社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－SLP－54 (13)
2004／12／20

Aggregate *a Posteriori* Linear Regression Adaptation of Hidden Markov Models

Jen-Tzung Chien and Chih-Hsien Huang

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan 70101, ROC
E-mail: jtchien@mail.ncku.edu.tw & acheron@chien.csie.ncku.edu.tw

Abstract We present a rapid and discriminative speaker adaptation algorithm for hidden Markov model (HMM) based speech recognition. The adaptation is based on the linear regression framework. Attractively, we estimate the regression matrices from speaker-specific adaptation data according to the aggregate *a posteriori* criterion, which is expressed in a form of classification error function. The aggregate *a posteriori* linear regression (AAPLR) is proposed to achieve discriminative adaptation so that the classification errors of adaptation data are minimized. The superiority of AAPLR to maximum *a posteriori* linear regression (MAPLR) is demonstrated. Different from minimum classification error linear regression (MCELR), AAPLR has closed-form solution to fulfill rapid adaptation. Experimental results reveal that AAPLR speaker adaptation does improve speech recognition performance with moderate computational cost compared to the maximum likelihood linear regression (MLLR), MAPLR and MCELR.

Key words: Hidden Markov model, MLLR, MAPLR, MCELR, Discriminative training, Aggregate *a posteriori* criterion, Speaker adaptation, Speech recognition

1. Introduction

In general, the speech hidden Markov model (HMM) parameters were estimated using two categories of approaches: the distribution estimation and the discriminative estimation. The popular algorithms for distribution estimation were based on maximum likelihood (ML) [14] and maximum *a posteriori* (MAP) criteria [7][9]. Also, the minimum classification error (MCE) [12] and maximum mutual information (MMI) [1] served as criteria for discriminative estimation. Using MCE discriminative estimation, the generalized probabilistic descent (GPD) algorithm was applied to iteratively estimate the HMM parameters. Implementation of MCE and MMI were time-consuming. Bahl et al. [1] addressed that the ML estimation could outperform MMI estimation when the data distributions were properly specified and the sufficiently large training samples were collected for distribution estimation. However, in real world, the problems of model assumption errors as well as sparse training data were inevitable. When estimating HMM distributions for speech recognition, the maximization of likelihood function or *a posteriori* probability was inferior to direct minimization of classification error.

In this study, we focus on developing a rapid and discriminative *speaker adaptation* algorithm where two categories of estimation approaches are considered. We would like to adapt the existing continuous-density HMM parameters to a new speaker and his/her operating environments. The speech recognition performance can be improved using speaker-adaptive HMM parameters. In the literature, the linear regression adaptation methods using maximum likelihood linear regression (MLLR) [14] and maximum *a posteriori* linear regression (MAPLR) [3][7] were popular and shown to be effective for batch adaptation. MLLR and MAPLR transformed clusters of HMM mean vectors using regression matrices, which were estimated via ML and MAP criteria, respectively. The regression classes should be assigned in advance. To facilitate the adaptation efficiency, we presented the quasi-Bayes linear regression (QBLR) [5] for sequential speaker adaptation. Using QBLR, the randomness of regression matrix was modeled by a matrix variate normal distribution so that the reproducible prior/posterior distributions were generated to build meaningful mechanism for sequential adaptation. Also, when considering the model assumption errors, the uncertainty in estimating linear regression matrices should be tackled to achieve robust classification. In [6], the linear regression based Bayesian predictive classification (LRBPC) was proposed for robust speech recognition. The robust classifier was constructed by averaging the randomness of regression parameters in Bayesian decision rule.

Here, we concern the issue of discriminative adaptation where the classification errors of adaptation data are minimized to attain the discriminative estimation. The most likely regression matrices are estimated by considering the likelihoods not only from target HMM's but also from competing HMM's. This approach is directly beneficial to reduce the classification error rates. Basically, the discriminative adaptation is different from MCE and MMI discriminative training [1][12][13] developed for model training. The MCE discriminative estimation has been applied for speaker adaptation [2][18]. In [2], the minimum classification error linear regression (MCELR) was explored. MCE criterion was merged to estimate the time-varying polynomial Gaussian mean functions in the trended HMM. Although the speech recognition performance was improved, the major weakness of MCELR came from the heavy costs on gradient calculation. He and Wu [11] proposed a "super" string model based MCELR adaptation where a special ratio of two positive functions was maximized to reduce the error rate and derive the closed-form solution.

To avoid extensive computation, we would like to derive a new closed-form solution to regression matrix for rapid adaptation. We properly incorporate the prior density of regression matrix and conduct the Bayesian distribution estimation. Interestingly, we

present the aggregate *a posteriori* linear regression (AAPLR) where the aggregate *a posteriori* (AAP) distribution [15] is maximized to find the optimal regression matrices. AAP probability is an aggregate of posterior probabilities, which can be expressed in a form of classification errors. Such estimation is referred as distribution estimation as well as discriminative estimation. Attractively, a closed-form solution to AAPLR is derived to achieve fast and discriminative speaker adaptation. In the experiments, we demonstrate the effectiveness and efficiency of applying AAPLR for speaker adaptation compared to other linear regression adaptation algorithms.

