FAST SPEAKER ADAPTION VIA MAXIMUM PENALIZED LIKELIHOOD KERNEL REGRESSION

Ivor W. Tsang, James T. Kwok, Brian Mak, Kai Zhang and Jeffrey J. Pan

Department of Computer Science
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

ABSTRACT

Maximum likelihood linear regression (MLLR) has been a popular speaker adaptation method for many years. In this paper, we investigate a generalization of MLLR using nonlinear regression. Specifically, kernel regression is applied with appropriate regularization to determine the transformation matrix in MLLR for fast speaker adaptation. The proposed method, called *maximum penalized likelihood kernel regression* adaptation (MPLKR), is computationally simple and the mean vectors of the speaker adapted acoustic model can be obtained analytically by simply solving a linear system. Since no nonlinear optimization is involved, the obtained solution is always guaranteed to be globally optimal. The new adaptation method was evaluated on the Resource Management task with 5s and 10s of adaptation speech. Results show that MPLKR outperforms the standard MLLR method.

1. INTRODUCTION

Current speaker adaptation methods fall into one of the following three categories: speaker-clustering-based methods [1], Bayesian-based methods such as the *maximum a posteriori* (MAP) adaptation [2], and transformation-based methods, most notably, *maximum likelihood linear regression* (MLLR) adaptation [3]. However, for many speech applications (e.g. telephone services), fast online speaker adaptation is needed as the amount of available adaptation speech can be really short — perhaps only a few seconds. There are at least three approaches for fast speaker adaptation:

- Improving from the speaker-clustering-based approach: In eigenspace-based adaptation methods [4, 7] and reference speaker weighting [5], a new speaker's model is constrained as a linear combination of a small set of eigenvectors and training speaker models chosen according to the adapting speaker respectively.
- Improving from MLLR: To further reduce the number of MLLR estimation parameters, techniques like the use of diagonal MLLR transformation [3], MAPLR [6], eigen-MLLR [7], eigenspace mapping [8], etc. have been tried with some success.
- Exploiting nonlinearity in eigenspace-based methods: The major procedure in eigenspace-based methods is linear PCA. [9, 10, 11] exploit possible nonlinearity in the speaker space by generalizing the procedure to nonlinear PCA using kernel method.

In this paper, we propose a novel approach for fast speaker adaptation called *maximum penalized likelihood kernel regression* (MPLKR) speaker adaptation by exploiting nonlinear regression of the maximum likelihood (ML) adapted mean vectors with the use of kernel methods [12]. The basic idea is to first transform the speaker-independent (SI) mean vectors to high-dimensional feature vectors via some nonlinear map φ, which are then used for linear regression with appropriate regularization. During the actual computation, the exact nonlinear map need not be known. The computational procedure depends only on the inner products of the high-dimensional feature vectors, which can be obtained efficiently with a suitable kernel function.

2. MAXIMUM LIKELIHOOD ADAPTATION

Let's consider a speaker-independent (SI) speech recognition system that uses hidden Markov models (HMMs) with a total of N Gaussian components to represent its acoustic models. Assume now there are T speech frames from a new speaker for adaptation. If the maximum likelihood (ML) criterion is used for adaptation, the ML mean vectors $\boldsymbol{\mu}_j^* \in \mathbb{R}^d, j = 1, \ldots, N$, are found by minimizing

$$\sum_{j=1}^{N} \sum_{t=1}^{T} \gamma_j(t)(\mathbf{o}_t - \boldsymbol{\mu}_j)' \mathbf{C}_j^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_j) \qquad (1)$$

w.r.t. $\boldsymbol{\mu}_j$'s, where $\mathbf{C}_j$ is a $d \times d$ covariance matrix of state j, $\mathbf{o}_t$ is the acoustic vector at frame t, and $\gamma_j(t)$ is the posterior

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

probability of state j at frame t given the T observations $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_T]$. The ML solution of $\boldsymbol{\mu}_j$ is

$$\boldsymbol{\mu}_j^* = \frac{\sum_{t=1}^T \gamma_j(t)\mathbf{o}_t}{\sum_{t=1}^T \gamma_j(t)} = \frac{\mathbf{O}\boldsymbol{\gamma}_j}{\mathbf{1}'\boldsymbol{\gamma}_j}, \tag{2}$$

where $\boldsymbol{\gamma}_j = [\gamma_j(1), \ldots, \gamma_j(T)]'$ and $\mathbf{1} = [1, \ldots, 1]' \in \mathbb{R}^T$.

