MAXIMUM A POSTERIORI LINEAR REGRESSION (MAPLR) VARIANCE ADAPTATION FOR CONTINUOUS DENSITY HMMS

Wu Chou † *and Xiaodong He* ‡

† Avaya Labs Research, 233 Mt. Airy Rd., Basking Ridge, NJ 07920
‡ CECS Department, University of Missouri, Columbia, Missouri 65211
Email: wuchou@avaya.com , xhb1a@mizzou.edu

ABSTRACT

In this paper, the theoretical framework of maximum *a posteriori* linear regression (MAPLR) based variance adaptation for continuous density HMMs is described. In our approach, a class of informative prior distribution for MAPLR based variance adaptation is identified, from which the close form solution of MAPLR based variance adaptation is obtained under its EM formulation. Effects of the proposed prior distribution in MAPLR based variance adaptation are characterized and compared with conventional maximum likelihood linear regression (MLLR) based variance adaptation. These findings provide a consistent Bayesian theoretical framework to incorporate prior knowledge in linear regression based variance adaptation. Experiments on large vocabulary speech recognition tasks were performed. The experimental results indicate that significant performance gain over the MLLR based variance adaptation can be obtained based on the proposed approach.

1. INTRODUCTION

Acoustic model adaptation is critical for speech recognition under mismatched or adverse ambient conditions. Unlike the regular model training, model adaptation is typically based on a small amount of training data and efficient adaptation methods often utilizes the structure of the model to share the adaptation information across similar model units. In particular, the linear transformation based model adaptation has become increasingly popular. In maximum likelihood linear regression (MLLR) based approach [1,2,3], a group of linear transformation matrices is estimated based on the principle of maximizing the likelihood on the adaptation data.

It is known that maximum likelihood estimation is a data driven parameter estimation method. When adaptation data is sparse, maximum likelihood estimation can lead to biased estimates. This is because the small amount of adaptation data may not be a good representation of the actual data distribution. Maximum *a posteriori* (MAP) based adaptation is a powerful approach. In MAP estimation, an appropriate prior distribution is used to incorporate prior knowledge into the model parameter estimation process, such that

$$\Lambda_{MAP} = \arg\max_{\Lambda} f(X \mid \Lambda) g(\Lambda) \qquad (1)$$

where the prior distribution $g(\Lambda)$ characterizes the distribution of the model parameter set Λ. The relation between ML and MAP estimation is through the Bayes' theorem where the posterior distribution $p(\Lambda \mid X) \propto f(X \mid \Lambda) g(\Lambda)$, and $f(X \mid \Lambda)$ is the likelihood function. In linear regression based model adaptation, the prior distribution $g(\Lambda)$ is a matrix variate distribution. It describes the distribution of the matrix Λ, which is assumed to be random. One of the advances in MAP based acoustic model adaptation is the maximum *a posteriori* linear regression (MAPLR) based mean vector adaptation [4,5], in which the linear transforms for mean vector adaptation are estimated based on the principle of MAP.

Variance adaptation is a challenging problem when the amount of training data is sparse. This is because variance is a second order statistics and often exhibits large variations . In [3], the MLLR based adaptation framework is extended to variance adaptation of continuous density HMMs. Although the use of linear regression in variance adaptation leads to an efficient use of the available training data, the problem of the biased estimation, related to data sparsity in maximum likelihood estimation, is even more acute than the situation in mean vector adaptation. This makes the need to incorporate prior knowledge in variance adaptation even more important. It should be pointed out that MAPLR variance adaptation is to estimate values of a special transform matrix, which does not even belong to the original HMM structure. Second, it is based on a structured parameter clustering, and the same transform matrix is used to adapt all variance parameters in the cluster.

In this paper, we develop the MAPLR based variance adaptation framework, and show that under a class of informative priors, MAPLR based variance adaptation has a close form solution under its EM formulation. From this close form solution, the effects of prior distribution in MAPLR based variance adaptation are characterized. We compare MAPLR based variance adaptation with MLLR based approach, and significant performance advantages are observed.

The novel contributions of this paper are:

- The theoretic framework of MAPLR based variance adaptation is developed and a class of informative prior distribution is identified.
- A close form solution of MAPLR for variance adaptation under the proposed prior is derived from its EM formulation. It is the real root of a special 4-th order equation, and the existence of such a real root solution is proved.
- Efficient prior distribution estimation methods are described based on the structured prior evolution under the empirical Bayes framework.
- The effects of prior distribution in MAPLR are characterized and compared with the conventional MLLR solution. Experimental results are given, which indicate that significant performance gain can be obtained comparing to MLLR based variance adaptation.

