# ROBUST FEATURE SPACE ADAPTATION FOR TELEPHONY SPEECH RECOGNITION

*Xin Lei* †

Electrical Engineering Department
Univ. of Washington, Seattle, WA 98195
leixin@ee.washington.edu

*Jon Hamaker, Xiaodong He*

Microsoft Speech and Natural Language Group
One Microsoft Way, Redmond, WA 98052
{jonham, xiaohe}@microsoft.com

**ABSTRACT**

Speaker adaptation is critical for modern speech recognition systems. Due to the computational and multi-channel model sharing considerations, the use of model adaptation techniques is limited in telephony speech recognition systems. On the other hand, feature space adaptation methods such as feature space maximum likelihood linear regression (fMLLR) are efficient approaches suitable for telephony systems. In this work, we first describe techniques for efficient implementation of online fMLLR adaptation. Then feature space maximum *a posteriori* linear regression (fMAPLR) is proposed to incorporate prior knowledge for the feature transform estimation and improve the robustness of the conventional fMLLR approach. Experiments on telephony data indicate that fMAPLR is significantly more robust than fMLLR, and outperforms fMLLR especially when the adaptation data is very limited.

**Index Terms**: Speaker adaptation, telephony, speech recognition.

## 1. INTRODUCTION

Speaker adaptation is essential for modern speech recognition systems, especially when there are significant mismatches between the training and decoding conditions. Many adaptation methods have been proposed to compensate for channel and speaker variations. Maximum likelihood linear regression (MLLR) [1] and maximum *a posteriori* (MAP) [2] based adaptation are the most popular approaches. MAP based adaptation incorporates prior knowledge about the distribution of the model parameters to help robust adaptation of model parameters, and it converges to maximum likelihood estimates when adaptation data increases. In MLLR adaptation, a set of linear transformation matrices is estimated to transform the model parameters and maximize the likelihood on the adaptation data. Furthermore, in order to obtain robust estimation of the linear transforms, maximum *a posteriori* linear regression (MAPLR) estimation has also been proposed to effectively adapt model parameters [3, 4, 5].

These methods work well for speaker adaptation. However, in a real time telephony speech recognition system, model space adaptation techniques are not preferred because application in the model space requires the expensive operations of saving, updating and re-quantizing the speaker-dependent (SD) model parameters. Therefore, as a dual of constrained MLLR adaptation, an efficient feature space maximum likelihood linear transform (fMLLR) based speaker adaptation method was first proposed in telephony applications [6]. By using fMLLR, it is only necessary to apply a linear transform to the feature vectors for every frame.

Maximum likelihood estimation is a data driven parameter estimation approach. When adaptation data is very limited, the estimated linear transform is often unreliable and may cause the adapted system to have even worse performance than the baseline system. In [6], in order to address this problem, a variant of discounted likelihood linear regression (DLLR) [7] with smoothing statistics obtained from the speaker independent acoustic model is used to smooth the fMLLR statistics. In this work, we derive the feature space MAP (fMAPLR) as a counterpart of the model space MAPLR. We find the fMAPLR formulation is similar to MAPLR in modifying the sufficient statistics with prior information.

The rest of the paper is organized as follows: In Section 2 we review the feature space MLLR algorithms and describe the implementation issues for real time telephony applications. Section 3 presents the fMAPLR solution and discusses its relationship to fMLLR. In Section 4, the hyperparameter estimation of the prior distribution of the linear transform is described. We show the experimental results in Section 5. The paper is concluded in Section 6.

## 2. FMLLR AND IMPLEMENTATION

### 2.1. Algorithm for fMLLR

Feature space maximum likelihood linear regression for online telephony application was proposed in [6]. As shown in [8], the constrained model space linear transform is equivalent to a feature space linear transform when a single transformation is used. Let $o_t$ be the $n$-dimensional feature vector at time $t$ in the original feature space, the transformed feature $\hat{o}_t$ is,

$$\hat{o}_t = Ao_t + b = W\xi_t, \tag{1}$$

where $A$ is the $n \times n$ rotation matrix, $b$ is the $n \times 1$ bias term, $\xi_t = [1 \quad o_t^T]^T$ is the $(n+1) \times 1$ extended observation vector

† The author was at Microsoft when this work was performed.

and $W = [b \quad A]$ is the $n \times (n+1)$ extended transformation matrix. The transform parameters are estimated by optimizing the following auxiliary Q-function,

$$Q_{ML} = -\frac{1}{2}\sum_{t,m}\gamma_m(t)\{\log|A|^2 + (W\xi_t - \mu^{(m)})^T\Sigma^{(m)-1}(W\xi_t - \mu^{(m)})\}, \quad (2)$$

where $\mu^{(m)}$ and $\Sigma^{(m)}$ are the mean and covariance for Gaussian component $m$ and $\gamma_m(t)$ is the posterior probability of being in Gaussian $m$ at time $t$.

