PAPER

Smoothing Method for Improved Minimum Phone Error Linear Regression

Yaohui QI[†,††,†††a)], **Fuping PAN**[††], **Fengpei GE**[††], ***Nonmembers***, **Qingwei ZHAO**[††], ***Member***, ***and*** **Yonghong YAN**[†,††], ***Nonmember***

SUMMARY A smoothing method for minimum phone error linear regression (MPELR) is proposed in this paper. We show that the objective function for minimum phone error (MPE) can be combined with a prior mean distribution. When the prior mean distribution is based on maximum likelihood (ML) estimates, the proposed method is the same as the previous smoothing technique for MPELR. Instead of ML estimates, maximum a posteriori (MAP) parameter estimate is used to define the mode of prior mean distribution to improve the performance of MPELR. Experiments on a large vocabulary speech recognition task show that the proposed method can obtain 8.4% relative reduction in word error rate when the amount of data is limited, while retaining the same asymptotic performance as conventional MPELR. When compared with discriminative maximum a posteriori linear regression (DMAPLR), the proposed method shows improvement except for the case of limited adaptation data for supervised adaptation.

key words: *speaker adaptation (SA), maximum likelihood linear regression (MLLR), maximum a posteriori linear regression (MAPLR), minimum phone error linear regression (MPELR), discriminative maximum a posteriori linear regression (DMAPLR)*

1. Introduction

Speaker adaptation is a effective way to improve the performance of speech recognition and has become an important component of automatic speech recognition systems. It reduces the mismatch between the training and testing data caused by the speaker variability. Model-based adaptation methods, which adjust the parameters of the original hidden Markov model (HMM) set to fit the actual acoustic characteristics by using some adaptation data from the target user, have been popular for many years.

Transformation-based maximum likelihood linear regression (MLLR) [1], [2] is one of model-based adaptation methods. MLLR is effective when the amount of adaptation data is limited and has a wide range of applications [3]. It uses affine transformation to map the original acoustic model to the speaker adaptation (SA) acoustic model. For MLLR, a transformation matrix can be used for large number of model parameters, which allowed those model parameters that have not been observed in the adaptation data to be adapted. MLLR uses maximum likelihood (ML) criterion to estimate the parameters of linear transforms and a sufficient amount of data is required before it begins to be effective. To address this problem, maximum a posteriori (MAP) [4] based linear regression, which estimates the transformation matrixes using the MAP criterion, has been proposed. Maximum a posteriori linear regression (MAPLR) [5]–[9] and structural MAPLR (SMAPLR) [10] are notable examples. These approaches incorporate the prior knowledge to address the potential over-fitting problem and have been shown to be successful. Another extension to MLLR is the use of discriminative criterion for transform parameter estimation. The previous methods include H-criterion based discriminative linear transform (H-cri DLT) [11], minimum phone error linear regression (MPELR) [12], [13], minimum classification error linear regression (MCELR) [14], [15], minimum word classification error linear regression (MWCELR) [16] and soft margin estimation linear regression (SMELR) [17]. These methods were proposed to increase the separation between parameters. Recently, discriminative maximum a posteriori linear regression (DMAPLR) [18] is proposed to increase the discriminative capability of MAP based linear regression estimation and has shown better performance.

In this paper we revisit the smoothing technique for MPELR. We show that the objective function for MPE can be combined with a prior mean distribution. The performance of MPELR can be improved by using MAP estimates of the Gaussian parameters as the center of prior. The use of MAP statistics to smooth the discriminative statistics is motivated by the idea of behind minimum phone error maximum a posterior (MPE-MAP) [19]. Considering there may not be enough data to estimate the ML Gaussian parameters in the context of adaptation, MPE-MAP uses MAP statistics to estimate the center of prior used to smooth the MPE-trained parameters. A large vocabulary continuous speech recognition task is used to assess the effectiveness of the proposed method for supervised and unsupervised adaptation and to compare its performance with that of MLLR, MAPLR, and DMAPLR.

The remainder of this paper is organized as below. In Sect. 2, we review the theory of model-based linear regression method for speaker adaptation. In Sect. 3, the method of using MAP statistics to smooth the MPE statistics in MPELR is given. Experiments are described in Sect. 4. The

Manuscript received November 15, 2013.
Manuscript revised March 4, 2014.

[†]The authors are with College of Information and Electronics, Beijing Institute of Technology, Beijing, 100081 China.
[††]The authors are with Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190 China.
[†††]The author is with College of Physics Science and Information Engineering, Hebei Normal University, Shijiazhuang, 050024 China.
a) E-mail: qiyaohui@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E97.D.2105

results are given on a large vocabulary recognition task for supervised and unsupervised adaptation. Finally, a summary and conclusion are presented in Sect. 5.

2. Model-Based Linear Regression Adaptation Methods

Given a set of acoustic models, λ, transformation-based model space adaptation approaches apply a transformation function F_φ to map λ to a new set of acoustic models, $\hat{\lambda}$. The parameters, φ, of the transformation function is derived from the adaptation data.

Affine transformation is used in linear regression adaptation methods. Mean transform is investigated initially since means are assumed to characterize the main differences between speakers [1]. Then, the variance parameters are also updated [20]. However, compared to mean transforms, the improvement of recognition performance is limited [21]. In this paper, we focus on mean adaptation. We obtain the adapted mean $\hat{\mu}$ according to

$$\hat{\mu} = A\mu + b = W\xi$$

where W is an $D \times (D+1)$ matrix $[b \;\; A]$; $\xi = [1 \;\; \mu^T]^T$ is the extended mean vector (D is the dimension of the features); μ is the original mean. In linear regression model-space adaptation, W is the transformation parameter.

Many optimization criteria have been used for estimation of the transformation parameters. In this paper, MLLR, MAPLR, MPELR and DMAPLR are examined.

