange multipliers, the adaptation parameters can be estimated by using

$$\hat{\lambda}_e^{(s)} = \frac{\sum_{t=1}^T \sum_{m=1}^M \gamma_{em}^{(s)}(t)}{\sum_{t=1}^T \sum_{m=1}^M \sum_{e=1}^E \gamma_{em}^{(s)}(t)}. \quad (15)$$

Thus, with this update formula, an EM scheme for finding the interpolation weights can be defined. Each iteration is guaranteed to not decrease the likelihood and to yield a valid probability distribution. It is worth noting that this scheme is valid for any form of kernel distribution representation. If the kernel is Gaussian, then the resulting model is a Gaussian mixture model with $M \times E$ components.

B. Relation to Kernel Eigenvoice Adaptation

Kernel eigenvoices [6] can be expressed in terms of a kernel distribution and speaker adaptation technique. In [6], only the mean parameters are adapted to the new speaker s ; the adapted mean μ_{sm}^Φ is defined as a point in a high-dimensional space spanned by E kernel eigenvoices μ_{em}^Φ , $e = 1, \dots, E$. The eigen-bases are obtained by applying kernel PCA to the centered mapped supervectors $\tilde{\mu}_k^\Phi \in \mathbb{R}^{M^d}$, which are formulated by concatenating the means of K speaker dependent M -component Gaussian mixture models. Thus,

$$\mu_{sm}^\Phi = \sum_{e=1}^E \lambda_e^{(s)} \tilde{\mu}_{em}^\Phi + \frac{1}{K} \sum_{k=1}^K \Phi(\mu_{km}), \quad (16)$$

$$\tilde{\mu}_{em}^\Phi = \sum_{k=1}^K \left(\beta_k^{(e)} - \frac{1}{K} \sum_{j=1}^K \beta_j^{(e)} \right) \Phi(\mu_{km}), \quad (17)$$

where $\beta_k^{(e)}$ are derived from eigen-decomposition of the centred Gram matrix. The form of kernel distribution in [6] can be written in a form similar to (7)¹

$$\begin{aligned} p(\mathbf{x}_t|m, s) &= \frac{1}{Z_m^{(s)}} \left(\langle \Phi(\mathbf{x}_t), \mu_{sm}^\Phi \rangle \right)^{1/\sigma_m} \\ &= \frac{1}{Z_m^{(s)}} \left(\sum_{k=1}^K \zeta_k^{(s)} \langle \Phi(\mathbf{x}_t), \Phi(\mu_{km}) \rangle \right)^{1/\sigma_m} \end{aligned} \quad (18)$$

¹There is no explicit preimage assumption for the adapted mean, only preimage assumption for the basis means $\Phi(\mu_{km})$. However, in [6], the likelihood function is derived using $\langle \Phi(\mathbf{x}_t), \mu_{sm}^\Phi \rangle = \exp(-\sigma_m \|\mathbf{x}_t - \mu_{sm}\|_{\Sigma_m}^2)$ that implies $\mu_{sm}^\Phi = \Phi(\mu_{sm})$. Though not directly used, this implicit assumption is not necessary.

where

$$\zeta_k^{(s)} = \sum_{e=1}^E \lambda_e^{(s)} \beta_k^{(e)} + \frac{1 - \sum_{e=1}^E \lambda_e^{(s)} \sum_{j=1}^K \beta_j^{(e)}}{K} \quad (19)$$

and the form of kernel function is

$$\langle \Phi(\mathbf{x}_t), \Phi(\boldsymbol{\mu}_{km}) \rangle = \exp(-\sigma_m \|\mathbf{x}_t - \boldsymbol{\mu}_{km}\|_{\Sigma_m}^2), \quad (20)$$

σ_m is the kernel-width. It is straightforward to show that $\sum_{k=1}^K \zeta_k^{(s)} = 1$. However, it is not necessary that $\zeta_k^{(s)} \geq 0$ without restricting the range of $\lambda_e^{(s)}$.