In practice, because the amount of adaptation speech from a new speaker is usually very limited, this ML solution results in a poorly adapted model with poor recognition performance.

3. MAXIMUM LIKELIHOOD LINEAR REGRESSION (MLLR)

To avoid the aforementioned problem with the ML solution in Eqn. (2), MLLR [3] obtains the adapted mean vectors by linear regression of the ML means. This is achieved by constraining the $\boldsymbol{\mu}_j$'s to be a linear transformation of the augmented SI mean vector $\boldsymbol{\xi}_j = [1\ \boldsymbol{\mu}_j']' \in \mathbb{R}^{(d+1)}$; i.e., $\boldsymbol{\mu}_j = \mathbf{W}\boldsymbol{\xi}_j$, where $\mathbf{W} \in \mathbb{R}^{d\times(d+1)}$. Without loss of generality, we will assume that only a single global transformation $\mathbf{W}$ is shared among all N Gaussians.

Similar to Eqn. (1), $\mathbf{W}$ can be estimated by maximizing the likelihood of the adaptation data, or equivalently, by minimizing the following:

$$\sum_{j=1}^N \sum_{t=1}^T \gamma_j(t)(\mathbf{o}_t - \mathbf{W}\boldsymbol{\xi}_j)'\mathbf{C}_j^{-1}(\mathbf{o}_t - \mathbf{W}\boldsymbol{\xi}_j). \tag{3}$$

Comparing Eqn. (1) and Eqn. (3), we see that $\mathbf{W}\boldsymbol{\xi}_j$ in Eqn. (3) plays the role of $\boldsymbol{\mu}_j$ in Eqn. (1). As the optimal solution of $\boldsymbol{\mu}_j$ in Eqn. (1) is $\boldsymbol{\mu}_j^*$, one can see that MLLR is in effect trying to learn a $\mathbf{W}$ such that

$$\mathbf{W}\boldsymbol{\xi}_j = \boldsymbol{\mu}_j^*,\ j = 1, \ldots, N. \tag{4}$$

One may solve this linear system directly as follows. Let $\mathbf{Y} = [\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_N^*]$, and $\Xi = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N]$, so that we may rewrite Eqn. (4) as

$$\mathbf{W}[\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N] = \mathbf{W}\Xi = \mathbf{Y}. \tag{5}$$

We may also denote the row vectors of $\mathbf{Y}$ and $\mathbf{W}$ by $\mathbf{y}_i'$ and $\mathbf{w}_i'$ respectively (i.e., $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_d]'$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d]'$). Thus, we have

$$\begin{aligned} \mathbf{W}\Xi = \mathbf{Y} \quad &\Rightarrow \quad \Xi'\mathbf{W}' = \mathbf{Y}' \\ &\Rightarrow \quad \Xi'\mathbf{w}_i = \mathbf{y}_i, \quad i = 1, \ldots, d. \end{aligned} \tag{6}$$

The "best" solution that minimizes the squared error is

$$\mathbf{w}_i = (\Xi')^+\mathbf{y}_i, \tag{7}$$

where $(\Xi')^+$ denotes the pseudo-inverse of Ξ'.

4. MAXIMUM LIKELIHOOD KERNEL REGRESSION ADAPTATION (MLKR)

Let's examine the linear system in Eqn. (5) in greater detail. Recall that each mean vector has a dimension of d. Hence, we have $d \times (d+1)$ unknowns (associated with matrix $\mathbf{W}$) and a total of dN constraints. When $d+1 \geq N$, a solution can be found[1] for $\mathbf{W}$; in fact, multiple solutions can be found when $d+1 > N$. Consequently, $\boldsymbol{\mu}_j$'s obtained from MLLR are the same as the ML solution in Eqn. (2). On the other hand, when $d+1 < N$ (which is the usual case in the context of speaker adaptation), the system in Eqn. (5) is over-constrained and the solution obtained by MLLR will, in general, be different from that obtained in Eqn. (2).