The organization of this paper is as follows. In section 2, the MAPLR based variance adaptation is formulated and the close form solution to this problem is presented. Section 3 is on prior parameter estimation. Experimental results and comparison to MLLR are given in section 4, and a summary is given in section 5.

2. MAPLR BASED VARIANCE ADAPTATION

In continuous density HMMs with mixture Gaussian densities, the Gaussian component is characterized by its mean and covariance matrix and denoted generically as $N(\mu_n, \Sigma_n)$. The covariance matrix Σ_n is a positive definite matrix that can be represented in the following form (Choleski factorization):

$$\Sigma_n = B_n^T B_n = \Sigma_n^{\frac{1}{2}} \Sigma_n^{\frac{1}{2}}. \quad (2)$$

In order to apply the linear regression in variance adaptation, a transformation matrix H_n is introduced and the adapted covariance matrix $\hat{\Sigma}_n$ is represented as follows:

$$\hat{\Sigma}_n = B_n^T H_n B_n, \quad (3)$$

where H_n is the transformation matrix to be estimated from the adaptation data. In the special case where no adaptation is made, H_n is just an identity matrix. Normally, in order to maintain the diagonal positive definiteness of Σ_n, H_n is assumed to be a diagonal matrix with positive elements along the diagonal. Due to the sparseness of the adaptation data, H_n is often tied to a cluster of Gaussian components, and the same transform H_n will be applied to all Gaussian components in the cluster during the process of linear regression based variance adaptation.

In MLLR based approach, for a Gaussian component cluster m, a linear transform matrix H_m is estimated based on the maximum likelihood principle [3], such that

$$H_m = \arg\max_H [\prod_k P(O_m^k \mid \lambda_m)], \quad (4)$$

where O_m^k (k=1,…) are data frames aligned to the cluster m and λ_m is the set of model parameters associated to the cluster m. In the MAPLR based variance adaptation, the transformation matrix is estimated according to

$$H_m = \arg\max_H [\prod_k P(O_m^k \mid \lambda_m)] p(H), \quad (5)$$

where $p(H)$ is the prior distribution of H. The use of prior distribution $p(H)$ allows incorporating prior knowledge into the variance adaptation process through a consistent Bayesian statistical framework. The estimated transformation matrix H_m is a combination of the information derived from the data and the prior knowledge about the distribution of H_m. The use of prior makes the parameter estimation problem more complex. This is because MAP solution is strongly dependent on the form of prior distributions being used. Therefore, finding a class of informative and yet solvable prior distributions for MAPLR becomes crucial.

In speech recognition, the covariance Σ of a Gaussian component is typically a diagonal matrix of the form $\Sigma = diag[\sigma_1^2,\ldots,\sigma_D^2]$, where $B = diag[\sigma_1,\ldots,\sigma_D]$. As in [3], in order to make the transformed variance Σ diagonal, we consider the linear regression matrix for MAPLR variance adaptation is of the form $H_m = diag[h_1^2,\ldots,h_D^2]$, whose prior distribution is denoted by $p(\mathbf{h})$, where $\mathbf{h}=[h_1,\ldots,h_D]^T$. Note H_m is an artificial hyper-structure in MAPLR based variance adaptation, and it does not belong to the original HMM structure. In our approach, H_m is parameterized as the square of another diagonal matrix $diag[h_1,\ldots,h_D]$, and the MAP estimation is performed on the transformed parameter space describing $diag[h_1,\ldots,h_D]$. Otherwise, the prior distribution is only on the positive values that H can take (e.g. chi-square, etc.), and a close form solution becomes difficult.

In order to solve Eq. (5), we consider the prior distribution $p(\mathbf{h})$ is from the family of vector normal distribution with mean vector $\mu_{\mathbf{h}} = [\mu_{\mathbf{h}}(1),\ldots,\mu_{\mathbf{h}}(D)]^T$ and the variance $\Sigma_{\mathbf{h}} = [\sigma_{\mathbf{h}}^2(1),\ldots,\sigma_{\mathbf{h}}^2(D)]^T$. Ignoring terms

which are irrelevant to maximization, the EM equation for (5) has the following form:

$$Q(M,\hat{M}) = -\frac{1}{2}\sum_{t,n,m}\gamma_t(n,m)[\log|\hat{\Sigma}_m| + (o_t-\mu_m)^T\hat{\Sigma}_m^{-1}(o_t-\mu_m)] + \log p(\mathbf{h}), \quad (6)$$

where $\gamma_t(n,m)$ is the probability of observing o_t in state n and mixture m. By expanding (6) according to the feature vector dimension, we have

$$Q(M,\hat{M}) = -\frac{1}{2}\sum_{t,n,m}\sum_{i=1}^{D}\gamma_t(n,m)[\log\sigma_m^{(i)2} + \log h_i^2 + \frac{h_i^{-2}}{\sigma_m^{(i)2}}(o_t^{(i)}-\mu_m^{(i)})^2] - \frac{1}{2}\log 2\pi\sigma_{\mathbf{h}}^2(i) - \frac{(h_i-\mu_{\mathbf{h}}(i))^2}{2\sigma_{\mathbf{h}}^2(i)}. \quad (7)$$