Because of the log determinant in the objective function, generally there is no explicit closed-form solution for the transformation matrix $W$. In this work we chose to follow the iterative solution in [8]. We assume the covariance matrices to be diagonal: $\Sigma^{(m)} = \text{diag}([1/\sigma_1^{(m)2} \quad 1/\sigma_2^{(m)2} \ldots 1/\sigma_n^{(m)2}])$. Let the $i$-th row of $W$ be $w_i = [W_{i1} \quad W_{i2} \ldots W_{in}]^T$. Taking the derivative of $Q_{ML}$ with respect to $w_i$ and equating to zero, we can get,

$$\frac{\partial Q_{ML}}{\partial w_i} = \beta\frac{p_i}{p_i^T w_i} - G^{(i)}w_i + k^{(i)} = 0, \quad (3)$$

where $\beta = \sum_{t,m}\gamma_m(t)$ is the total count, $p_i$ is the extended cofactor vector $[0 \quad \text{cof}(A_{i1}) \ldots \text{cof}(A_{in})]^T$ and the sufficient statistics of $G^{(i)}$ and $k^{(i)}$ are as follows:

$$G^{(i)} = \sum_t \xi_t\xi_t^T \sum_m \frac{\gamma_m(t)}{\sigma_i^{(m)2}} \quad (4)$$

$$k^{(i)} = \sum_t \xi_t \sum_m \frac{\gamma_m(t)\mu_i^{(m)}}{\sigma_i^{(m)2}}. \quad (5)$$

By using the direct method over rows [8], we get an iterative solution,

$$w_i = G^{(i)-1}(\alpha p_i + k^{(i)}), \quad (6)$$

where $\alpha$ is solved from the following quadratic equation and the root that maximizes the Q-function is selected.

$$\alpha^2 p_i^T G^{(i)-1} p_i + \alpha p_i^T G^{(i)-1} k^{(i)} - \beta = 0 \quad (7)$$

Note that equation 3 and 6 are slightly different from [8] since in our notation all vectors are column vectors unless transposed explicitly.

### 2.2. Implementation Issues

For the real time application of fMLLR in multiple parallel channels, memory usage and computational complexity need to be optimized. Unlike in the model space mean MLLR, the sufficient statistic $G^{(i)}$ in equation 4 is a second-order statistic. If we store the statistics at the Gaussian level, we need to store $\mathcal{O}(n^2)$ parameters for each Gaussian, which amounts to $n$ times the model size. This is not affordable for reasonably large acoustic models. Therefore, we chose to store the statistics in a global $G^{(i)}$ matrix and $k^{(i)}$ vector. The disadvantage of global storage is that we need $\mathcal{O}(n^3)$ multiply accumulates per Gaussian component per frame [8].

In addition to the statistics accumulation, the fMLLR transform estimation itself is also computationally expensive. Various methods have been proposed to decrease the computational complexity of feature space adaptation for real-time applications. For example, in [9] the author proposed to use stochastic gradient descent. Some engineering trade-offs can also speed up fMLLR significantly. First, by using block diagonal transforms we can decrease the computation of the feature transforms and significantly decrease the complexity of accumulating the global sufficient statistics. In all our experiments, we have used block diagonal transforms with 3 blocks. Second, considering that $G^{(i)}$ is real symmetric and positive definite, Choleski decomposition can be used to speed up the matrix inversion process. Finally, it is possible to accumulate statistics only from the Gaussian component with the highest posterior probability. In practice, we have found that the performance difference is minor but the speedup is significant.

## 3. FEATURE SPACE MAXIMUM A POSTERIORI LINEAR REGRESSION

In some telephony applications like name dialing, each phone call lasts for only a few utterances and the data available for adaptation is very limited. This usually leads to a biased fMLLR adaptation due to overtraining. In order to address this robustness issue, we apply the maximum *a posteriori* framework and derive the feature space maximum *a posteriori* linear regression (fMAPLR) based speaker adaptation as a counterpart of the model space MAPLR.

The auxiliary Q-function for fMAPLR with prior matrix distribution of $p(W)$ is given by

$$Q_{MAP} = Q_{ML} + \log p(W), \quad (8)$$

We assume the feature transformation matrix $W$ follows an elliptically symmetric matrix variate distribution [3],

$$p(W) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(w_i - \mathcal{M}_i)^T\mathcal{V}_i^{-1}(w_i - \mathcal{M}_i)\right], \quad (9)$$

where $\mathcal{M}_i$ is the location parameter and $\mathcal{V}_i$ is the scale parameter for $w_i$. $\mathcal{M}_i$ and $\mathcal{V}_i$ are called the hyperparameters of the prior distribution.