2.1 MLLR

The ML criterion is initially used to estimate the transformation matrix because of its simplicity. Given R adaptation observation sequences $\{O_1, O_2, \ldots, O_r, \ldots O_R\}$, we estimate transformation matrix W based on the ML criterion [1]:

$$\hat{W}_{\mathrm{ML}} = \arg\max_W \{\sum_{r=1}^{R} \log P(O_r \mid s_r^{\mathrm{ref}}, W, \lambda)\} \tag{1}$$

where s_r^{ref} is the corresponding reference of adaptation observation O_r.

MLLR estimates the transformation matrix W to maximize the likelihood of the adaptation data given the adapted model. The values of W are found by optimizing the following auxiliary function which ignores the other parameters independent of W [1]:

$$Q_{\mathrm{ML}}(W, \hat{W}) = \sum_{r=1}^{R} \sum_{m=1}^{M} \sum_{t=1}^{T_r} \gamma_m(t) \log N(O_r(t), \hat{W}\xi_m, \Sigma_m) \tag{2}$$

Where ξ_m and Σ_m are the extended mean vector and covariance matrix for Gaussian component m; $O_r(t)$ is the observation vector at time t; $\gamma_m(t)$ is the posterior probability of Gaussian component m at time t. For the diagonal covariance case a closed form of $\hat{W}$ can be obtained. The inverse of the i-th row of $\hat{W}$ is given by [1]

$$\hat{W}^{(i)T} = G^{(i)-1} K^{(i)} \tag{3}$$

$$G^{(i)} = \sum_{m=1}^{M} \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m(t)}{\sigma_m^{(i)2}} \xi_m \xi_m^T \tag{4}$$

$$K^{(i)} = \sum_{m=1}^{M} \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m(t) O_r^{(i)}(t)}{\sigma_m^{(i)2}} \xi_m \tag{5}$$

where $\sigma_m^{(i)2}$ is the i-th element of the diagonal variance and $O_r^{(i)}(t)$ is the i-th element of the feature vector at time t.

2.2 MAPLR

ML estimation assumes that W is fixed but unknown parameters. W depends only on the original acoustic model and the adaptation data. When very small amount of adaptation data is available the MLLR adapted acoustic model performs even worse than the SI acoustic model. So MAP criterion is used for the estimation of the transformation parameters to take into consideration the prior density. The prior of the transformation matrix itself [5] and the mean parameters [6]–[9] can both use for the MAPLR estimation. In this paper, we use mean prior.

MAPLR estimates the transformation matrix W by [7]:

$$\hat{W}_{\mathrm{MAP}} = \arg\max_W \{\sum_{r=1}^{R} \log[P(O_r \mid s_r^{\mathrm{ref}}, W, \lambda)\, p(W, \lambda)]\} \tag{6}$$

where $p(W, \lambda)$ is the joint prior distribution of W and λ. The multivariate Normal distribution is generally used as the prior distribution for the m-th Gaussian [7]

$$p(W, \xi_m) = \frac{\exp[-\frac{1}{2}(W\xi_m - \eta_m)^T (\beta V_m)^{-1} (W\xi_m - \eta_m)]}{(2\pi)^{D/2} |\beta V_m|^{1/2}} \tag{7}$$

where η_m and V_m are hyper parameters; β is a scaling factor that controls the contribution of the prior distribution. The ML estimation is a special case of MAP estimation when β approaches infinity. The auxiliary function is written by ignoring the parameters independent of W [6]:

$$Q_{\mathrm{MAP}}(W, \hat{W}) = Q_{\mathrm{ML}}(W, \hat{W}) + \sum_{m=1}^{M} \frac{1}{\beta} [-\frac{1}{2} (\hat{W}\xi_m - \eta_m)^T V_m^{-1} (\hat{W}\xi_m - \eta_m)] \tag{8}$$

where $Q_{\mathrm{ML}}(W, \hat{W})$ is the auxiliary function for MLLR estimation. With the prior as in Eq. (7), the MAP estimation of transformation matrix $\hat{W}$ can be obtained with a similar derivation to that in the MLLR approach. For the diagonal covariance case, the calculation for $\hat{W}$ is as Eq. (3) but with

[7]:

$$G^{(i)} = \sum_{m=1}^{M} \left(\frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m(t)}{\sigma_m^{(i)2}} + \frac{\frac{1}{\beta}}{v_m^{(i)2}} \right) \xi_m \xi_m^T \tag{9}$$

$$K^{(i)} = \sum_{m=1}^{M} \left(\frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_m(t) O_r^{(i)}(t)}{\sigma_m^{(i)2}} + \frac{\frac{1}{\beta}\eta_m^{(i)}}{v_m^{(i)2}} \right) \xi_m \tag{10}$$

where $v_m^{(i)2}$ and $\eta_m^{(i)}$ are the i-th element of V_m and η_m respectively. For small amounts of adaptation data the prior statistics are important. As more data becomes available, the adaptive statistics will become dominant.

2.3 DMAPLR

MAP estimation has the learning advantages of being effective and stable. In order to increase the discriminative power on the MAP-based linear regression adaptation method a new objective function is proposed. This function combines the MAP objective function and likelihood ratio (LR) score [18]:

$$\begin{aligned} &\hat{W}_{\mathrm{DMAP}} = \\ &\arg\max_{W} \{ \alpha_1 \sum_{r=1}^{R} \log[P(O_r \mid s_r^{\mathrm{ref}}, W, \lambda)\, p(W, \lambda)] \\ &+ \alpha_2 \sum_{r=1}^{R} \sum_{n=1}^{N} \varepsilon_n \log \frac{p(O_r \mid s_r^{\mathrm{ref}}, W, \lambda)}{p(O_r \mid s_r^n, W, \lambda)} \} \end{aligned} \tag{11}$$

where α_1 and α_2 are weighting parameters; ε_n is a scaling factor which is used to control the weight of each of the N-best hypothesis and satisfies $\sum_{n=1}^{N} \varepsilon_n = 1$; s_r^n is the n-th competing word sequence corresponding to observation O_r.