Set $\frac{\partial Q(M,\hat{M})}{\partial h_i} = 0$ for each i,

$$\sum_{t,n,m}\gamma_t(n,m)[-h_i^{-1} + h_i^{-3}\frac{(o_t^{(i)}-\mu_m^{(i)})^2}{\sigma_m^{(i)2}}] - \frac{h_i}{\sigma_{\mathbf{h}}^2(i)} + \frac{\mu_{\mathbf{h}}(i)}{\sigma_{\mathbf{h}}^2(i)} = 0. \quad (8)$$

Multiplying both sides of (8) by $-h_i^3$ and rearranging terms, it reduces to a 4-th order equation about h_i of the following form:

$$A\cdot h_i^4 - B\cdot h_i^3 + C\cdot h_i^2 - D = 0 \quad (9)$$

where $A = 1/\sigma_{\mathbf{h}}^2(i)$, $B = \mu_{\mathbf{h}}(i)/\sigma_{\mathbf{h}}^2(i)$, $C = \sum_{t,n,m}\gamma_t(n,m)$ and $D = \sum_{t,n,m}\gamma_t(n,m)(o_t^{(i)}-\mu_m^{(i)})^2/\sigma_m^{(i)2}$.

Since $A > 0$ and $D > 0$, the left part of (9) will tend to positive infinity when $h_i \to \infty$, and tend to $-D < 0$ when $h_i \to 0$. Therefore, (9) must have at least one positive real root, which is the close form solution of the MAPLR based variance adaptation under the EM formulation. It should be pointed out that although the general root formula does exist, the roots of a 4-th order equation can have two complex conjugate root pairs. The close form solution to the MAPLR based variance adaptation hinges on whether there is always a real root. By proving that there must be a real root to (9), the close form solution to the MAPLR based variance adaptation is therefore established.

It is noteworthy that, when a very loose prior distribution is applied so that $\sigma_{\mathbf{h}}^2(i)$ is very large, or data amount is very large so that $C >> A$, (9) will give a similar result as in maximum likelihood based approach; and on the other hand, when data amount is very small, or a very peaky prior distribution is selected, the root of (9) is tend to be around the prior mean $\mu_{\mathbf{h}}(i)$.

3. PRIOR HYPERPARAMETER ESTIMATION

In MAPLR based approach, additional parameters are needed to describe the prior distribution. These additional parameters are called hyperparameters. In a strict Bayes approach, hyperparameters in the prior density is assumed known based on a common or subjective knowledge about the stochastic process. But in most cases, these hyperparameters cannot be derived from the subjective knowledge and alternative approaches are needed. One popular solution to adopt is based on empirical Bayes (EB) approach in which the prior parameters are also estimated from the data. However in EB approach, additional data points are often required.

The prior estimation methods used in our approach for variance adaptation are based on combinations of the structural information of the model and the EB based estimation process. In MAPLR based variance adaptation, adaptation transforms are shared according to a tree structure, which is derived from the model units based on certain similarity measures. This tree structure is used in our approach so that the prior estimation can be evolved from the tree. Prior evolution is an efficient prior estimation method. In this approach, the mode of the prior distribution at each tree node is taken to be the MAPLR solution of its predecessor node [5].

We studied two methods of prior estimation for MAPLR based variance adaptation. The first method estimates a global transformation as the global prior mean and uses EB approach to estimate the variance vector of the prior distribution at each tree node using a lower sample count cut-offs. The second method is based on the prior evolution approach so that the mean vector of the prior distribution is evolved from the root of the tree to its leaves [5]. This is in addition to the EB based estimation of the variance vector of the prior distribution.

4. EXPERIMENTS

The speech recognition experiments were performed on the Wall Street Journal (WSJ) speaker adaptation task using the official 1993 Spoke 3 non-native speaker adaptation and evaluation data (ET_S3). The standard 5k-trigram language model specified for the evaluation was used. The speech feature vector is MFCC based with standard 39 dimensions (c, Δc, $\Delta\Delta c$, e, Δe, $\Delta\Delta e$). There are 10 speakers in the database, with 40 adaptation sentences and 40~43 testing sentences for each speaker. The speaker independent model was trained on the standard speaker independent WSJ SI-84 portion of the training corpus. Crossword triphones were used and the baseline speaker independent model was obtained by using decision tree based state tying. In adaptation, the silence model was not adapted.