The eigenvoice weights are estimated in [6] with gradient descent methods using the derivative of the auxiliary function

$$\frac{\partial Q(\hat{\Lambda}; \Lambda)}{\partial \lambda_e^{(s)}} = \sum_{t=1}^T \sum_{m=1}^M \gamma_m^{(s)}(t) \left(\frac{\partial \log Z_m^{(s)}}{\partial \lambda_e^{(s)}} + \frac{\sum_{k=1}^K (\beta_k^{(e)} - \frac{1}{K} \sum_{j=1}^K \beta_j^{(e)}) \langle \Phi(\mathbf{x}_t), \Phi(\boldsymbol{\mu}_{km}) \rangle}{\sigma_m \langle \Phi(\mathbf{x}_t), \boldsymbol{\mu}_{sm}^\Phi \rangle} \right). \quad (21)$$

The normalisation constant $Z_m^{(s)}$ in equation (18) is a function of the component and speaker parameters. In [6], an approximate normalisation term, $Z_m^{(s)} \approx (2\pi^d |\Sigma_m|)^{1/2}$, is used. This ignores the dependence of the normalisation term on the point in speaker-space, but simplifies the estimation of the derivative, as the first term in (21) disappears. Moreover, this approximate normalisation term has ignored additional dependencies on the component parameters; thus, it will also impact the recognition result. This approximation is also not considered in [6].

An interesting case is when $\sigma_m = 1$. In this case, the normalisation term can be simply written as $Z_m^{(s)} = (2\pi^d |\Sigma_m|)^{1/2}$. Thus, the form used in [6] both for estimating the point in speaker-space and the likelihoods are exact. However, as $\zeta_k^{(s)}$ is not restricted to be non-negative, the constraints in (10) and (11) are not satisfied.

Given the constraint that $\sigma_m = 1$, the form of likelihood expression in (18) has exactly the same form as (7) with a Gaussian kernel distribution representation. It is therefore possible to use the EM optimisation detailed in the previous section to find the point (interpolation weights) in the speaker-space, $\zeta_k^{(s)}$. Optimising $\zeta_k^{(s)}$, and satisfying the sum-to-one and non-negative constraints ensures that the probability function is valid and each EM iteration will not decrease the likelihood. It is also possible to optimise the values of $\beta_k^{(e)}$ based on EM, as likelihood (18) may also be expressed as a mixture model in terms of $\beta_k^{(e)}$. In this work, the form investigated will use the constraint $\sigma_m = 1$ and have a Gaussian kernel distribution. This will be referred to as kernel eigenvoice adaptation (KEA).

IV. COMPACT REPRESENTATIONS FOR LVCSR

One issue related to applying kernel eigenvoice adaptation to large vocabulary speech recognition systems is that the number of model parameters can become very large as the number of components, M , and eigen-bases, E , increase. If each of the components of each basis is kept distinct, then the number of model parameters will scale as $E \times M$. To address this issue, a compact representation of the model parameters for each of the basis is required. The approach adopted here

is similar to the transform-based CAT formulation [3]. Here, linear transformations of an underlying canonical model are used as the basis representations.

In this work, a form of representation related to constrained MLLR (CMLLR) is investigated for kernel eigenvoice adaptation. The basis mean for e , $\boldsymbol{\mu}_{em}$, has the form

$$\boldsymbol{\mu}_{em} = \mathbf{A}_e^{-1} \boldsymbol{\mu}_m - \mathbf{A}_e^{-1} \mathbf{b}_e, \quad (22)$$

where $\boldsymbol{\mu}_m$ is the canonical mean and the form of the inner product used in kernel representation is given by

$$\langle \Phi(\mathbf{x}_t), \Phi(\boldsymbol{\mu}_{em}) \rangle = \exp(-\|\mathbf{x}_t - \boldsymbol{\mu}_{em}\|_{\mathbf{A}_e^{-1} \Sigma_m \mathbf{A}_e^{-T}}^2) \quad (23)$$

Σ_m is the canonical covariance matrix, and the normalisation constant $Z_m^{(e)} = \frac{|\mathbf{A}_e|}{(2\pi^d |\Sigma_m|)^{1/2}}$. In the form used here, there is a preimage for the basis parameters such that $\boldsymbol{\mu}_{\Phi}^{em} = \Phi(\boldsymbol{\mu}_{em})$. It is possible to generalise this using an initial projection (the equivalent of the KPCA projection in kernel eigenvoices). It should be noted that, as before, there is no requirements that the preimage of the adapted speaker $\boldsymbol{\theta}_{sm}^\Phi$ exists.