2. Discriminative Training and Linear Regression Adaptation

Before describing the new AAPLR adaptation algorithm, we survey the discriminative training using MCE and MMI criteria and the linear regression adaptation using MLLR, MAPLR and MCELR.

2.1 MCE and MMI discriminative training

Juang et al. [12] presented the MCE discriminative training algorithm with a three-step procedure. For the case of M-category classification, the first step is to determine the discriminant functions $\{g_m(X;\lambda_m), m=1,\cdots,M\}$, which are usually represented by probabilistic models. Second, a misclassification measure is introduced as follows

$$d_m(X) = -g_m(X;\lambda_m) + \log\left[\frac{1}{M-1}\sum_{j\neq m}\exp[\eta g_j(X;\lambda_j)]\right]^{1/\eta}. \quad (1)$$

where η is a positive number and λ_m is the model parameter. This measure is continuous and flexible with varying η. Notably, all competing classes $j\neq m$ are used during parameter estimation. At the third step, the loss function measuring the classification errors is formulated by

$$\ell(X;\lambda_m) = \ell(d_m(X)) = \frac{1}{1+\exp(-\gamma d_m(X)+\theta)}, \quad (2)$$

using sigmoid function $\ell(\cdot)$ with parameters γ and θ. In (1), the case $d_m(X)>0$ reflects the misclassification while $d_m(X)<0$ implies the correct classification. Loss function is a smoothed zero-one function used for recognition error rate minimization. Then, GPD algorithm based on MCE criterion is developed via minimizing the expected loss. The iterative learning rule of $\Lambda=\{\lambda_m\}$ is given by

$$\lambda_m^{(i+1)} = \lambda_m^{(i)} - \varepsilon U \nabla \ell(X;\lambda_m^{(i)}). \quad (3)$$

Here, i is the iteration index, X are the training samples, U is the positive definite matrix and ε is the learning rate.

On the other hand, Bahl et al. [1] presented the MMI discriminative training for HMM based speech recognition. The HMM parameters were estimated by maximizing the mutual information between the observation sequence X and the associated word sequence with parameter λ_m. The resulting MMI objective function is yielded as

$$\begin{aligned} I(W_m, X) &= \log(p(X|W_m)/p(X)) \\ &= \log p(X|W_m) - \log\sum_{j=1}^{M} p(W_j)p(X|W_j) \\ &= \log(\ell(-(d_m(X)+\log(M-1))) + \log M \end{aligned} \quad (4)$$

Interestingly, the mutual information was arranged as a function of logarithm of loss function in (2) [16]. MMI training was corresponding to MCE training. The model parameters $\Lambda=\{\lambda_m\}$ using MMI criterion are then estimated through the learning rule of (3) using the gradient of $I(W_m, X)$ with respect to λ_m.

When realizing MCE and MMI discriminative training, the standard forward-backward algorithm was replaced by N-best algorithm [8] or the beam search algorithm [13] so that the complexity of calculating likelihoods of competing models $j\neq m$ was alleviated. Differently, this paper concerns the discriminative linear regression adaptation rather than discriminative HMM training. In what follows, we investigate several linear regression adaptation methods and the conceptual evolution from discriminative training to the discriminative linear regression adaptation.

2.2 MLLR, MAPLR and MCELR adaptation

The linear regression speaker adaptation aims to estimate the cluster-dependent regression matrices, which are used to adapt the speaker-independent HMM parameters to a new speaker. By properly controlling the sharing of regression matrices, maximum likelihood linear regression (MLLR) can effectively find the maximum likelihood estimate of regression matrices for adaptation of HMM mean vectors. Assume that the HMM λ_m having a $d\times 1$ mean vector μ_m, the adapted mean vector $\hat{\mu}_m$ using $d\times(d+1)$ regression matrix $\mathbf{W}_{r(m)}$ is expressed by

$$\hat{\mu}_m = \mathbf{W}_{r(m)}\xi_m, \quad (5)$$

Here, $r(m)$ is the regression/cluster class for model m and ξ_m is the extended mean vector $[1,\mu_m^T]^T$. The maximum likelihood estimate of regression matrices $\mathbf{W}=\{\mathbf{W}_{r(m)}\}$ using adaptation data $X=\{\mathbf{x}_t\}=\{x_{t,i}\}$ is determined by

$$\mathbf{W}_{\mathrm{ML}} = \arg\max_{\mathbf{W}} p(X|\mathbf{W},\Lambda). \quad (6)$$