For the case where $d+1 < N$, if we could introduce more variables into the linear system in Eqn. (5) so that the number of variables is greater than or equal to the number of constraints, then we would be able to get back the optimal $\boldsymbol{\mu}_j^*$'s in the ML sense. In the following, we consider first mapping $\boldsymbol{\xi}_j \in \mathbb{R}^{d+1}$ to $\varphi(\boldsymbol{\xi}_j) \in \mathbb{R}^N$. The linear system in Eqn. (5) then becomes

$$\tilde{\mathbf{W}}\varphi(\boldsymbol{\xi}_j) = \boldsymbol{\mu}_j^*,\ j = 1, \ldots, N, \tag{8}$$

or

$$\tilde{\mathbf{W}}\boldsymbol{\Phi} = [\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_N^*] = \mathbf{Y},$$

where $\tilde{\mathbf{W}}$ is of size $d \times N$ and $\boldsymbol{\Phi}_{N\times N} \equiv [\varphi(\boldsymbol{\xi}_1), \ldots, \varphi(\boldsymbol{\xi}_N)]$. The least-square solution is obtained by minimizing the Frobenius norm,

$$\|\mathbf{Y} - \tilde{\mathbf{W}}\boldsymbol{\Phi}\|_F. \tag{9}$$

On assuming that $\boldsymbol{\Phi}$ is symmetric and invertible, the solution is given by

$$\tilde{\mathbf{W}} = \mathbf{Y}\boldsymbol{\Phi}^{-1}, \tag{10}$$

and the new adapted means will be equal to those obtained in Eqn. (2), $\boldsymbol{\mu}_1^*$'s.

The mapping φ can be achieved with the use of kernels k [12], and the method will be called *maximum likelihood kernel regression* (MLKR). In particular, one may use a positive definite kernel (such as the Gaussian kernel) and the corresponding empirical kernel map[2] for φ. Then, we have

$$\boldsymbol{\Phi} = \begin{bmatrix} k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_1) & \cdots & k(\boldsymbol{\xi}_1, \boldsymbol{\xi}_N) \\ \vdots & \cdots & \vdots \\ k(\boldsymbol{\xi}_N, \boldsymbol{\xi}_1) & \cdots & k(\boldsymbol{\xi}_N, \boldsymbol{\xi}_N) \end{bmatrix} \equiv \mathbf{K}, \tag{11}$$

which is usually called the kernel matrix. From [12], we know that when a positive definite kernel is used, the kernel matrix

[1] A solution can always be found for $\mathbf{W}$ unless the system is inconsistent. From Eqn. (6), we will have an inconsistent system when for some i, $\boldsymbol{\xi}_a = \boldsymbol{\xi}_b$ but $\mathbf{y}_{ia} \neq \mathbf{y}_{ib}$ (i.e., $\boldsymbol{\mu}_{ai}^* \neq \boldsymbol{\mu}_{bi}^*$) for some a, b. For an inconsistent system, a solution can still be found by using the pseudo-inverse.

[2] For a given set $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N\}$, the empirical kernel map [12] maps a pattern $\boldsymbol{\xi}$ to $(k(\boldsymbol{\xi}_1, \boldsymbol{\xi}), \ldots, k(\boldsymbol{\xi}_N, \boldsymbol{\xi}))'$, where k is the kernel function.

in Eqn. (11) always has a full rank (assuming that the $\boldsymbol{\xi}_j$'s, $j = 1, \ldots, N$, are distinct). Hence, $\boldsymbol{\Phi}$ is symmetric and invertible, and the desired $\tilde{\mathbf{W}}$ can be obtained as in Eqn. (10). Thus, we have shown that the ML solution can *always* be obtained with MLKR when a positive definite kernel is used. In other words, although the $\tilde{\mathbf{W}}$ matrix is tied across all Gaussians, the use of kernels allows the ML solution to be perfectly recovered.

4.1. Maximum Penalized Likelihood Kernel Regression Adaptation (MPLKR)

However, as discussed in Section 2, the use of this ML solution is undesirable and so, equivalently, this MLKR solution is also not useful. From the regression perspective, linear regression used by MLLR can only capture linear characteristics in the data; on the other hand, nonlinear regression without regularization (analogous to our unregularized MLKR described above) can be overly flexible, and can attain zero training error (which is analogous to our situation here where the ML solution can be perfectly recovered) and thus suffers from over-fitting. Finally, nonlinear regression with appropriate regularization is then able to capture possible nonlinearity in the data, and at the same time, effectively control the degree of freedom. This thus leads to the *maximum penalized likelihood kernel regression* adaptation method (MPLKR).