By rearranging the terms in (23), this is equivalent to the following observation likelihood

$$p(\mathbf{x}_t | m, s) = \sum_{e=1}^E \tilde{\lambda}_e^{(s)} |\mathbf{A}_e| \mathcal{N}(\mathbf{A}_e \mathbf{x}_t + \mathbf{b}_e; \boldsymbol{\mu}_m, \Sigma_m). \quad (24)$$

The number of model parameters to be trained reduces from $M \times E \times 2d$ parameters for $M \times E$ distinct means and diagonal variances to $M \times 2d + E \times d^2$ for full CMLLR transforms per basis in the proposed scheme (24). Parameter estimation is performed using the auxiliary function discussed in section III. As transform, model and eigenvoice weight parameters cannot be estimated simultaneously, an iterative solution similar to cluster adaptive training [3] is applied. The update formulae to estimate the transforms and canonical models are small modifications to the forms presented in [4].

The form of kernel eigenvoice adaptation (24) is related to other adaptation schemes. In cluster adaptive training [3] (CAT) the new speaker is defined as the linear interpolation of cluster means $\boldsymbol{\mu}_{em}$. For transform-based CAT

$$p(\mathbf{x}_t | m, s) = \mathcal{N}(\mathbf{x}_t; \sum_{e=1}^E \lambda_e^{(s)} (\mathbf{A}_e \boldsymbol{\mu}_m + \mathbf{b}_e), \Sigma_m). \quad (25)$$

This form of adaptation operates in the model-space. The KEA scheme described here can be viewed as interpolation in the likelihood space.

More closely related is Maximum likelihood stochastic transforms [9] (MLST). The CMLLR form of MLST in [9] is

$$p(\mathbf{x}_t | m, s) = \sum_{e=1}^E \lambda_e^{(s)} \mathcal{N}(\mathbf{x}_t; (\mathbf{A}_e^{(s)} \boldsymbol{\mu}_m + \mathbf{b}_e^{(s)}), \mathbf{A}_e^{(s)} \Sigma_m \mathbf{A}_e^{(s)T}). \quad (26)$$

Both the transforms and weights are estimated from the adaptation data (though the option of estimating these from training data is briefly mentioned). In KEA, full transforms are estimated from training data. In CMLLR-MLST, diagonal transformations are used and no adaptive training of the canonical model is described. Although MLST form is closely related to KEA, the motivation for it is very different. KEA is

motivated from a kernel distribution and adaptation perspective. The form in (24) is simply one example with a Gaussian kernel representation and the use of linear transforms to limit the number of model parameters. MLST is motivated from an extension to standard linear transform adaptation approaches. In common with MLST, but not CAT, the decoding cost increases (linearly in terms of Gaussian calculations) with the number of bases, E . This issue is not addressed in this paper.

V. EXPERIMENTS

Kernel eigenvoice adaptation was evaluated on a large vocabulary conversational telephone speech (CTS) task. A 76-hour training dataset (h5etrainsub) containing 1118 speakers and 77201 utterances in three corpora was used. A 3-hour subset of the 2001 development data for CTS (dev01sub) was used for evaluation. This test data has 59 speakers (30 female, 29 male) and 2663 utterances. The feature vectors were 12-dimensional MF-PLP with c_o energy with first, second and third derivatives. Cepstral mean and variance normalisation was applied. The data were projected to a 39 dimensional space by an HLDA transform. Vocal tract length normalisation (VTLN) was also used. Note for consistency with previous systems VTLN, and mean and variance normalisation, were estimated on a speaker level. This limits the potential gains from adaptation, particularly for utterance-level rapid adaptation. All acoustic models were state-clustered triphone HMMs with 12 components per state, maximum likelihood trained. The supervision hypotheses for estimating the adaptation parameters were generated using the SI model.

