Geometric Constrained Maximum Likelihood Linear Regression On Mandarin Dialect Adaptation[1]

Huayun Zhang[*] *Bo Xu*[*+]

High-Tech Innovation Center[*], National Laboratory of Pattern Recognition[+]
Institute of Automation, Chinese Academy of Sciences, Beijing
{hyzhang, xubo}@hitic.ia.ac.cn

Abstract

This paper presents a geometric constrained transformation approach for fast acoustic adaptation, which improves the modeling resolution of the conventional Maximum Likelihood Linear Regression (MLLR). For this approach, the underlying geometry difference between the seed and the target spaces is exposed and quantified, and used as a prior knowledge to reconstruct refiner transforms. Ignoring dimensions that have minor affections to this difference, the transform could be constrained to a lower rank subspace. And only distortions within this subspace are to be refined in a cascaded process. Compared to previous cascade method, we employ a different parameterization and obtain a higher resolution. At the same time, since the geometric span for refiner transforms is highly controlled, it could be adapted quickly. So, it could achieve a better tradeoff between resolution and robustness. In Mandarin dialect adaptations, this approach provides 4~9% word-error-rate relative decrease over MLLR and 3~5% over previous cascade method correspondingly with varying amounts of data.

1. Introduction

It is well known that today's automatic speech recognition (ASR) systems are sensitive to variations between the training and testing conditions. Such mismatches caused by speaker/ channel/dialect variations inevitably occur in real applications. Previous studies prove that adapting acoustic model to target conditions can efficiently address this problem [1,2]. Acoustic adaptation is an essential part for state-of-the-art ASR systems.

Many adaptation techniques have been extensively studied. One of the most successful techniques is transform-based adaptation, such as MLLR [3]. Standard MLLR applies unconstrained linear transforms for Gaussian components, using Expectation-Maximization (EM) to find Maximum-Likelihood estimation of the transforms. To obtain a robust estimation, transforms are shared across Gaussians into classes.

With many transforms and enough data, MLLR could obtain speaker/environment dependent performance. But MLLR fairs poorly if we try to compute too many parameters from too little data [2]. Since sparse data problem exists in many applications, it appears necessary to introduce constraints on the possible values for these parameters to avoid unreasonable estimates that might perturb the underlying structure of the acoustic space. Typically, this problem is handled by limiting the transform number or by imposing some types of compact parameterization for the transform. Alternatively, some kind of prior statistic knowledge could be utilized to constrain the transform [5,8,9]. All these schemes obtain a fast adaptation with limited data, but also saturate quickly [2].

Additionally, MLLR assumes that all Gaussians in the same class are transformed identically and the transforms for different classes are estimated independently. Studies show that additional gains can be made if we do not constrain components of the same class to use identical transform. As in Maximum-Likelihood Stochastic Transformation (MLST) [6], multiple transform sets are obtained beforehand from large amounts of training data; only weights are to be estimated for adaptation. However, the pre-determined multiple transform sets are unreliable or even unavailable in some cases. And resolution is also limited by the pre-determined transform sets. Another cascade transform is developed to improve resolution gradually in refiner classes [10]. However, not all parameters could be refined in the cascaded process due to limited data for refiner classes (only the bias is adapted in refinement of [10]), which prevent its resolution from further improved.

In this paper, we propose a geometric constrained approach for improving the resolution of conventional MLLR. The MLLR transform is analyzed and approximated upon its geometric span. A new simplified parameterization is employed for the refinement. Geometric Constrained MLLR (GC-MLLR) achieves a better trade-off between robustness and resolution. In experiments on Mandarin dialect adaptation, GC-MLLR was proved superior to MLLR, especially when the available data is limited. In the next part, we describe the details of GC-MLLR. In the 3rd part, we present our experiments on adapting a mainland Mandarin acoustic model to be used for Taiwanese telephony speech recognition. In the 4th part, we will give some discussions on practical issues for GC-MLLR. Finally we will give our conclusion in the 5th part.

