# A Combination of Maximum Likelihood Bayesian Framework and Discriminative Linear Transforms for Speaker Adaptation

Sh. Pirhosseinloo and Sh. Javadi

*Abstract*—**Linear transforms are one of the most commonly used methods to speaker adaptation. In this paper, we present a combinational method of Bayesian framework and maximum likelihood linear regression as well as discriminative method for speaker adaptation. Furthermore significant gains can be obtained using discriminative training for acoustic models. Experiments on supervised adaptation on Persian data show that the combinational method outperforms both Maximum likelihood linear regression and Bayesian framework. Also the proposed method with discriminative adaptation outperforms previously proposed methods for transform estimation and discriminative training outperforms ML training.**

*Index Terms*—**Discriminative linear transforms, maximum-a-posterior adaptation, maximum likelihood linear regression adaptation, speech recognition, speaker adaptation.**

## I. Introduction

Speaker adaptation is an important part of automatic speech recognition systems. Linear transforms are widely used for model adaptation in HMM-based systems. MLLR is a popular method estimating the parameters by maximizing the likelihood of generating the adaptation data given the transformed model [1], [2]. Audio data and transcription are required for estimating the linear transforms. If the correct transcriptions are available, the adaptation operates in supervised mode. If there is no transcription available for data, the adaptation is unsupervised. In this paper we focus on supervised adaptation.

Maximum likelihood (ML) criterion was used to estimate linear transforms. Discriminative criteria such as Maximum Mutual Information (MMI) and Minimum Phone Error (MPE) [3], [4] are commonly used to train HMM systems. Training models with discriminative criteria reduce Word Error Rate (WER) significantly [5], [6]. Discriminative training such as MPE has been successfully used to train acoustic models, hence it is expected that this criterion is able to improve the estimation of the linear transforms for speaker adaptation [7]. In unsupervised adaptation, the performance results of discriminative adaptation have been limited, as these criteria are sensitive to errors in the hypotheses rather than the ML criterion.

Manuscript received April 19, 2012; revised May 31, 2012.

Sh. Pirhosseinloo is the member of Scientific Association of Electrical and Electronic Engineering, Islamic Azad University Central Tehran Branch, and Tehran, Iran (e-mail: sh.pier@ yahoo.com).

Sh.Javadi is the assistant Professor of Electrical Engineering Department Islamic Azad University Central Tehran Branch, Tehran, Iran (e-mail: sh.javadi@iauctb.ac.ir).

In this paper acoustic models are trained with MPE discriminative criterion. The maximum-a-posterior (MAP) estimation has been proposed for robustly estimating MLLR transforms with small amount of adaptation data. Then discriminative linear transforms (DLT) are estimated to adapt Gaussian means. The optimization of the transform parameters is employed to minimize the errors on the adaptation data. Furthermore it is necessary to smooth the discriminative criteria with statistics used in ML estimation. The I-smoothing improve the generalization of MPE-based discriminative linear transform [8].

The rest of this paper is organized as follows. In section 2 MLLR and maximum likelihood Bayesian framework are described. In section 3 the theory of MPE-based DLT estimation including the use of weak-sense auxiliary function for optimization is presented. Experiment results on Persian data are described in section 4. Finally a summery and conclusion are presented in section 5.

## II. Maximum-Likelihood Bayesian Adaptation

Linear transform based speaker adaptation was initially investigated with ML estimation. In MLLR adaptation, the mean μ of the model parameters is transformed to speaker-adapted mean μ̃ (s) as:

$$\tilde{\mu}^{(s)} = A^{(s)}\mu + b^{(s)} = W^{(s)}\xi \tag{1}$$

where $W^{(s)}=[A^{(s)}\ b^{(s)}]$ is the linear transform and $\xi=[\mu^T\ 1]^T$ is the extended mean vector .W is a n×(n+1) matrix (n is the dimension of the features). The parameters of transform, $W^{(s)}$ are estimated using ML criterion,

$$w^{(s)} = \arg\max_{w}\{p(o^{(s)} \mid H^{(s)}, W; \lambda)\} \tag{2}$$

where $O^{(s)}$ and $H^{(s)}$ are the observations and reference of the adaptation data for speaker s respectively.