To compare the adaptation results on variance adaptation, we performed two sets of experiments. One is the baseline using MLLR based mean adaptation and MLLR based variance adaptation. Table 1 tabulates the

performance of MLLR based adaptation using 20 and 40 adaptation sentences, respectively. As illustrated in the table, the MLLR based mean+variance adaptation only provides a slight performance improvement over the MLLR based mean adaptation. Its relative error rate reduction is around 1.9% ~ 2.0% with 20 and 40 adaptation sentences.

Table 2 illustrates the performance of MLLR based mean adaptation plus MAPLR based variance adaptation. The set of experiments in Table 2 differs from the set of experiments in Table 1 only at the variance adaptation where MAPLR variance adaptation was applied to replace the MLLR based approach. The relative word error rate reduction of MLLR based mean adaptation plus MAPLR based variance adaptation over MLLR based mean adaptation is between 3.0% ~ 4.4%. In all variance adaptation experiments, the sample count threshold of generating a transform matrix was set to 1000 and lower threshold of 200 was used to generate additional data points to estimate the variance of the prior distribution. However, because the estimated prior variance from the low sample count matrices tended to be too large and made the prior distribution too loose, a factor of $\frac{1}{15}$ was multiplied to shrink the estimated prior variance. Prior variance was estimated if there are at least 10 lower count matrices. Otherwise, the variance of the prior at the parent node is used.

TABLE I: WORD ERROR RATES OF MLLR BASED MODEL ADAPTATION METHODS (%)

Adaptation Method	20 adpt. utterances	40 adpt. utterances
MLLR mean only	16.88	14.74
MLLR mean + vari	16.55	14.45

TABLE II: WORD ERROR RATES OF MLLR BASED AND MAPLR BASED MODEL ADAPTION METHODS (%)

Adaptation Method	20 adpt. utterances	40 adpt. utterances
MLLR mean only	16.88	14.74
MLLR-mean + MAPLR-vari (1)	16.37	14.09

Table 3 tabulates the performance comparison of MLLR based mean adaptation plus MAPLR based variance adaptation using different prior distribution estimation methods described in the previous section. Compared with MLLR based variance adaptation, MAPLR based variance adaptation is more stable. We found experimentally that a global prior mean based prior hyperparameter estimation gave a better performance than the more data driven prior mean evolution method. This can be an indication that variance requires more data to estimate than the mean and even the prior evolution can be unreliable at the lower tree nodes.

TABLE III: WORD ERROR RATES OF MLLR/MAPLR BASED MODEL ADAPTION WITH DIFFERENT PRIOR DISTRIBUTION ESTIMATION METHODS (%)

Adaptation Method	20 adpt. utterances	40 adpt. utterances
MLLR-mean + MAPLR-vari (1)	16.37	14.09
MLLR-mean + MAPLR-vari (2)	16.43	14.31

Although the absolute performance gain in variance adaptation, as observed by other experiments as well [3], is relative small, the proposed MAPLR variance adaptation almost doubled the gain of variance adaptation introduced by the MLLR approach (1.9% ~ 2.0% vs. 3.0% ~ 4.4%) in our experimental study. The close form solution derived from the selected priors makes the MAPLR based variance adaptation suitable for large vocabulary speech recognition tasks.

5. SUMMARY

The theoretical framework of MAPLR based variance adaptation for continuous density HMMs was presented. We showed that there exists a class of informative prior distribution, from which the MAPLR based variance adaptation has close form solution under its EM formulation. These findings provided a consistent Bayesian theoretical framework to incorporate prior knowledge in linear regression based variance adaptation. Our experimental results indicated that the proposed MAPLR based variance adaptation approach is suitable for large vocabulary speech recognition tasks, and significant performance gain over the MLLR based variance adaptation could be obtained.

6. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Speaker Adaptation of HMMs Using Linear Regression," *CUED/F-INFENG/TR. 181*, 1994.

[2] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp.171 – 185. 1995.

[3] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework," *Computer Speech and Language*, Vol. 10, pp. 249-264. 1996.

[4] W. Chou, "Maximum A Posteriori Linear Regression with Elliptically Symmetric Matrix Variate Priors," *Proc. EuroSpeech*, Vol. 1, pp. 1-4, Budapest, Hungary. September 1999.

[5] W. Chou, O. Siohan, T. A. Myrvoll and C-H Lee, "Extended Maximum A Posteriori Linear Regression (EMAPLR) Model Adaptation for Speech Recognition," *Proc. ICSLP*, Beijing, China. October 2000.