With the multivariate Normal distribution as in Eq. (7), DMAPLR calculates the transformation matrix for the case of diagonal covariance also using Eq. (3) but with [18]:

$$\begin{aligned} G^{(i)} = &\sum_{r=1}^{R} \sum_{m \in s_r^{\mathrm{ref}}} \left(\frac{(\alpha_1 + \alpha_2) \sum_{t=1}^{T_r} \gamma_m(t)}{\sigma_m^{(i)2}} + \frac{\frac{\alpha_1}{\beta}}{v_m^{(i)2}} \right) \xi_m \xi_m^T \\ &- \sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{l \in s_r^n} \varepsilon_n \frac{\alpha_2 \sum_{t=1}^{T_r} \gamma_l(t)}{\sigma_l^{(i)2}} \xi_l \xi_l^T \end{aligned} \tag{12}$$

$$\begin{aligned} K^{(i)} = &\sum_{r=1}^{R} \sum_{m \in s_r^{\mathrm{ref}}} \left(\frac{(\alpha_1 + \alpha_2) \sum_{t=1}^{T_r} \gamma_m(t) O_r^{(i)}(t)}{\sigma_m^{(i)2}} + \frac{\frac{\alpha_1}{\beta}\eta_m^{(i)}}{v_m^{(i)2}} \right) \xi_m \\ &- \sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{l \in s_r^n} \varepsilon_n \frac{\alpha_2 \sum_{t=1}^{T_r} \gamma_l(t) O_r^{(i)}(t)}{\sigma_l^{(i)2}} \xi_l \end{aligned} \tag{13}$$

where $l \in s_r^n$ and refers to the Gaussian belonging to a competing model in the n-th best hypothesis.

2.4 MPELR

Due to the successful application in acoustic model training, the minimum phone error (MPE) criterion is also used for the estimation of transformation parameters to increase the discriminative ability of model.

MPELR estimates the transformation matrix W by [13]:

$$\hat{W}_{\mathrm{MPE}} = \arg\max_{W} \{ \sum_{r=1}^{R} \sum_{s_i \in s_r^{\mathrm{lat}}} P(s_i \mid O_r, W, \lambda) A(s_i, s_r^{\mathrm{ref}}) \} \tag{14}$$

where s_r^{lat} is the corresponding word lattice of O_r and is used as an approximation to the hypothesis space; s_i is one of hypothesized word sequences in s_r^{lat}; $P(s_i \mid O_r, W, \lambda)$ is the posteriori probability of hypothesis s_i given O_r; $A(s_i, s_r^{\mathrm{ref}})$ is the phone accuracy of hypothesis s_i compared with the corresponding reference phones and could usually be approximated to the sum of the phone accuracy over all phones in s_i.

The weak-sense auxiliary function is proposed [19] for the optimization of discriminative criteria. For the case of MPE, the auxiliary function is based on the log likelihood of phone arc q, $\log p(q)$. The auxiliary function for MPE-based mean transform estimation is written by ignoring the parameters independent of the transform W [12]:

$$\begin{aligned} &Q_{\mathrm{MPE}}(W, \hat{W}) \\ &= \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{latt}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \gamma_q^{\mathrm{MPE}} \log N(O_r(t), \hat{W}\xi_m, \Sigma_m) \\ &+ \sum_{m=1}^{M} D_m [-\frac{1}{2} (W\xi_m - \hat{W}\xi_m)^T \hat{\Sigma}_m^{-1} (W\xi_m - \hat{W}\xi_m)] \\ &= \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{latt}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, \gamma_q^{\mathrm{MPE}}) \log N(O_r(t), \hat{W}\xi_m, \Sigma_m) \\ &- \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{latt}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, -\gamma_q^{\mathrm{MPE}}) \log N(O_r(t), \hat{W}\xi_m, \Sigma_m) \\ &+ \sum_{m=1}^{M} D_m [-\frac{1}{2} (W\xi_m - \hat{W}\xi_m)^T \hat{\Sigma}_m^{-1} (W\xi_m - \hat{W}\xi_m)] \end{aligned} \tag{15}$$

where $q \in s_r^{\mathrm{lat}}$ denotes a phone q arc that belongs to the word lattice s_r^{lat}; s_q and e_q are the start and end times of phone arc q respectively; $\gamma_{qm}(t)$ is the posterior probability at state j, mixture component m on the condition of arc q at frame t; $\gamma_q^{\mathrm{MPE}} = \frac{1}{k} \frac{\partial F_{\mathrm{MPE}}(\lambda)}{\partial \log p(q)}$ is a quantity defined for MPE training and k is the acoustic scale; D_m is the smoothing factor with a constant E, $D_m = E \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{lat}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, -\gamma_q^{\mathrm{MPE}})$.

Eq. (15) has two parts, which are analogous to the numerator and denominator terms in the maximum mutual information (MMI) auxiliary function. But the definition of the numerator and the denominator are different from those in the MMI. The numerator statistics is accumulated with the arcs whose γ_q^{MPE} is positive, while the denominator statistics is accumulated with arcs whose γ_q^{MPE} is negative.