MAP adaptation [9] involves the use of prior knowledge about the model parameter distribution. One obvious drawback to MAP adaptation is that it requires more adaptation data to be effective compared to MLLR. In fact the two adaptation processes can be combined to improve performance by using MLLR transformed means as the priors for MAP adaptation. The MAP estimation can be seen as a Bayesian estimation: given a set of n speech feature vectors $O=(O_1,...,O_n)$ , if H is the parameter vector to be estimated from O , with probability density function (pdf) given by f(O| H) and g is the prior pdf of H , it is possible to

estimate $H_{MAP}$ as:

$$H_{MAP} = \arg\max_{H}\{f(O|H)g(H)\} \tag{3}$$

where the acoustic score is marginal likelihood given as :

$$f(O|H) = \int f(O|H,W)p(W|\varphi_{ML})dW \tag{4}$$

The transform prior p(W| $\varphi_{ML}$) is a Gaussian for mean MLLR transforms. The MAP points estimates of ML transforms are obtained as:

$$\hat{W}_{MAP}^{(H)} = \arg\max_{w}\{p(O|H,W)p(W|\phi_{ML})\} \tag{5}$$

Bayesian method yield robust estimates of ML-based transforms and lead to reduction in WER [10].

## III. MPE Criterion for Discriminative Linear Transforms

MPE criterion had been proposed to evaluate the phone accuracy in the word context. The MPE objective function was defined in [8], [11]:

$$F_{MPE}(\lambda) = \sum_{r=1}^{R}\frac{\sum_{w'}P_{\lambda}(O_r|\mathrm{M}^{w'})^{\kappa}P(w')RawAccuracy(w')}{\sum_{w}P_{\lambda}(O_r|\mathrm{M}^{w})^{\kappa}P(w)} \tag{6}$$

where $M^w$ is the model corresponding to the word sequence w , P(w) is the probability of the word sequence w and κ is the acoustic scale. The RawAccuracy(w`) measures the number of phones correctly recognized in the sentence according to reference phones. Transforms estimated with discriminative criteria are referred to discriminative linear transforms (DLTs). The form of adaptation remains as MLLR:

$$\tilde{\mu}^{(s)} = A_{dl}^{(s)}\mu + b_{dl}^{(s)} = W_{dl}^{(s)}\xi \tag{7}$$

where $W_{dl}^{(s)}=[A_{dl}^{(s)}\ b_{dl}^{(s)}]$ is the DLT of speakers. DLTS are estimated using MPE criterion which can be expressed as:

$$W_{dl}^{(s)} = \arg\max_{w}\{\sum_{H}P(H|O^{(s)},w;\lambda)L(H,H^{(s)})\} \tag{8}$$

where $P(H|O^{(s)},w)$ is the posterior probability of hypothesis H from speaker s and $L(H,H^{(s)})$ is the loss function of H given the supervision $H^{(s)}$ measured at the phone level [11].

For optimization of discriminative criteria the weak-sense auxiliary function was proposed [8].Given the objective function F (λ), the weak-sense auxiliary function is defined to satisfy the following condition,

$$\frac{\partial}{\partial\lambda'}Q(\lambda,\lambda')|_{\lambda=\lambda'} = \frac{\partial}{\partial\lambda'}F(\lambda')|_{\lambda=\lambda'} \tag{9}$$

where λ is the original parameter set and λ′ is the newly estimated parameter. Optimizing the weak-sense auxiliary function doesn't guarantee an increase in the objective function [8]. This auxiliary function is based on log likelihood of phone arc to make the optimization tractable according to the phone accuracy in the objective function,

$$Q_{MPE}(\lambda,\lambda') = \sum_{r=1}^{R}\sum_{q=1}^{Q_r}\frac{\partial F_{MPE}}{\partial\log p(q)}|_{(\lambda=\lambda')}\log p(q) \tag{10}$$