Taking the derivative of equation 8 and substituting in equation 3 and equation 9, we have:

$$\frac{\partial Q_{MAP}}{\partial w_i} = \beta\frac{p_i}{p_i^T w_i} - \hat{G}^{(i)}w_i + \hat{k}^{(i)} = 0, \quad (10)$$

where

$$\hat{G}^{(i)} = G^{(i)} + \mathcal{V}_i^{-1} \quad (11)$$

$$\hat{k}^{(i)} = k^{(i)} + \mathcal{V}_i^{-1}\mathcal{M}_i \quad (12)$$

Equation 10 has the same form as equation 3. Therefore, we can estimate the fMAPLR transform in the same iterative way as in fMLLR (equation 6), but with different statistics of $\hat{G}^{(i)}$ and $\hat{k}^{(i)}$. In practice, we have found that four iterations is sufficient for the convergence of both fMLLR and fMAPLR.

The new statistics are a smoothed version of the fMLLR statistics with the prior knowledge about the transform distribution incorporated. Moreover, if we compare fMAPLR and fMLLR, it is clear that, when the adaptation data amount is very small, equation 11 of fMAPLR is dominated by the statistics from the prior distribution. When more adaptation data is available, equation 11 converges to equation 4. In this way, the fMAPLR provides robustness to small amounts of adaptation data.

## 4. PRIOR DISTRIBUTION HYPERPARAMETER ESTIMATION

One issue of MAP estimation is the estimation of the hyperparameters of the prior distribution. In a strict Bayesian approach, the hyperparameters are assumed known based on a common or subjective knowledge about the stochastic process. In most cases it is difficult to obtain this common knowledge about the informative prior distribution. As a popular solution, the empirical Bayesian approach is widely used where the hyperparameters are learned from the data.

Assume there are $K$ observations of the transformation matrices $\{W^{(1)}, \ldots W^{(K)}\}$, the hyperparameters can be estimated by,

$$\mathcal{M}_i = \frac{1}{K}\sum_{r=1}^{K} w_i^{(r)} \quad (13)$$

$$\mathcal{V}_i = \frac{1}{K}\sum_{r=1}^{K}(w_i^{(r)} - \mathcal{M}_i)(w_i^{(r)} - \mathcal{M}_i)^T \quad (14)$$

where $w_i^{(r)}$ is the $i$-th row of the matrix $W^{(r)}$.

In model space MAPLR, researchers have proposed to derive the set of transformation matrices from the speaker-independent (SI) acoustic models [4] or from the adaptation data [3, 5]. Both methods make use of the adaptation class tree structure to generate a set of matrices for each adaptation class. However, in fMAPLR a single class is used. Thus, it is straightforward to derive the set of transformation matrices directly from SI model or the data. In our experiments, an efficient algorithm is developed to learn the prior distribution of the feature transformation matrices. Speech data from the disjoint devleopment set is extracted and one fMLLR feature transform matrix is estimated for each of these speakers. Then equations 13 and 14 are used to estimate the prior distribution used in fMAPLR in future tests.

## 5. EXPERIMENTS

Experiments are conducted using two internal telephony speech recognition databases, D1 and D2 respectively. Both test sets are US English telephony databases over diverse topics such as digits, letters, names and dates and covering landline, speakerphone and cell phone conditions. In D1, the development set used for prior distribution estimation contains about 300 speakers, each of them provided about 45 seconds of speech data. The corresponding test set contains about 100 speakers. For simpicity, diagonal scale factor matrices, $\mathcal{V}_i$, were used in the prior distribution. In order to further verify the robustness of the proposed approach, we used the prior distributions derived from D1 to test fMAPLR on D2 which contains data in a very noisy environment. Test set D2 contains 82 dialogs, with around 19 utterances on average for each dialog. Each utterance typically contains around 1 second of speech data. A speaker independent, triphone-based model is used. Senone clustering is performed based on a phonetic decision tree and the model has about 50K Gaussians.

We experimented with two different methods for organizing the adaptation data and applying the adapted transforms: batch mode and incremental mode. In batch adaptation mode, the first $T$ frames of speech data are used to derive the feature transformation matrix. The estimated feature transform is then applied to the remainder of the data in the dialog. In incremental adaptation mode, as in batch mode, the adaptation transform is estimated after $T$ frames of data. However, in incremental mode, we continuously update and apply the adaptation transform as new data is seen. In our incremental mode experiments, we updated and applied the new transform after each utterance in the dialog.