By calculating the partial differential of the Eq. (15) with respect to each row of the transformation matrix $\hat{W}$ a closed form can be obtained from Eq. (3) for the case of diagonal covariance but with [12]:

$$G^{(i)} = \sum_{m=1}^{M} \left(\frac{\gamma_m^{\mathrm{num}} - \gamma_m^{\mathrm{den}} + D_m}{\sigma_m^{(i)2}}\right)\xi_m \xi_m^T \tag{16}$$

$$K^{(i)} = \sum_{m=1}^{M} \left(\frac{\theta_m^{\mathrm{num}}(O^{(i)}) - \theta_m^{\mathrm{den}}(O^{(i)}) + D_m \tilde{\mu}_m^{(i)}}{\sigma_m^{(i)2}}\right)\xi_m \tag{17}$$

where $\tilde{\mu}_m$ is the adapted mean vector with the initial MLLR transform matrix W; γ_m^{num}, $\theta_m^{num}(O^{(i)})$ and γ_m^{den}, $\theta_m^{den}(O^{(i)})$ are the numerator and denominator statistics with the following forms [22],

$$\gamma_m^{\mathrm{num}} = \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{lat}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, \gamma_q^{\mathrm{MPE}})$$

$$\theta_m^{\mathrm{num}}(O^{(i)}) = \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{lat}}} \sum_{t=s_q}^{e_q} r_{qm}(t) \max(0, \gamma_q^{\mathrm{MPE}}) O_r^{(i)}(t) \tag{18}$$

$$\gamma_m^{\mathrm{den}} = \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{lat}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, -\gamma_q^{\mathrm{MPE}})$$

$$\theta_m^{\mathrm{den}}(O^{(i)}) = \sum_{r=1}^{R} \sum_{q \in s_r^{\mathrm{lat}}} \sum_{t=s_q}^{e_q} \gamma_{qm}(t) \max(0, -\gamma_q^{\mathrm{MPE}}) O_r^{(i)}(t) \tag{19}$$

Here, the calculation of γ_q^{MPE} is as follows:

$$\gamma_q^{\mathrm{MPE}} = \gamma_q(c_q - c_{\mathrm{avg}})$$

where γ_q is the posterior probability for phone q in the lattice; c_{avg} is the average phone accuracy over all word sequences in the lattice; c_q is the expected phone accuracy over all word sequences containing a phone arc q. The calculation of c_q and c_{avg} is based on the phone accuracy of each phone arc in the word lattice.

3. The Smoothing Technique for MPELR

For MPELR, in order to prevent overfitting, ML statistics is used to smooth the discriminative statistics over each Gaussian component. To achieve this end, an extra term associated with the ML estimate is added to the auxiliary function, which is given by ignoring the parameters independent of $\hat{W}$:

$$\log P(\hat{W}) = \sum_{m=1}^{M} -\frac{1}{2}\left[\frac{\tau^{\mathrm{I}}}{\gamma_m^{ml}} \sum_t \gamma_m^{ml}(t)(O(t) - \hat{W}\xi_m)^T \Sigma_m^{-1}(O(t) - \hat{W}\xi_m)\right] \tag{20}$$

where $\gamma_m^{ml}(t)$ is the Gaussian occupation probability at time t and is obtained by ML training.

In this paper, we consider the prior of mean parameters. The log joint prior distribution of W and λ is added to the objective function for MPE. The multivariate Normal distribution as in Eq. (7) is used. Furthermore, hyper parameter V_m is set to be the unadapted variance Σ_m and β is set to be $\frac{1}{\tau^{\mathrm{I}}}$. Any function is both a weak and strong-sense auxiliary function of itself around any point [19]. So the auxiliary function for MPELR is written by ignoring the parameters independent of $\hat{W}$:

$$Q(W, \hat{W}) = Q_{\mathrm{MPE}}(W, \hat{W}) + \sum_{m=1}^{M} \tau^{\mathrm{I}}\left[-\frac{1}{2}(\hat{W}\xi_m - \eta_m)^T \hat{\Sigma}_m^{-1}(\hat{W}\xi_m - \eta_m)\right] \tag{21}$$

where τ^{I} is the smoothing factor. If η_m is obtained by ML training, the result is the same as the above smoothing method. The numerator statistics will be altered as follow:

$$\gamma_m^{\mathrm{num}'} = \gamma_m^{\mathrm{num}} + \tau^{\mathrm{I}}$$

$$\theta_m^{\mathrm{num}'}(O^{(i)}) = \theta_m^{\mathrm{num}}(O^{(i)}) + \tau^{\mathrm{I}} \frac{\theta_m^{\mathrm{ml}}(O^{(i)})}{\gamma_m^{\mathrm{ml}}} \tag{22}$$

where $\theta_m^{\mathrm{ml}}(O^{(i)})$ and γ_m^{ml} are the statistics calculated by ML training.

In the context of adaptation, it may not be robust to use ML estimates of state means to define the mode of the prior distribution. This is because there is limited data to estimate Gaussian parameters. In this case, it is preferable to use MAP estimates of the Gaussian parameters to smooth the MPE statistics. Hence, the numerator statistics for MPE-based mean transform estimation will be modified as follows:

$$\gamma_m^{\mathrm{num}'} = \gamma_m^{\mathrm{num}} + \tau^{\mathrm{I}}$$

$$\theta_m^{\mathrm{num}'}(O^{(i)}) = \theta_m^{\mathrm{num}}(O^{(i)}) + \tau^{\mathrm{I}} \frac{\theta_m^{\mathrm{ml}}(O^{(i)}) + \tau^{\mathrm{MAP}} \mu_m^{\mathrm{orig}}}{\gamma_m^{\mathrm{ml}} + \tau^{\mathrm{MAP}}} \tag{23}$$

where μ_m^{orig} is the speaker independent (SI) mean; τ^{MAP} is the prior weight factor.

4. Experiments

In this section, we describe the evaluation of the proposed smoothing method for MPELR for speaker adaptation on a large vocabulary continuous speech recognition task.

4.1 Experimental Setup

All data used in the experiment is from the National 863 High-Tech Project. The training data is about 65 hours from

140 speakers. The test data is from 6 speakers, 3 female and 3 male. There are 260 sentences for each speaker.