Each sentence r consist of phone arcs q=1,...,$Q_r$, and p(q) is the likelihood of arc q. The auxiliary function consists of three parts as:

$$\begin{aligned}&Q_{MPE}(W,W')\\&=\sum_{r=1}^{R}\sum_{q=1}^{Q_r}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(\gamma_q^{MPE})\log\mathrm{N}(o(t),\hat{W}\xi_m,\Sigma_m)\\&-\sum_{r=1}^{R}\sum_{q=1}^{Q_r}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(-\gamma_q^{MPE})\log\mathrm{N}(o(t),\hat{W}\xi_m,\Sigma_m)\\&+Q_{sm}(W,\hat{W})\end{aligned} \tag{11}$$

where $\gamma_q^{MPE} = \frac{1}{\kappa}\frac{\partial F_{MPE}}{\partial\log p(q)}$ as defined for MPE training. $\gamma_{qm}(t)$ is the posterior probability over time t at state j, mixture component m of arc q. The function f ($\gamma_q^{MPE}$) defined as below:

$$\begin{cases}f(\gamma_q^{MPE}) = \max(0,\gamma_q^{MPE})\\f(-\gamma_q^{MPE}) = \max(0,-\gamma_q^{MPE})\end{cases} \tag{12}$$

With calculating the differential of (11) with respect to each row of the linear transforms $\hat{w}^{(i)}$ we have:

$$\begin{aligned}\hat{w}^{(i)} &= G^{(i)^{-1}}k^{(i)}\\G^{(i)} &= \sum_{m}\frac{1}{\sigma_{m(i)}^2}((\gamma_m^{num}-\gamma_m^{den})+D_m)\xi_m^T\xi_m\\k^{(i)} &= \sum_{m}\frac{1}{\sigma_{m(i)}^2}((\theta_m^{num}(O_{(i)})-\theta_m^{den}(O_{(i)})+D_m\tilde{\mu}_{m(i)})\xi_m^T\end{aligned} \tag{13}$$

where $D_m$ is the smoothing factor with a constant E.

$$D_m = E\sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(-\gamma_q^{MPE}) \tag{14}$$

The numerator statistics [8] to estimate MPE-based DLT had following forms:

$$\begin{aligned}\gamma_m^{num} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(\gamma_q^{MPE})\\\theta_m^{num} &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(\gamma_q^{MPE})o(t)\\\theta_m^{num}(o^2) &= \sum_{q=1}^{Q}\sum_{t=s_q}^{t=e_q}\gamma_{qm}(t)f(\gamma_q^{MPE})o^2(t)\end{aligned} \tag{15}$$

The denominator statistics had following forms [8]:

$$\gamma_m^{den} = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE})$$

$$\theta_m^{den} = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) o(t)$$

$$\theta_m^{den}(o^2) = \sum_{q=1}^{Q} \sum_{t=s_q}^{t=e_q} \gamma_{qm}(t) f(-\gamma_q^{MPE}) o^2(t) \qquad (16)$$

The I-smoothing is used to prevent over-training and improve model generalization. The I-smoothing use ML statistics as a "prior" to smooth the discriminative statistics over each Gaussian component. The ML statistics are calculated using the numerator lattices according to the correct transcription. Hence, only the numerator statistics for MPE-based mean transform estimation are altered in [8]:

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

$$\theta_m^{num'}(O) = \theta_m^{num}(O) + \frac{\tau}{\gamma_m^{ml}} \theta_m^{ml}(O)$$

$$\theta_m^{num'}(O^2) = \theta_m^{num}(O^2) + \frac{\tau}{\gamma_m^{ml}} \theta_m^{ml}(O^2) \qquad (17)$$

where $\gamma_m^{ml}$ is the state occupation probability calculated by ML training. This prior is proportional to the likelihood of $\tau$ observation points. The ML statistics can be calculated using the numerator lattices corresponding to the correct transcriptions.

## IV. Experiments

The acoustic models used in experiments are gender-independent continuous mixture density, tied state cross-word triphone HMMs. The training dataset consisted of 250 speakers about 4 hours of data. The test set consists of 50 speakers about 1 hour. All systems used a 39-dimensional MFCC frond-end with $C_0$ energy and its first, second and third derivatives. The gender independent cross-word triphone HMMs consist of 4499 tied states. Speaker independent (SI) model sets were obtained using ML, MMI and MPE criteria. A mean transform was used in all experiments for supervised adaptation.

The lattice-based framework as used in MPE training is employed here for estimating MPE-based DLT. Initially, word lattices are generated on adapted models (using maximum-likelihood Bayesian framework) with unigram language model. The lattices used were generated by the HTK recognition system. The numerator and denominator lattices were generated. Then the denominator and numerator phone-level lattices are created by aligning the recognized word lattices and correct transcription separately with a unigram language model. The appropriate statistics for the MPE-based DLT were gathered via a forward-backward pass through the lattice marked with the phone starting/ ending times. For optimization of discriminative criteria the I-smoothing is employed.