The expectation-maximization (EM) algorithm was applied to find optimal $\mathbf{W}_{\mathrm{ML}}$ [14]. Assuming that HMM covariance matrices are diagonal, i.e. $\Sigma_m = \mathrm{diag}(\sigma_{m1}^2,\cdots,\sigma_{md}^2)$, the ith row of $\mathbf{W}_{\mathrm{ML}}$ for regression class r was derived as [5]

$$\mathbf{w}_{ri}^{\mathrm{ML}} = \left(\sum_t\sum_{j,k\in\Omega_r}\frac{\zeta_t(j,k)}{\sigma_{jki}^2}x_{t,i}\xi_{jk}^T\right)\left(\sum_t\sum_{j,k\in\Omega_r}\frac{\zeta_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T\right)^{-1}, \quad (7)$$

where $\zeta_t(j,k)=P(j,k|X,\mathbf{W},\Lambda)$ is the posterior probability of $\mathbf{x}_t$ staying in state j and mixture component k given that current regression parameter $\mathbf{W}$ generates X. Notably, class label m is changed to the HMM state j and mixture component k in HMM framework. We assume that class labels j and k correspond to regression set Ω_r.

However, when the amount of adaptation data is sparse, the estimated regression matrices $\mathbf{W}_{\mathrm{ML}}$ are biased. It is helpful to achieve desirable adaptation performance by constraining the distribution shape of regression matrices using prior densities. In [3][5][7], the matrix-variate normal density served as the prior distribution for $\mathbf{W}$. The maximum *a posteriori* linear regression (MAPLR) was accordingly built by the MAP distribution estimation

$$\mathbf{W}_{\mathrm{MAP}} = \arg\max_{\mathbf{W}} p(\mathbf{W}|X,\Lambda) = \arg\max_{\mathbf{W}} p(X|\mathbf{W},\Lambda)g(\mathbf{W}). \quad (8)$$

Prior distribution of a regression matrix $\mathbf{W}_r$ is defined by

$$g(\mathbf{W}_r) \propto |\Delta_r|^{-1/2} \cdot q\left(\sum_{i=1}^{d}(\mathbf{w}_{ri}-\mathbf{m}_{ri})\Sigma_{ri}^{-1}(\mathbf{w}_{ri}-\mathbf{m}_{ri})^T\right), \quad (9)$$

where $\mathbf{m}_{ri}$ and Σ_{ri} are the mean vector and covariance matrix for regression row vector $\mathbf{w}_{ri}$, respectively, and the matrix Δ_r is defined by a $d(d+1)\times d(d+1)$ block diagonal matrix $\Delta_r = \mathrm{diag}(\Sigma_{r1},\ldots,\Sigma_{rd})$ with each diagonal block element Σ_{ri} being a $(d+1)\times(d+1)$ covariance matrix. Usually, q is an exponential function. The MAP estimate of the ith regression row vector was obtained by [3][5][7]

$$\mathbf{w}_{ri}^{\mathrm{MAP}} = \left(\sum_t \sum_{j,k\in\Omega_r} \frac{\zeta_t(j,k)}{\sigma_{jki}^2} x_{ti}\xi_{jk}^T + \mathbf{m}_{ri}\Sigma_{ri}^{-1}\right) \times \left(\sum_t \sum_{j,k\in\Omega_r} \frac{\zeta_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T + \Sigma_{ri}^{-1}\right)^{-1} \quad (10)$$

Also, Chengalvarayan proposed the first study on discriminative adaptation called minimum classification error linear regression (MCELR) adaptation algorithm to estimate the time-varying polynomial Gaussian mean functions in trended HMM [2]. The discriminative estimation of regression matrix was achieved through the gradient descent algorithm. Wu and Huo [18] further performed MCELR adaptation of MCE-trained HMM parameters using multiple regression classes. Under the same MCE criterion, the supervised speaker adaptation performance was better than that using MLLR adaptation for ML-trained HMM parameters. For both works, the learning rule of $\mathbf{W}$ was established by minimizing the expected loss function $\ell(X;\mathbf{W})$. The gradient of $\ell(X;\mathbf{W})$ with respect to $\mathbf{W}$ should be calculated for parameter learning. By adopting log likelihood function as the discriminant function,

$$g_{jk}(X;W_r,\lambda_{jk}) = \sum_t \sum_{j,k\in\Omega_r} \zeta_t(j,k)\log p(\mathbf{x}_t|W_r,\lambda_{jk}), \quad (11)$$

the parameter learning rule was formed by [18]

$$\mathbf{w}_{ri}^{(i+1)} = \mathbf{w}_{ri}^{(i)} - \varepsilon\ell(X;\mathbf{w}_{ri}^{(i)})(1-\ell(X;\mathbf{w}_{ri}^{(i)}))\cdot\left\{-\sum_t\sum_{j,k\in\Omega_r}\zeta_t(j,k) \times\left(\frac{x_{ti}-\mathbf{w}_{ri}^{(i)}\xi_{jk}}{\sigma_{jki}^2}\right)\xi_{jk}^T + \sum_t\sum_{j,k\in\Omega_r}\zeta_t(j,k)\left(\frac{x_{ti}-\mathbf{w}_{ri}^{(i)}\xi_{jk}}{\sigma_{jki}^2}\right)\xi_{jk}^T\right\} \quad (12)$$

3. Aggregate *a Posteriori* Linear Regression Adaptation

Although the MCELR algorithms [2][18] are able to accomplish discriminative speaker adaptation, the gradient descent implementation makes MCELR computationally expensive. Subsequently, we are introducing the generalized minimum error rate (GMER) [15] algorithm, which was proposed for discriminative model training by Li and Juang. An aggregate *a posteriori* (AAP) distribution was defined and arranged to express classification error measure. Under some meaningful assumptions, a closed-form solution to HMM training was obtained. Such solution was used efficiently perform model training. In this study, we adopt AAP probability as the objective function to estimate the regression matrices for discriminative speaker adaptation rather than estimate HMM parameters for discriminative training.