Speech signals were sampled at 8 kHz. The analysis frame is 25 ms wide with a 15 ms overlap. Each speech frame was parameterized into a 52-dimensional feature vector composed of 13 Mel-frequency-based perceptual linear prediction coefficients (MF-PLP, [23]) and the first, second and third order time derivatives of these features. Cepstral mean and variance normalization [24] was performed for all frames. Then a heteroscedastic linear discriminant analysis (HLDA) [25] transformation was applied to project these normalized features to a 39 dimensional space.

Acoustic models (AMs) used in experiments were state-tied, cross-word triphone HMM. The phone set is composed of 179 phonemes: 27 initials, 150 tonal finals, a silence (sil) and a short pause (SP). All acoustic units had a left-to-right topology. The SP model consisted of a single emitting state. The other models had three emitting states. The system had 5955 shared states resulted from a decision tree state tying [26]. Each state observation density was represented by an 8-component Gaussian mixture model (GMM). Each Gaussian component had a diagonal covariance. Two "initial" speaker independent (SI) models were trained: an MLE-trained system and an MPE-trained system. For the purpose of adaptation, the SI acoustic models provide both bases for transformation and the parameters of the prior distributions. A trigram language model was employed in experiments.

The lattice-based framework was employed in MPE training and MPE-based mean transform estimation. To generate lattices, a ML-trained HMM set was used for MPE training, the adapted HMM set from MLLR adaptation was used for MPELR. And a unigram language model was used to improve model generalization. The "exact-match" approach [27] was used to perform the forward-backward alignment to accumulate the statistics.

In the following experiments, supervised and unsupervised batch adaptation was performed. During parameter estimation, the smoothing values for MPELR adaptation and MPE training was chosen as $E = 2$.

4.2 Experimental Results

Firstly, the performance of the proposed smoothing technique for MPELR was evaluated. Table 1 shows the word error rate (WER) of adapting an ML-trained or MPE-trained initial HMMs set with MPELR. MPELR→ML and MPELR→MAP refer to the use of ML and MAP estimates of state means to define the mode of prior distribution respectively. Figure 1 gives the WER of MPELR→ML and MPELR→MAP adaptation from each test speaker with 4 utterances. We did preliminary experiments to determine the value of τ^I and τ^{MAP}. Firstly, in the MPELR→ML adaptation experiment, we tested performance to determine the value of τ^I. Then, τ^I was fixed, we did MPELR→MAP adaptation experiments to find the value of τ^{MAP} that gives the best performance. If $\tau^{\mathrm{MAP}} = 0$, MPELR→MAP adaptation is the same as MPELR→ML adaptation. Our setup in this set of experiments is: for MPELR→ML, $\{\tau^I = 50,\ \tau^{MAP} = 0\}$; for MPELR→MAP, $\{\tau^I = 50,\ \tau^{MAP} = 0.002\}$. Note that, during adaptation, only those Gaussians with a large value of γ_m^{ml} are used to smooth the MPE statistics. Gaussians with γ_m^{ml} below a certain threshold are not used for smoothing.

The results show that the values of smoothing statistics

Table 1 Word error rate (%) for batch supervised and unsupervised experiments for MPELR adaptation of ML and MPE trained model with various amount of data.

Training, Adaptation			Number of sentence for adaptation			
			0	4	5	6
supervised	ML	MPELR→ML	11	10	9.4	9.0
		MPELR→MAP	11	9.7	9.2	8.9
	MPE	MPELR→ML	8.1	8.3	7.5	7.2
		MPELR→MAP	8.1	7.6	7.2	7.0
unsupervised	ML	MPELR→ML	11	10.4	9.6	9.4
		MPELR→MAP	11	10	9.5	9.3
	MPE	MPELR→ML	8.1	8.5	7.6	7.3
		MPELR→MAP	8.1	7.7	7.2	7.1

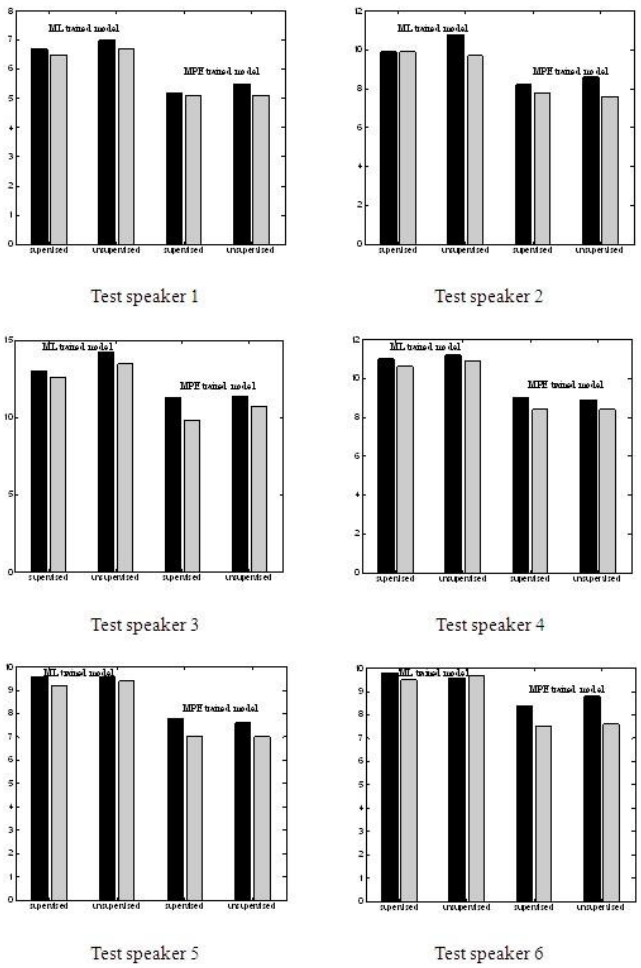

Fig. 1 Word error rate (%) for batch supervised and unsupervised experiments for MPELR→ML (black) and MPELR→MAP (white) of ML and MPE trained models with 4 adaptation utterances.