Given that the SI model is often a fail-safe model in speech recognition, we require $\tilde{\mathbf{W}}$ to be close to the SI model transformation $\tilde{\mathbf{W}}^{(si)} = \Xi\mathbf{K}^{-1}$ (by replacing $\mathbf{Y}$ by Ξ in Eqn. (10)). Therefore, the cost function in Eqn. (9) is modified by adding the following regularizer,

$$\|\mathbf{Y} - \tilde{\mathbf{W}}\mathbf{K}\|_F^2 + \beta\|\tilde{\mathbf{W}} - \Xi\mathbf{K}^{-1}\|_F^2, \qquad (12)$$

where β is a regularization parameter. It can be shown that the solution of Eqn. (12) is given by

$$\begin{aligned}\tilde{\mathbf{W}} &= (\mathbf{Y}\mathbf{K}' + \beta\Xi\mathbf{K}^{-1})(\mathbf{K}\mathbf{K}' + \beta\mathbf{I})^{-1} \\ &= (\mathbf{Y}\mathbf{K} + \beta\Xi\mathbf{K}^{-1})(\mathbf{K}^2 + \beta\mathbf{I})^{-1}, \qquad (13)\end{aligned}$$

making use of the fact that the kernel matrix must be symmetric. Finally, the mean vectors of the new speaker-adapted (SA) model can be recovered as

$$\boldsymbol{\mu} = \tilde{\mathbf{W}}\mathbf{K} = (\mathbf{Y}\mathbf{K} + \beta\Xi\mathbf{K}^{-1})(\mathbf{K}^2 + \beta\mathbf{I})^{-1}\mathbf{K}. \qquad (14)$$

Compared to kernel-based speaker adaptation techniques such as kernel eigenvoice (KEV) [9], MPLKR has the advantage that the final mean vectors of the SA model can be computed analytically by simply solving a linear system. As no nonlinear optimization is involved, unlike KEV adaptation, the solution obtained by MPLKR adaptation is always globally optimal (w.r.t. Eqn. (12)).

5. EXPERIMENTAL EVALUATION

Maximum penalized likelihood kernel regression (MPLKR) speaker adaptation method was evaluated on the DARPA Resource Management continuous speech database RM1. RM1 consists of 3990 SI training utterances from 109 speakers, and 12 speakers in the SD section, each having 600 utterances for training, 100 utterances for development, and 100 utterances for evaluation.

5.1. Feature Extraction and Acoustic Modeling

Forty-seven context-independent phoneme models were trained using the SI training set. Each phoneme model was a strictly left-to-right 3-state hidden Markov model (HMM) with 10 Gaussian mixtures per state. In addition, there were a 1-state short pause model and a 3-state silence model. The acoustic vector has a dimension $d = 13$, consisting of 12 MFCCs and the normalized log energy extracted from speech frames of 25 ms long at the frame rate of 100Hz[3].

5.2. Experimental Procedure

Experiments were performed with either 5s or 10s adaptation data. (If we exclude the silence portion, there are about 4s or 8s of speech in the adaptation utterances.) To improve reliability of the results, for each test speaker, 3 sets of adaptation data were randomly chosen from his 100 development utterances. All reported results are the averages of experiments over the 3 adaptation sets of all speakers, and the adapted models were tested on their 100 evaluation utterances using word-pair grammar.

The following models or adaptation methods are compared:

SI: speaker-independent model;

MLLR: MLLR adaptation using either diagonal or full transformation;

EMLLR : eigenspace based MLLR adaptation [7];

MPLKR: maximum penalized likelihood kernel regression;

eKEV: embedded kernel eigenvoice [10];

KEMLLR : eigenspace based MLLR adaptation [11].

MLLR adaptation was done using the HTK software with a regression tree of 32 classes, and the best results obtained with either diagonal or full transformation is reported. Notice that, by default, HTK requires at least 700 frames of speech for each regression class; as some configurations had

[3]Notice that as a first trial, we tested our new method on smaller acoustic models so that many experiments might be run during the exploration. If we used context-dependent HMM and the conventional 39-dimensional MFCC acoustic vectors, the word accuracy was about 95% on the SD test set.