A number of KEA systems and contrast systems were built. Two forms of standard cluster-based systems were trained. First, a gender-dependent (GD) system (constructed using MAP adaptation of the SI model) was used, note the gender labels were obtained from the references. Second, a 2-cluster CAT system was built (CAT). In addition, CMLLR adaptation was used to adapt the SI and GD model (SI, GD+CMLLR). The baseline KEA system (SI+KEA) was constructed using an iterative transform splitting approach, similar to the iterative mixture splitting scheme used in HTK. The transform biases were perturbed and the system trained updating both the points in speaker-space for the training speakers and the transforms. The number of CMLLR transforms (hence number of eigenbases) was increased to 32. Moreover, adaptive training of the canonical was performed (SI+KEA+SAT). The final KEA system was based on the GD canonical models with gender-specific basis (GD+KEA). Again for this system the reference gender labels were used. Adaptation was either run at the utterance or speaker level. The average amount of data per speaker was 180 s; per utterance level, the average is around 4 s, with average minimum of 0.8 s and maximum of 11 s.

Table I shows the performance of the baseline systems. Both the GD and CAT systems show gains over the SI system. SI+CMLLR adaptation gave gains over the SI, GD and CAT systems for speaker-level adaptation. However, there was not sufficient data at the utterance level to robustly estimate the transforms. Transform priors could have been used to improve the robustness, but this was not investigated in this work.

TABLE I
WER-PERFORMANCE OF ADAPTATION APPROACHES

| Sys | WER | |
|----------|------|-------|
| | Spkr | Utter |
| SI | 34.6 | |
| +CMLLR | 33.0 | 34.4 |
| +CAT | 34.2 | 34.3 |
| +KEA | 33.3 | 33.2 |
| +KEA+SAT | — | 32.9 |
| GD | 34.2 | |
| +CMLLR | — | 34.3 |
| +KEA | — | 32.7 |

SI+KEA gave similar results to speaker level SI+CMLLR, but were robust for both speaker and utterance level; hence, KEA is suitable for adaptation in cases of little data availability.

Table I also shows the performance of the SAT and GD KEA systems on the utterance levels. Both yielded gains over the baseline KEA system. The best performance was using GD+KEA, which gave a gain of 1.9% absolute over the baseline SI system, in contrast to GD+CMLLR, which can not be robustly trained from utterance-level data.

VI. CONCLUSION

This paper has examined kernel eigenvoice adaptation (KEA). The general requirements for kernel representations of distributions have been described, and ways for incorporating speaker adaptation into this framework have been analyzed. An EM-based estimation scheme for the speaker parameters has also been detailed. The form of distribution representation described in standard KEA uses an approximation of the required normalisation, which can impact both the estimation of the speaker parameters and inference. Restricting the form of distributional representation addresses this problem. Furthermore, the parameters of the point in speaker-space can then be estimated using EM, rather than gradient-based approaches. Initial results on a conversational telephone speech task show that the modified version of KEA can be used for rapid adaptation on large vocabulary speech recognition tasks.

REFERENCES

- [1] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, 1995.
- [2] R. Kuhn, P. Nguyen, J.-C. Jungua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP'98*, 1998.
- [3] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," in *IEEE Transactions on Speech and Audio Processing*, 2000.
- [4] —, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, 1998.
- [5] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions Speech and Audio Processing*, 1995.
- [6] B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," in *IEEE Transactions on Speech and Audio Processing*, 2005.
- [7] B. Mak and R. Hsiao, "Kernel eigenspace-based MLLR adaptation," in *IEEE Transactions on Audio, Speech and Language Processing*, 2007.
- [8] N. Young, *An introduction to Hilbert space*. Cambridge University Press, 1998.
- [9] V. D. Diakouloulas and V. V. Digalakis, "Maximum-likelihood stochastic transformation adaptation of hidden Markov models," in *IEEE Transactions on Speech and Audio Processing*, 1999.