Table 2 Effect of τ^I and τ^{MAP} on MPELR→MAP.

		Number of sentence for adaptation			
		4	5	6	14
$\tau^{MAP} = 0.002$	$\tau^I = 10$	9.7	9.2	8.9	8.4
	$\tau^I = 50$	9.7	9.2	8.9	8.4
	$\tau^I = 100$	9.7	9.2	8.9	8.4
$\tau^I = 50$	$\tau^{MAP} = 0.1$	9.6	9.3	9.2	8.8
	$\tau^{MAP} = 0.01$	9.6	9.2	9.1	8.5
	$\tau^{MAP} = 0.002$	9.7	9.2	8.9	8.4

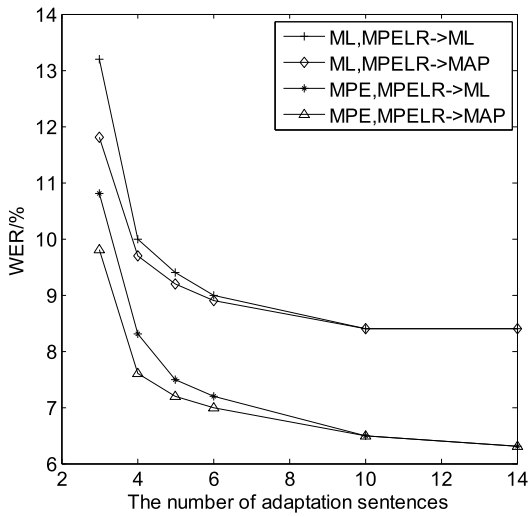

Fig. 2 Asymptotic property of MPELR.

are essential to the performance of MPELR adaptation given limited adaptation data. For example, in supervised experiment with only 4 (about 6.8s) adaptation sentences, 8.4% relative improvement can be obtained with MAP statistics over ML statistics by using an MPE-trained HMM set. For ML-trained HMM set, less improvement is obtained. It may be because more accurate statistics can be achieved by using MPE-trained HMM set. Secondly, with increasing amounts of adaptation sentences, the improvement from using MAP statistics for MPELR is partly lost. Third, in unsupervised adaptation experiments, MPELR→MAP still outperforms MPELR→ML. But there is a loss in accuracy compared to the supervised adaptation.

We investigate the effect of τ^I and τ^{MAP} on MPELR→MAP. Supervised adaptation experiments were performed. The results are showed in Table 2. It can be seen that τ^{MAP} is more influential on smoothing.

In the second experiment, the asymptotic property of the proposed smoothing method was evaluated. Figure 2 illustrates the effect of supervised MPELR adaptation with different smoothing statistics. From Fig. 2, we can see that the use of MAP statistics has the same asymptotic property as the use of ML statistics.

Hereafter, MAP statistics are used for smoothing in MPELR adaptation experiments.

Finally, MPELR adaptation was compared with the SI model and the flowing common adaptation methods:

Table 3 Word error rate (%) for batch supervised and unsupervised experiments for MLLR, MAPLR, DMAPLR and MPELR adaptation of ML and MPE trained model with various amount of data.

Training, Adaptation			Number of sentence for adaptation			
			0	4	5	6
supervised	ML	MLLR	11	11.1	9.9	9.6
		MAPLR	11	10.4	9.7	9.3
		DMAPLR	11	10.0	9.6	9.2
		MPELR→MAP	11	9.7	9.2	8.9
	MPE	MLLR	8.1	8.8	7.8	7.3
		MAPLR	8.1	7.9	7.3	7.1
		DMAPLR	8.1	7.4	7.2	6.9
		MPELR→MAP	8.1	7.6	7.2	7.0
unsupervised	ML	MLLR	11	11.7	10.2	10.1
		MAPLR	11	10.6	10	9.7
		DMAPLR	11	10.2	9.8	9.6
		MPELR→MAP	11	10	9.5	9.3
	MPE	MLLR	8.1	8.9	7.8	7.6
		MAPLR	8.1	8.1	7.4	7.2
		DMAPLR	8.1	7.5	7.2	7.2
		MPELR→MAP	8.1	7.7	7.2	7.1

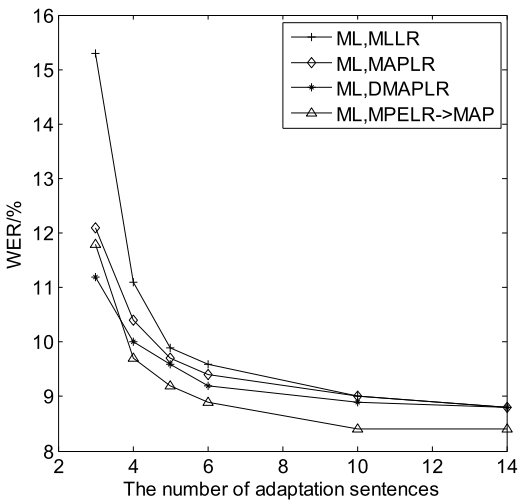

Fig. 3 Supervised MLLR, MAPLR, DMAPLR and MPELR adaptation of ML-trained model.

MLLR: the SA model from MLLR adaptation.

MAPLR: the SA model from MAPLR adaptation.

DMAPLR: the SA model from DMAPLR adaptation.