3.1 GMER algorithm and AAP criterion

In GMER training algorithm, the AAP probability is defined by aggregating the posterior probability $P(\lambda_m|X_{m,n})$ for all classes and their training samples

$$J_{\mathrm{AAP}}(\Lambda) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(X_{m,n}|\lambda_m)}{p(X_{m,n})} = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(X_{m,n}|\lambda_m)P_m}{\sum_{j=1}^{M}p(X_{m,n}|\lambda_j)P_j}, \quad (13)$$

where $X_{m,n}$ is the nth training sentence from the mth model λ_m with length T_n, $X_{m,n}=\{\mathbf{x}_{m,n,t}\}$ and P_m represents the prior probability of class m. Assume that training data are i.i.d., i.e.

$$p(X_{m,n}|\lambda_m) = \prod_{t=1}^{T_n} p(\mathbf{x}_{m,n,t}|\lambda_m), \quad (14)$$

the AAP probability can be arranged as

$$J_{\mathrm{AAP}}(\Lambda) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\ell(d_m^{\mathrm{AAP}}). \quad (15)$$

Here, $\ell(\cdot)$ is a loss function in (2) under the case of using $\gamma=1$ and $\theta=0$. The misclassification measure becomes

$$d_m^{\mathrm{AAP}} = \log p(X_{m,n}|\lambda_m)P_m - \log\sum_{j\neq m}p(X_{m,n}|\lambda_j)P_j. \quad (16)$$

In [15], a closed-form AAP solution was derived to estimate the HMM parameters

$$\Lambda_{\mathrm{AAP}} = \arg\max_{\Lambda} J_{\mathrm{AAP}}(\Lambda). \quad (17)$$

3.2 AAPLR criterion

However, in linear regression adaptation framework, we deal with the estimation of regression matrices $\mathbf{W}=\{\mathbf{W}_r\}$. To facilitate the parameter estimation for insufficient adaptation data, it is meaningful to introduce the prior density of regression matrix $g(W_r)$. To activate the capability of discriminative adaptation, we adopt the GMER algorithm where the evidence term $p(X_{m,n})$ expresses the likelihood of observations $X_{m,n}$ matching with all HMM's $\Lambda=\{\lambda_m\}$. Accordingly, optimizing AAP criterion is feasible to achieve discriminative capability. To combine the advantages of MAP estimation and AAP discriminative function for linear regression adaptation, we present a new aggregate *a posteriori* linear regression adaptation (AAPLR) algorithm for rapid and discriminative speaker adaptation. The AAPLR criterion is generated through merging regression parameter W_r and its prior density $g(W_r)$ as follows

$$J_{\mathrm{AAPLR}}(\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\frac{p(X_{m,n}|\mathbf{W}_r,\lambda_m)P_m g(\mathbf{W}_r)}{p(X_{m,n})}. \quad (18)$$

3.2 Relation between AAPLR and MAPLR

Before finding solutions to AAPLR, we illustrate the relation between the criteria of AAPLR and MAPLR from the perspective of expectation-maximization (EM) algorithm. When solving MAPLR in (8) via EM algorithm, we calculate an expectation function (E-step) as the objective criterion to be optimized. The expectation is operated over the logarithm of posterior density

$$P(\mathbf{W}|X,\Lambda) = \prod_t\left(\frac{p(\mathbf{x}_t|\mathbf{W},\Lambda)}{p(\mathbf{x}_t)}\right)g(\mathbf{W}). \quad (19)$$

The posterior expectation function $R(\hat{\mathbf{W}}|\mathbf{W})$ of new estimate $\hat{\mathbf{W}}$ given current estimate $\mathbf{W}$ is yielded as

$$R(\hat{\mathbf{W}}|\mathbf{W}) = E\{\log P(\hat{\mathbf{W}}|X,\mathbf{q},\Lambda)\Big|X,\mathbf{W}\} = \sum_m\sum_t \varsigma_t(m)\log\frac{p(\mathbf{x}_t, q_t = m\Big|\hat{\mathbf{W}}_{r(m)},\lambda_m)g(\hat{\mathbf{W}}_{r(m)})}{p(\mathbf{x}_t)}, \quad (20)$$

to resolve the missing data problem in HMM framework. In (20), the missing label sequence $\mathbf{q} = (q_1, q_2, \ldots, q_T)$ has element $q_t = m = (j,k)$ representing model m or state and mixture component (j,k). Also, $\varsigma_t(m) = P(q_t = m|X,\mathbf{W},\Lambda)$ denotes the posterior probability of observation $\mathbf{x}_t$ staying at model m under current estimate $\mathbf{W}$. Practically, we use Viterbi algorithm to align X using parameters $(\mathbf{W},\Lambda)$ and obtain the posterior probability as $\varsigma_t(m) = \delta(q_t - m)$ where $\delta(\cdot)$ is Kronecker delta function.