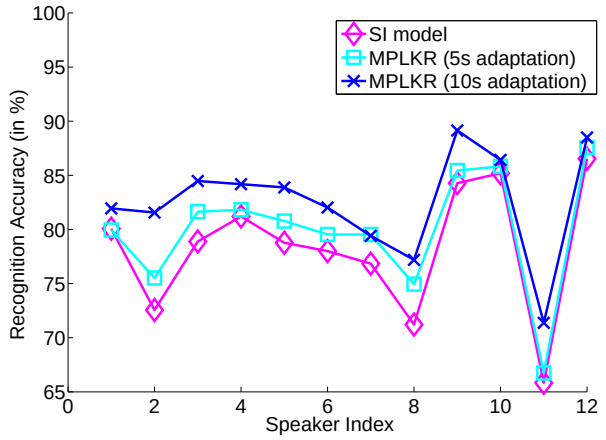

Fig. 1. Performance of MPLKR on each speaker of RM.

very few data, this threshold was lowered in order to force HTK to perform MLLR whenever the situation permits. For MPLKR, we use the following Gaussian kernel: $k(\mathbf{x}, \mathbf{y}) = \exp(\gamma\|\mathbf{x} - \mathbf{y}\|^2)$ with $\gamma = 0.01$, and β in Eqn. (12) is set to 5.0. Preliminary experiments suggest that the results are not sensitive to changes in these parameters.

Figure 1 shows the performance of MPLKR on each test speaker. It can be seen that the SA model obtained by MPLKR is consistently better than the original SI model on all speakers by about 2% and 5%. Table 1 compares its average performance with other speaker adaptation methods. It can be seen that MPLKR is the best among all the non-kernelized methods. Among the kernelized methods, the previously proposed eKEV and KEMLLR are slightly better than our new MPLKR on 5s adaptation. With 10s of adaptation speech, eKEV saturates very fast with little improvement over its 5s-adaptation performance; on the other hand, MPLKR is comparable with KEMLLR which has the best adaptation performance. Unlike KEV, eKEV or KEMLLR which are nonlinear optimization problems and are currently solved by gradient-based methods in [9, 10, 11], MPLKR has the advantage that it is not an iterative procedure, and the transformation (and hence the new model means) can be obtained analytically by solving a linear system.

Table 1. Performance of various adaptation methods on RM.

Model/Method	5s	10s
SI	78.28%	78.28%
MLLR	78.43%	82.10%
EMLLR	79.92%	80.51%
MPLKR	79.94%	82.51%
eKEV	80.58%	80.70%
KEMLLR	80.57%	82.62%

6. CONCLUSION

In this paper, we improve the standard maximum likelihood linear regression speaker adaptation method by using kernel methods to capture possible nonlinearity in the data. Computationally, the proposed method is simple and the solution can be analytically obtained by simply solving a linear system. No nonlinear optimization is involved, and thus the solution obtained here is always globally optimal. In the Resource Management task, it is found that the proposed maximum penalized likelihood kernel regression (MPLKR) adaptation method outperforms MLLR.

7. REFERENCES

[1] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," *Journal of CSL*, vol. 10, pp. 55–74, 1996.

[2] 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.

[3] 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.

[4] 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.

[5] Tim J. Hazen, "A comparison of novel techniques for rapid speaker adaptation," *Speech Communications*, vol. 31, pp. 15–33, May 2000.

[6] C. Chesta, O. Siohan, and C. H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. of Eurospeech*, 1999, vol. 1, pp. 211–214.

[7] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. of ICSLP*, 2000, vol. 3, pp. 742–745.

[8] B. Zhou and J. Hansen, "Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation," *IEEE Trans. on SAP*, vol. 13, no. 4, pp. 554–564, July 2005.

[9] B. Mak, J. T. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Trans. on SAP*, vol. 13, no. 5, pp. 984–992, September 2005.

[10] B. Mak, S. Ho, and J. T. Kwok, "Speedup of kernel eigenvoice speaker adaptation by embedded kernel PCA," in *Proc. of ICSLP*, 2004, vol. IV, pp. 2913–2916.

[11] B. Mak and R. Hsiao, "Improving eigenspace-based MLLR adaptation by kernel PCA," in *Proc. of ICSLP*, 2004, vol. I, pp. 13–16.

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