Table 3 and Figures 3–6 show the WER of adapting an ML-trained or MPE-trained initial HMM set with MLLR, MAPLR, DMAPLR and MPELR. For MAPLR, the scaling factor β was set to 5 (the prior weight was 0.2) and we chose the m-th mean vector and covariance matrix from SI HMM set as the hyper parameters η_m and V_m. For DMAPLR, we set $\varepsilon_n = 1/N$, $N = 8$ and $\{\alpha_1 = 0.4,\ \alpha_2 = 0.6,\ \beta = 2\}$. We can observe that:

(1) MPELR→MAP gives higher word error rates than DMAPLR with very limited adaptation data for supervised adaptation. But with increasing amounts adaptation data it outperforms DMAPLR.

(2) When limited adaptation data is available, DMAPLR outperforms MPELR→MAP for unsupervised adaptation. For both ML-trained and MPE-trained models,

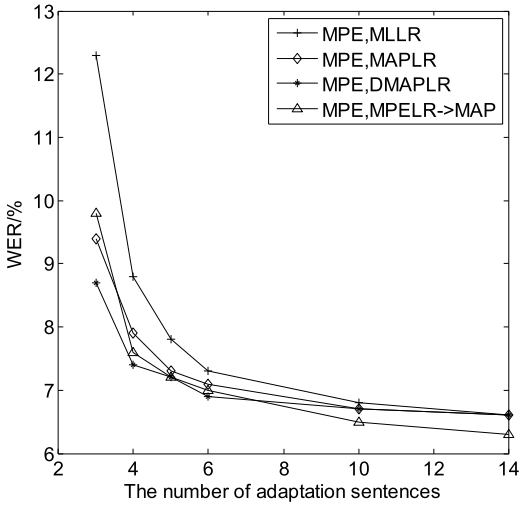

Fig. 4 Supervised MLLR, MAPLR, DMAPLR and MPELR adaptation of MPE-trained model.

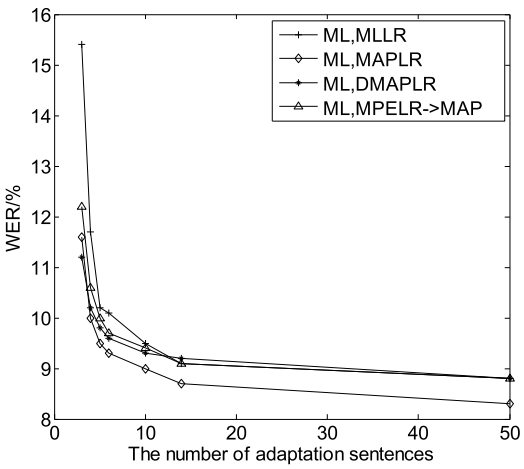

Fig. 5 Unsupervised MLLR, MAPLR, DMAPLR and MPELR adaptation of ML-trained model.

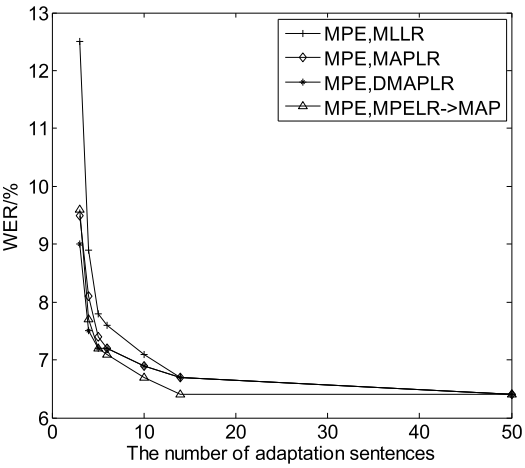

Fig. 6 Unsupervised MLLR, MAPLR, DMAPLR and MPELR adaptation of MPE-trained model.

as more adaptation data is available, MPELR→MAP shows improvement over DMAPLR. With large amounts of data, MPELR→MAP still shows better performance than DMAPLR in ML-trained system. While, for MPE-trained model, MPELR → MAP demonstrates no improvement.

(3) When the amount of speech data is limited MAPLR outperforms MLLR and DMAPLR outperforms MAPLR for both supervised and unsupervised adaptation. With increasing amounts of adaptation data DMAPLR and MAPLR converge asymptotically to MLLR. It may be because that by using the combination of LR-based objective function and ML criterion for transformation estimation, the improvement is not clear. With more adaptation data is available, MAP estimation converges to ML estimation. Therefore, the LR score has less effect on transformation estimation.

5. Conclusions

This paper has proposed a smoothing method for MPELR adaptation. A log prior mean distribution is combined with the objective function for MPE. The use of ML estimates of state means as 'prior' will lead to the conventional smoothing method for MPELR. Due to limited adaptation data, we proposed the use of MAP estimates to get more robust 'prior' and in turn to improve the performance of adaptation. Supervised and unsupervised speaker adaptation experiments were conducted on a large vocabulary continuous speech recognition task. Results show that, with limited amount of adaptation data, the use of MAP statistics for smoothing can considerately improve the performance of MPELR adaptation. Moreover, for MPELR, the use of MAP estimates of state means as 'prior' has the same asymptotic property as that of ML estimates. Moreover the proposed method outperforms MLLR and shows better recognition performance that MAPLR and DMAPLR with increasing amounts of adaptation data for both supervised and unsupervised adaptation. For ML-trained model, the proposed method still shows improvement with large amounts of data. But, for MPE-trained model, the improvement is not clear for unsupervised adaptation when large amounts of data is available.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001, 61271426) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500).

References

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang., vol.9, no.2, pp.171–185, 1995.