Next, the M-step of implementing MAPLR is to maximize the criterion $R(\hat{\mathbf{W}}|\mathbf{W})$ with respect to $\hat{\mathbf{W}}$. To illustrate the relation between MAPLR and AAPLR, we rewrite expectation function by expressing observation $\mathbf{x}_t$ as frame t of nth observation sentence $X_{m,n}$ aligned to model m as follows

$$J_{\text{MAPLR}}(\hat{\mathbf{W}}) = R(\hat{\mathbf{W}}|\mathbf{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N_m}\log\frac{p(X_{m,n},m\Big|\hat{\mathbf{W}}_r,\lambda_m)g(\hat{\mathbf{W}}_r)}{p(X_{m,n})} \quad (21)$$

The posterior probability $\varsigma_t(m)$ is also removed because the new equation is written using the aligned samples. When comparing AAPLR criterion in (18) and MAPLR criterion in (21), it is interesting to find that the *logarithm* is involved in MAPLR while the class prior probability P_m is only included in AAPLR. Also, MAPLR is an EM iterative procedure of optimizing expectation function of new estimate $\hat{\mathbf{W}}$ given the current estimate $\mathbf{W}$ while AAPLR performs *single optimization step* according to GMER algorithm. Actually, AAPLR can be modified to perform iterative EM steps. Another different point of these two criteria is the *treatment of the evidence* $p(\mathbf{X}_{m,n})$. Using MAPLR, the evidence is ignored for parameter estimation because this term is independent of regression matrix $\mathbf{W}_r$. However, in case of AAPLR, the evidence is used to indicate the likelihood from all classes including correct class and competing classes. This is critical to enable the discriminant power of AAPLR. In what follows, we show how AAPLR criterion is feasible to derive closed-form solution for rapid adaptation.

3.4 Derivation of AAPLR solution

Using AAPLR for speaker adaptation, we aim to adapt HMM mean vectors to a new speaker using linear regression matrices $\mathbf{W} = \{\mathbf{W}_{r(m)}\}$. The AAPLR criterion in (18) should be maximized to find regression matrices

$$\mathbf{W}_{\text{AAP}} = \arg\max_{\mathbf{W}} J_{\text{AAPLR}}(\mathbf{W}). \quad (22)$$

Similar to GMER algorithm, AAPLR criterion can be arranged as

$$J_{\text{AAPLR}}(\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\ell(d_m^{\text{AAPLR}}), \quad (23)$$

where

$$d_m^{\text{AAPLR}} = g_m(X;\lambda_m,\mathbf{W}_{r(m)}) - \log\left\{\frac{1}{M-1}\sum_{j\neq m}\exp[\eta g_j(X;\lambda_j,\mathbf{W}_{r(j)})]\right\}^{1/\eta}, \quad (24)$$

and

$$g_m(X;\lambda_m,\mathbf{W}_r) = \log\{p(X_{m,n}|\mathbf{W}_r,\lambda_m)g(\mathbf{W}_r)\}. \quad (25)$$

Importantly, the discriminant function is a *logarithm of posterior density of aligned observations* $X_{m,n}$.

Again, we adopt the prior density of regression matrix defined in (9) and assume that HMM covariance matrix is diagonal. The frame-based likelihood function turns out to be

$$p(\mathbf{x}_{m,n,t}|\mu_m,\Sigma_m,\mathbf{W}_r) = (2\pi)^{-d/2}\prod_{i=1}^{d}(\sigma_{mi}^2)^{-1/2} \times\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_{m,n,t,i} - \mathbf{w}_{ri}\xi_m)^2}{\sigma_{mi}^2}\right]. \quad (26)$$

AAPLR regression matrix is then estimated individually for each row vectors $\{\mathbf{w}_{ri}^{\text{AAP}}, i = 1,\cdots,d\}$. When maximizing AAPLR criterion, we take gradient of $J_{\text{AAPLR}}(\mathbf{W})$ with respect to $\mathbf{w}_{ri}$

$$\nabla_{\mathbf{w}_{ri}} J_{\text{AAPLR}}(\mathbf{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N_m}\ell(d_m^{\text{AAPLR}})(1-\ell(d_m^{\text{AAPLR}}))\nabla_{\mathbf{w}_{ri}} d_m^{\text{AAPLR}} \quad (27)$$

where the gradient of misclassification measure d_m^{AAPLR} consisted of two terms

$$\nabla_{\mathbf{w}_{ri}} d_m^{\text{AAPLR}} = \nabla_{\mathbf{w}_{ri}} g_m(X;\lambda_m,\mathbf{W}_r) - \nabla_{\mathbf{w}_{ri}} G_{\overline{m}}(X;\Lambda,\mathbf{W}_r). \quad (28)$$