### 5.1. Batch mode experiments

Batch mode results for fMLLR and fMAPLR adaptation experiments on D1 and D2 are shown in Table 1 and Table 2 respectively. In Table 1, it is observed that for fMLLR, when only 100 frames of adaptation data are available, poor feature transforms are estimated causing a dramatic degradation. Even when the adaptation data amount increases to 300 frames per speaker, the fMLLR adaptation still did not perform as well as the baseline without adaptation for D1. On the other hand, fMAPLR provides a much more stable result even at the extreme of only 100 frames of adaptation data. fMAPLR is able to achieve a 3.2% relative gain after estimating transforms from only 300 frames of data.

To test the robustness of fMAPLR and the generalization of the prior distribution estimation, we applied the prior distribution estimated on D1 to the fMAPLR estimation for the test set of D2. These results are shown in Table 2. The re-

**Table 1**. *WER on D1 test set in batch mode.*

| | $T = 100$ | $T = 300$ |
|---|---|---|
| No adaptation | 11.33 | 11.33 |
| fMLLR | 21.55 | 11.84 |
| fMAPLR | 11.36 | 10.97 |

sults show that fMAPLR is more robust than fMLLR even with the hyperparameters learned from from a database with a mismatched environment. The reader should also notice that, a larger $T$ will not necessarily lead to a lower overall WER since that reduces the amount of data to be decoded using speaker adapted system. This effect is more significant if the dialog provided by each speaker is very short. Of course, the system could be modified to re-decode the input after adaptation to regain some of this loss.

**Table 2**. *WER on D2 test set in batch mode.*

| | $T = 100$ | $T = 300$ |
|---|---|---|
| No adaptation | 14.27 | 14.27 |
| fMLLR | 18.09 | 13.52 |
| fMAPLR | 13.14 | 13.61 |

### 5.2. Incremental mode experiments

Incremental mode results for fMLLR and fMAPLR on D1 and D2 are shown in Table 3 and Table 4 respectively. As in batch mode, Table 3 shows that fMAPLR incremental adaptation performs better than the fMLLR adaptation. Unlike batch mode, fMLLR now gives some improvement over the baseline at $T = 100$ case because the poor transform estimated from the first 100 frames of data has been corrected when more data are decoded and used for transform estimation.

**Table 3**. *WER on D1 test set in incremental mode.*

| | $T = 50$ | $T = 100$ | $T = 300$ |
|---|---|---|---|
| No adaptation | 11.33 | 11.33 | 11.33 |
| fMLLR | 12.97 | 10.86 | 10.51 |
| fMAPLR | 10.48 | 10.45 | 10.45 |

Table 4 shows incremental adaptation at different time delays. In this experiment $T = 0$ means the incremental adaptation is started as soon as the first utterance is decoded. The prior distribution estimated on D1 has generalized very well to D2. Moreover, fMAPLR performs consistently better than fMLLR, and it shows the best performance by conducting fMAPLR as soon as the first utterance is decoded and available for adaptation, this is critical for the applications such as name dialing, where the whole dialog contains only a few dialog turns.

**Table 4**. *WER on D2 test set in incremental mode.*

| | $T = 0$ | $T = 100$ | $T = 300$ |
|---|---|---|---|
| No adaptation | 14.27 | 14.27 | 14.27 |
| fMLLR | 15.83 | 13.10 | 13.05 |
| fMAPLR | 12.72 | 12.81 | 12.95 |

## 6. CONCLUSION

Feature space adaptation such as fMLLR is preferred to model space adaptation in telephony applications. However, with very sparse adaptation data, the fMLLR estimated feature transforms could be unstable. In this paper, we apply MAP framework to feature space transformation and derive the fMAPLR as a counterpart of the model space MAPLR. Experimental results show that fMAPLR performs more robustly than fMLLR when the adaptation data is very limited.

## Acknowledgments

The authors would like to thank Patrick Nguyen at Microsoft Research for providing valuable technical advice. Thanks also to Jian Wu and Yifan Gong in the Microsoft Speech and Natural Language Group for many useful discussions during this work.

## 7. REFERENCES

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[2] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[3] W. Chou, "Maximum a posteriori linear regression with elliptically symmetric matrix variate priors," in *EUROSPEECH*, 1999.

[4] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden markov model adaptation," in *EUROSPEECH*, 1999.

[5] W. Chou and X. He, "Maximum a posteriori linear regression based variance adaptation of continuous density HMMs," in *EUROSPEECH*, 2003.

[6] Y. Li, H. Erdogan, Y. Gao, and E. Marcheret, "Incremental online feature space MLLR adaptation for telephony speech recognition," in *ICSLP*, 2002.

[7] W. Byrne and A. Gunawardana, "Discounted likelihood linear regression for rapid adaptation," in *EUROSPEECH*, 1999.

[8] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., Cambridge University Engineering Department, May 1997.

[9] S.V. Balakrishnan, "Fast incremental adaptation using maximum likelihood regression and stochastic gradient descent," in *EUROSPEECH*, 2003.