The smoothing values for I-smoothing were chosen as E=2 and $\tau$=100. The scale factor is chosen as $\kappa$ =1.

The experimental results for ML and discriminative training with MMI and MPE criteria are given in Table 1.

TABLE I: WER(%) ON PERSIAN DATA AFTER ML , MMI AND MPE TRAINING

| System | Training | | |
|---|---|---|---|
| | ML | MMI | MPE |
| SI | 14.05 | 12.32 | 11.85 |

It is observed that MPE discriminative training can reduce the WER in comparison with ML and MMI training. MPE discriminative training gave 2.2% reduction in WER over standard ML trained model and 0.47% reduction over MMI discriminative trained model. The smoothing values for estimation of discriminative linear transforms were chosen same as discriminative training.

The experimental results for different adaptation methods on MPE-based discriminative trained models are given in Table 2. Different speaker adaptation such as MLLR, MAP, MLLRMAP and MPE-based DLT were used for speaker adaptation.

TANLE II :WER(%) ON PERSIAN DATA AFTER MLLR , MAP AND MLLRMAP ADAPTATION

| System | Adaptation | WER (%) | |
|---|---|---|---|
| | Testing | ML | MPE |
| SI | MAP | 14.41 | 11.27 |
| | MLLR | 13.39 | 9.2 |
| | MAPMLLR | 12.9 | 8.8 |
| | MAPMLLR+DLT | 12.4 | 8.2 |

It is observed that MAP estimates of ML transforms reduced the WER in comparison on MLLR and MAP. As it can be observed the MPE-based DLT with MLLRMAP gave a 1% reduction in WER over standard MLLR and 0.6% reduction over the MAPMLLR system.

## V. Conclusions

This paper has investigated a combinational method of Bayesian framework with MLLR for speaker adaptation. Furthermore discriminative linear transforms had been used for improving the speaker adaptation results. The experimental results on Persian data have shown that the ML Bayesian framework can improve the supervised adaptation performance in comparison with MAP and MLLR adaptation. Also the discriminative linear transforms have been applied to MAPMLLR adaptation. Experiments illustrated that DLT on combinational adaptation outperform MAPMLLR in supervised adaptation.

## Acknowledgment

Authors are grateful to Dr. A .Kashaninia and Dr. F. Farokhi for their constant patient support and useful suggestions.

## References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp.171–186, 1995.

[2] M. J. F.Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and Language*, vol.12,pp.75–98,1998.

[3] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc.ICASSP*, Orlando ,2002.
[4] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–48, 2002.
[5] L.Wang and P. C. Woodland, "Discriminative adaptive training using the MPE criterion," *in Proc, ASRU*,St. Thomas, 2003.
[6] D. Povey, "Discriminative training for large vocabulary speech recognition," PHD. dissertation, Cambridge University, 2003.
[7] L. F. Uebel and P. C. Woodland, "Discriminative linear transforms for speaker adaptation," in *Proc. ICSA ITR-Workshop on Adaptation Methods for Speech Recognition*, 2001.
[8] L. Wang and P. C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Computer Speech and Language*, vol. 22,no. 3, pp. 256–272, 2008.
[9] C. Chesta, O. Siohan and C. Lee, " Maximum a posterior linear regression for hidden Markov model adaptation," *in Proc.Eurospeech*, vol.1, pp.211–214, 1999.
[10] C. K. Raut and M. J. F. Gales, " Bayesian discriminative adaptation for speech recognition ," *in Proc. ICASSP*, Taipei, 2009.
[11] K.Yu, M.J.F. Gales and P.C. Woodland, " Unsupervised discriminative adaptation using discriminative mapping transforms," in *Proc. ICASSP*, Las Vegas, 2008.

**Shadi Pirhosseinloo** received her Bachelor of Engineering in Electrical Engineering from Islamic Azad University Central Tehran Branch, Tehran, Iran. Her research interests include speech recognition systems and speaker adaptation. She works as an electrical engineer in design of power electrical systems for factories. She is the member of Scientific Association of Electrical & Electronic Engineering in Azad university of Iran.