The first term is due to the contribution of the log posterior distribution from the target model m

$$\nabla_{\mathbf{w}_{ri}} g_m(X;\lambda_m,\mathbf{W}_r) = \nabla_{\mathbf{w}_{ri}}\log\{p(X_{m,n}|\mathbf{W}_r,\lambda_m)g(\mathbf{W}_r)\} = \sum_t\left(\frac{x_{m,n,t,i} - \mathbf{w}_{ri}\xi_m}{\sigma_{mi}^2}\right)\xi_m^T + 2(\mathbf{w}_{ri} - \mathbf{m}_{ri})\Sigma_{ri}^{-1}, \quad (29)$$

and the second term is that from the competing models $\overline{m}$

$$\nabla_{\mathbf{w}_{ri}} G_{\overline{m}}(X;\Lambda,\mathbf{W}_r) = \nabla_{\mathbf{w}_{ri}}\log\left\{\frac{1}{M-1}\sum_{j\neq m}\exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})]\right\} = \frac{1}{\sum_{j\neq m}\exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})]}\cdot\left\{\sum_{j\neq m}\exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})] \times\left\{\sum_t\left(\frac{x_{m,n,t,i} - \mathbf{w}_{ri}\xi_j}{\sigma_{ji}^2}\right)\xi_j^T + 2(\mathbf{w}_{ri} - \mathbf{m}_{ri})\Sigma_{ri}^{-1}\right\}\right\} \quad (30)$$

By setting functions $\Psi_{\overline{m}}(X_{m,n}) = \frac{1}{\sum_{j\neq m}\exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})]}$, $\Phi_j(X_{m,n}) = \exp[g_j(X;\lambda_j,\mathbf{W}_{r(j)})]$ and $L(d_m) = \ell(d_m^{\text{AAPLR}})\times(1-\ell(d_m^{\text{AAPLR}}))$, the gradient in (27) is expressed by

$$\nabla_{\mathbf{w}_{ri}} J_{\text{AAPLR}}(\mathbf{W}) = \sum_{m=1}^{M}\sum_{n=1}^{N_m}L(d_m)\cdot\left\{\sum_t\left(\frac{x_{m,n,t,i} - \mathbf{w}_{ri}\xi_m}{\sigma_{mi}^2}\right)\xi_m^T + 2(\mathbf{w}_{ri} - \mathbf{m}_{ri})\Sigma_{ri}^{-1} - \Psi_{\overline{m}}(X_{m,n})\sum_{j\neq m}\Phi_j(\mathbf{X}_{m,n}) \times\left\{\sum_t\left(\frac{x_{m,n,t,i} - \mathbf{w}_{ri}\xi_j}{\sigma_{ji}^2}\right)\xi_j^T + 2(\mathbf{w}_{ri} - \mathbf{m}_{ri})\Sigma_{ri}^{-1}\right\}\right\} \quad (31)$$

By equating (31) to zero, we can obtain a closed-form solution $\mathbf{w}_{ri}^{\text{AAP}}$ to AAPLR adaptation through

$$\mathbf{w}_{ri}^{AAP}\cdot\left\{\begin{array}{l}\sum_{m=1}^{M}\sum_{n=1}^{N_m}L(d_m)\left[T_n(\frac{\xi_m\xi_m^T}{\sigma_{mi}^2})-\right.\\ T_n\Psi_{\overline{m}}(X_{m,n})\sum_{j\neq m}\Phi_j(X_{m,n})(\frac{\xi_j\xi_j^T}{\sigma_{ji}^2})\\ \left.-2(1-\Psi_{\overline{m}}(X_{m,n})\sum_{j\neq r}\Phi_j(X_{m,n}))\Sigma_{ri}^{-1}\right]\end{array}\right\}$$

$$=\left\{\begin{array}{l}\sum_{m=1}^{M}\sum_{n=1}^{N_m}L(d_m)\left[\sum_{t=1}^{T_n}(\frac{x_{m,n,t,i}}{\sigma_{mi}^2})\xi_m^T-\right.\\ \Psi_{\overline{m}}(X_{m,n})\sum_{j\neq m}\Phi_j(X_{m,n})\sum_{t=1}^{T_n}(\frac{x_{m,n,t,i}}{\sigma_{ji}^2})\xi_j^T\\ \left.-2(1-\Psi_{\overline{m}}(X_{m,n})\sum_{j\neq r}\Phi_j(X_{m,n}))\mathbf{m}_{ri}\Sigma_{ri}^{-1}\right]\end{array}\right\}. \quad (32)$$

Without loss of generality, we can express (32) in matrix equation $\mathbf{w}_{ri}^{AAP}\cdot\mathbf{L}=\mathbf{r}$ with left-hand-side $(d+1)\times(d+1)$ matrix $\mathbf{L}$ and right-hand-side $1\times(d+1)$ vector $\mathbf{r}$. In most cases, matrix $\mathbf{L}$ is nonsingular during AAPLR implementation. Finally, we solve a linear equation to find the regression vector

$$\mathbf{w}_{ri}^{AAP}=\mathbf{r}\cdot\mathbf{L}^{-1}. \quad (33)$$

With the closed-form solution to regression matrices $\mathbf{W}_{AAP}=\{\mathbf{w}_{ri}^{AAP}\}$, we avoid heavy computation of applying gradient descent algorithm. The rapid speaker adaptation can be achieved.