[2] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Comput. Speech Lang., vol.12, no.2, pp.75–98, 1998.
[3] K. Shinoda, "Speaker adaptation techniques for automatic speech recognition," Proc. APSIPA ASC, Xi'an, China, 2011.
[4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Guassian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[5] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," Proc. EUROSPEECH, pp.211–214, Budapest, Hungary, 1999.
[6] C.-H. Lin and W.-J. Wang, "Maximum a posteriori linear regression for speaker adaptation with the prior of mean," Proc. EUPSICO, 2000.
[7] Y. Tsao, R. Isotani, H. Kawai, and S. Nakamura, "An environment structuring framework to facilitating suitable prior density estimation for MAPLR on robust speech recognition," Proc. ISCSLP, pp.29–32, Tainan, Taiwan, 2010.
[8] T.-Y. Hu, Y. Tsao, and L.-S. Lee, "Discriminative fuzzy clustering maximum a posteriori linear regression for speaker adaptation," Proc. Interspeech, Portland, USA, 2012.
[9] Y. Tsao, C.-L. Huang, S. Matsuda, C. Hori, and H. Kashioka, "A linear projection approach to environment modeling for robust speech recognition," Proc. ICASSP, pp.4329–4332, Kyoto, Japan, 2012.
[10] O. Siohan, T.A. Myrvoll, and C.-H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," Comput. Speech Lang., vol.16, pp.5–24, 2002.
[11] L.F. Uebel and P.C. Woodland, "Discriminative linear transforms for speaker adaptation," Proc. ISCA ITRW on Adaptation Methods for Automatic Speech Recognition, pp.61–63, Sophia-Antipolis, France, 2001.
[12] L. Wang and P.C. Woodland, "MPE-based discriminative linear transform for speaker adaptation," Proc. ICASSP, pp.321–324, Quebec, Canada, 2004.
[13] Sh. Pirhosseinloo and Sh. Javadi, "A combination of maximum likelihood Bayesian framework and discriminative linear transforms for speaker adaptation," Int. J. Information and Electronics Engineering, vol.2, no.4, pp.552–555, 2012.
[14] J. Wu and Q. Huo, "A study of minimum classification error (MCE) linear regression for supervised adaptation of MCE-trained continuous-density hidden Markov models," IEEE Trans. Audio Speech Language Process., vol.15, no.2, pp.478–488, 2007.
[15] X. He and W. Chou, "Minimum classification error linear regression for acoustic model adaptation of continuous density HMMs," Proc. ICASSP, pp.556–559, HongKong, 2003.
[16] B. Zhu, Z.-J. Yan, Y. Hu, Z.-G. Wang, L.-R. Dai, and R.-H. Wang, "Investigation on adaptation using different discriminative training criteria based linear regression and MAP," Proc. ISCSLP, pp.93–96, Kunming, China, 2008.
[17] S. Matsuda, Y. Tsao, J. Li, S. Nakamura, and C.-H. Lee, "A study on soft margin estimation of linear regression parameters for speaker adaptation," Proc. Interspeech, pp.1603–1606, Brighton, UK, 2009.
[18] Y. Tsao, R. Isotani, H. Kawai, and S. Nakamura, "Increasing discriminative capability on MAP-based mapping function estimation for acoustic model adaptation," Proc. ICASSP, pp.5320–5323, Prague, Czech, 2011.
[19] D. Povey, M.J.F. Gales, D.Y. Kim, and P.C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," Proc. Interspeech, pp.1981–1984, Geneva, Switzerland, 2003.
[20] M.J.F. Gales, D. Pye, and P.C. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," Proc. ICSLP, vol.3, pp.1832–1835, Philadelphia, PA, 1996.
[21] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," Comput. Speech Lang., vol.10, pp.249–264, 1996.
[22] S.-H. Liu, F.-H. Chu, Y.-T. Lo, and B. Chen, "Improved minimum phone error based discriminative training of acoustic model for Mandarin large vocabulary continuous speech recognition," Computational Linguistics and Chinese Language Processing, vol.13, no.3, pp.343–362, 2008.
[23] P.C. Woodland, M.J.F. Gales, D. Povey, and S.J. Young, "Broadcast news transcription using HTK," Proc. ICASSP, Munich, Germany, vol.2, pp.719–722, 1997.
[24] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Commun., vol.25, pp.133–147, 1998.
[25] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," Speech Commun., vol.26, pp.283–297, 1998.
[26] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," Proc. Workshop on Human Language Technology, pp.307–312, 1994.
[27] P.C. Woodl and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," Comput. Speech Lang., vol.16, pp.25–47, 2002.

Yaohui Qi received her M.S. degree in Communication and Information System from Hebei University. Now she is a Ph.D. candidate in College of Information and Electronics at Beijing Institute of Technology. Her research interests include speech recognition, speech signal processing.

Fuping Pan received his Ph.D. in Information and Signal Processing from Institute of Acoustics (IOA), Chinese Academy of Science (CAS), in 2007. Now he is an Associate Researcher in IOA. His research interests include automatic pronunciation evaluation, speech signal processing and speech recognition.

Fengpei Ge received her Ph.D. in Information and Signal Processing from IOA, CAS, in 2010. Now she is an Assistant Researcher in IOA. Her research interests include automatic pronunciation evaluation, speech signal processing and speech recognition.

Qingwei Zhao received his Ph.D. in Electronic Engineering from Tsinghua University in 1999. Now he is an Associate Professor at Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS. Before joining CAS, he was with Intel as senior researcher. His research interests include voice search, spontaneous speech recognition and spoken term detection.

Yonghong Yan received his B.E. From Tsinghua University in 1990 and his Ph.D. From Oregon Graduate Institute (OGI) in 1995. From 1995 to 1998, he worked in OGI as Assistant Professor, Associate Director and Associate Professor of the Center for Spoken Language Understanding and from 1998 to 2001 he worked as Principal Engineer of Intel Microprocessors Research Lab, Director and Chief Scientist of Intel China Research Center. Now he is a professor and director of Key Laboratory of Speech Acoustics and Content Understanding, IOA, CAS. His research interests include large vocabulary speech recognition, speaker/language recognition and audio signal processing. He has published more than 100 papers and holds 40 patents.