4. Experiments

4.1 Speech database and experimental setup

In the experiments, a Mandarin broadcast news transcription task was performed to examine the performance of speaker adaptation. We carried out four linear regression adaptation algorithms, MLLR, MAPLR, MCELR and AAPLR, to evaluate the adaptation performance in terms of syllable error rates and adaptation times under different adaptation data lengths. We prepared two speech corpora for HMM training and adaptation. The speaker-independent (SI) seed HMM's were trained using the benchmark Mandarin speech corpus TCC300 [4] which was recorded in office environments using close-talking microphones. We sampled 14266 sentences (about 16 hours) recorded by 100 males and 100 females for training. On the other hand, the adaptation and test data were sampled from the MATBN database [17]. MATBN database contained Mandarin Chinese broadcast news utterances, which were shared by the Public Television Service Foundation of Taiwan and collected by the Institute of Information Science at Academia Sinica, Taiwan. The total length of broadcast news utterances in MATBN database was about 220 hours. In this study, we sampled the preceding 40-hour speech data. This news data set was collected during the period from late 2001 to 2002. There were totally 779 stories, 104 headlines, 40 weather reports, and 40 ending sections. Only the stories were considered in the experiments. The length of 779 stories was about 30 hours and that of anchor speech was around 300 minutes. The anchor speech of the stories was segmented by hand. We performed two-pass adaptation prior to speech recognition; task adaptation and speaker adaptation. In task adaptation, we used 200 utterances (about 30 minutes) randomly sampled from MATBN database to adapt the SI seed HMM's to fit the broadcast news transcription task. MAP adaptation [9] algorithm was adopted for task adaptation. In speaker adaptation, there were two reporters (one male and one female). We collected 60 utterances (about 14 minutes) for each reporter and performed the linear regression adaptation. The other 40 utterances (about 9 minutes) from the same speaker were used for speech recognition. We averaged the adaptation results of two speakers

We built subsyllable HMM's for large-vocabulary continuous Mandarin speech recognition. Mandarin is a tonal and syllabic language. Without considering the tonal information, there are 408 Mandarin syllables. Each Mandarin syllable is composed of an initial (consonant) part and a final (vowel) part. We adopted the context-dependent subsyllable modeling to construct the HMM's for Mandarin speech. Totally, there were 94 context-dependent (CD) initials, 40 context-independent (CI) finals and 6 null initials to serve as HMM's. The HMM's of CD initials, CI finals and null initials contained three, five and three states, respectively. Each HMM state had at most 32 mixture components. In the experiments, we used 26-dimension feature vectors consisted of twelve Mel-frequency cepstral coefficients, one log energy coefficient and their first derivatives. We reported base syllable error rates (%) for comparative study. To evaluate the computational costs, we measured the processing time per regression class (second) for different algorithms on a personal computer with CPU Pentium IV 2.0 GHz and RAM 256 MB. The number of regression classes was fixed to be four; two for initials and two for finals. We performed five-fold cross-validation over the adaptation data set for all experiments.

4.2 Comparison of different linear regression adaptation

Several sets of experiments on supervised adaptation were reported. To evaluate the effect on adaptation data length, we performed speaker adaptation using five, ten, fifteen, twenty, forty and sixty adaptation utterances. Roughly, the length of each utterance was ranged from ten to twenty seconds. Table 1 lists the syllable error rates (%) for MLLR, MAPLR, MCELR and AAPLR.

		MLLR	MAPLR	MCELR	AAPLR
Number of Adaptation Data	5	34.6	32.9	32.1	31.8
	10	32.5	31.5	31.1	30.6
	15	31.1	30.6	30.1	29.6
	20	30.5	29.7	29.6	29.2
	40	29.6	29.1	28.8	28.5
	60	29.0	28.4	28.1	27.9

Table 1 Syllable error rates (%) of supervised adaptation using different adaptation algorithms

		MLLR	MAPLR	MCELR	AAPLR
Number of Adaptation Data	5	9.4	9.7	11.7	10.3
	10	10.7	10.9	13.1	11.7
	15	12.4	11.9	14.8	13.1
	20	13.2	13.4	16.4	14.3
	40	15.3	15.5	19.1	16.5
	60	21.7	21.3	26.5	23.3

Table 2 Averaged adaptation time (sec) for each regression class using different adaptation algorithms.

Without performing adaptation, the baseline syllable error rate (SER) is 53.3%. After performing task adaptation, SER is

greatly reduced to 41.6%. This implies that there is significant environmental mismatch between TCC300 with ordinary read speech and MATBN with broadcast news speech. Namely, it is important to perform environmental adaptation for a new task of broadcast news transcription. Further, when performing linear regression speaker adaptation, we find that all linear regression adaptation algorithms using different number of adaptation data do reduce SER's . For the case of five adaptation utterances, AAPLR obtains SER of 31.8%, which is better than those of MLLR (34.6%), MAPLR (32.9%) and MCELR (32.1%). As the number of adaptation data increases to ten, twenty and sixty, all recognition results are improved accordingly. After using sixty adaptation utterances, the performance of AAPLR (27.9%) still outperforms those of MLLR (29%), MAPLR (28.4%) and MCELR (28.1%). In general, MAPLR is better than MLLR due to the incorporation of prior regression information. MCELR performs better than MLLR and MAPLR due to merging the discriminant capability. However, the superiority of AAPLR to MLLR, MAPLR and MCELR is caused by merging the contributions of prior information and discriminant power. Besides, we list the computational costs of MLLR, MAPLR, MCELR and AAPLR adaptation in Table 2. We find that roughly MCELR spends additional 23% computational cost compared to MLLR and MAPLR. Also, the computational cost of AAPLR is about 8% higher than those of MLLR and MAPLR. But, AAPLR is computationally efficient than MCELR. In summary, the proposed AAPLR is good for rapid and discriminative speaker adaptation.

5. Conclusion

We have presented a new AAPLR algorithm for rapid and discriminative speaker adaptation. The AAP criterion was introduced to achieve model discriminability and simultaneously derive a closed-form solution for rapid parameter estimation. The adapted speech HMM's using discriminative regression matrices were able to enhance the speech recognition performance for broadcast news transcription. More importantly, we established AAPLR algorithm in which a closed-form solution to regression matrices was derived to obtain desirable adaptation performance. Also, the prior information of transformation matrix was incorporated in AAPLR criterion for constrained Bayesian estimation. The robustness of estimating regression matrices was guaranteed according to AAPLR Bayesian approach. The continuous-density HMM parameters of all acoustic units were effectively adapted. From the experiments, we found that using AAPLR for supervised speaker adaptation achieved the best SER among four adaptation algorithms. This fact was validated for different numbers of adaptation data. Also, the computation cost of AAPLR was smaller than MCELR and moderately increased compared to MLLR and MAPLR.

6. References

[1] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", in *Proc. ICASSP*, vol. 11, pp. 49-52, 1986.

[2] R. Chengalvarayan, "Speaker adaptation using discriminative linear regression on time-varying mean parameters in trended HMM", *IEEE Signal Processing Letters*, vol. 5, pp. 63-65, 1998.

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

[4] J.-T. Chien and C.-H. Huang, "Bayesian learning of speech duration models", *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 558-567, 2003.

[5] J.-T. Chien, "Quasi-Bayes linear regression for sequential learning of hidden Markov models", *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 268-278, 2002.

[6] J.-T. Chien, "Linear regression based Bayesian predictive classification for speech recognition", *IEEE Trans. Speech Audio Processing*, vol. 11, no. 1, pp. 70-79, 2003.

[7] W. Chou, "Maximum a posteriori linear regression with elliptically symmetric matrix variate priors", in *Proc. EUROSPEECH*, vol. 1, pp. 1-4, 1999.

[8] Y.-L. Chow, "Maximum mutual information estimation of HMM parameters for continuous speech recognition using the N-best algorithm", in *Proc. ICASSP*, pp. 701-704, 1990.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chains", *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 291-298, 1994.

[10] P. S. Gopalakrishnan, D. Kanevsk, A. Nádas and D. Nahamoo, "An inequality for rational function with applications to some statistical estimation problems", *IEEE Trans. Information Theory*, vol. 37, pp. 107-113, 1991.

[11] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs", in *Proc. ICASSP*, vol. 1, pp. 556-559, 2003.

[12] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

[13] S. Kapadia, V. Valtchev and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *Proc. ICASSP*, vol. 2, pp. 491-494, 1993.

[14] C. J. Leggeter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, pp. 171-185, 1995.

[15] Q. Li and B.-H. Juang, "Fast discriminative training for sequential observations with application to speaker identification", in *Proc. ICASSP*, vol. 2, pp. 397-400, 2003.

[16] W. Reichl and G. Ruske, "Discriminative training for continuous speech recognition", in *Proc. EUROSPEECH*, vol. 1, pp. 537-540, 1995.

[17] H. M. Wang, "MATBN 2002: A Mandarin Chinese broadcast news corpus", in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003.

[18] J. Wu and Q. Huo, "Supervised adaptation of MCE-trained CDHMMs using minimum classification error linear regression", in *Proc. ICASSP*, vol. 1, pp. 605-608, 2002.