2. Geometric Constrained MLLR

Standard MLLR involves computing transforms to update both the mean vectors and the variance matrices of Gaussians.

$$\begin{aligned}\hat{\mathbf{m}}_g &= \mathbf{A}_c * \mathbf{m}_g + \mathbf{B}_c \\ \hat{\mathbf{\Sigma}}_g^{-1} &= \mathbf{A}_c \mathbf{\Sigma}_g^{-1} \mathbf{A}_c^T\end{aligned} \quad g = 1,\cdots,G_C; c = 1,\cdots,C \tag{1}$$

All Gaussians are clustered into C classes and Gaussians belonging to class c, $N(\mathbf{m}_g, \mathbf{\Sigma}_g)$ $g = 1,2,\cdots,G_c$, are adapted with the same transform $(\mathbf{A}_c, \mathbf{B}_c)$. Here $\mathbf{A}_c$ is a $D \times D$ rotation matrix, $\mathbf{B}_c$ is a D-dimensional bias vector, D is dimensionality of observations. Since means adaptation has the most contribution to performance improvement [7], only mean vector transforms is considered in this study.

[1] This work was supported by the National Key Fundamental Research Program of China under grant 19980300504.

The regression classes are determined according to the amount and the context of the data available using a regression tree [4]. Given a task, there are two practical decisions: (A) Depth of the tree. The deeper tree, the more classes, and the higher resolution achieved, but the larger amount of data required. (B) Parameterization. $\mathbf{A}_c$ could be considered as identity, diagonal, block-diagonal, or full matrix. Obviously, full matrices model inter-dimensional correlation among the means more precisely and provide superior description for acoustic variations [2, 7]. However, it requires more data for estimation. In many cases, resolution has to be sacrificed in exchange for robust estimates.

2.1. Geometric Analysis of Linear Transform

Since linear transform has clear geometric interpretation in the hyperspace, we are motivated to analyze underlying distortion of the acoustic space in detail to find a more compact and more effective parameterization. Linear transform parameters consist of a $D \times D$ matrix $\mathbf{A}$ and a vector $\mathbf{B} = [b_1; b_2; \cdots; b_D]$.

It is well known that singular value decomposition (SVD) is a powerful tool to expose the geometric structure of the square matrix $\mathbf{A}$. SVD of matrix $\mathbf{A}$ is any factorization of the form:

$$\mathbf{A} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^T \quad (2)$$

$\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. Denote them in columns:

$$\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_D], \mathbf{U} \times \mathbf{U}^T = \mathbf{I}_D; \quad \mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_D], \mathbf{V} \times \mathbf{V}^T = \mathbf{I}_D \quad (3)$$

$\mathbf{S}$ is a $D \times D$ matrix with $s_{ij} = 0 \; if \; i \neq j \; and \; s_{ii} = s_i \geq 0$. Its diagonal elements are decreasingly ordered singular values of $\mathbf{A}$. We can express $\mathbf{A}$ by the sum of D matrices of rank 1.

$$\mathbf{A} = s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + s_D \mathbf{u}_D \mathbf{v}_D^T \quad (4)$$

The geometric interpretation of the transform can be viewed more clearly in terms of the singular values of $\mathbf{A}$. If we assume a set of vectors $\mathbf{x}$ for $\|\mathbf{x}\| = 1$ in a D-dimensional space, which defines a hyper unit circle, then $\mathbf{y} = \mathbf{A} * \mathbf{x} + \mathbf{B}$ results a hyper ellipsoid geometrically. Let $D = 2$, this is depicted in Figure (1). The lengths of axes of the ellipsoid are the singular values $\mathbf{A}$. $\mathbf{B}$ is a bias between the seed and target vectors.

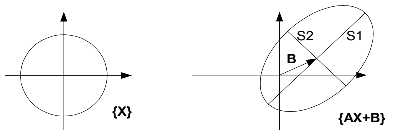

Figure 1:*Mapping by* $\mathbf{A}$ *of Unit Circle.*

The decomposition shows the action of $\mathbf{A}$ can be described as a rotation of $\mathbf{V}$ followed by a stretch of $\mathbf{S}$ followed by another rotation of $\mathbf{U}$. These singular values describe the extent to which multiplication by $\mathbf{A}$ distorts the original vector.

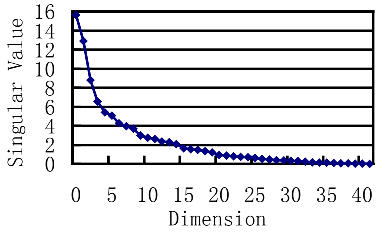

Figure 2:*The singular values of full MLLR Transform.*

Figure (2) shows the SVD on MLLR full matrix. The sharp decline of the singular values suggests that the full matrix $\mathbf{A}$ is nearly deficient. And SVD of $\mathbf{A}$ produces a sequence of approximations to $\mathbf{A}$ of successive ranks $\mathbf{A}_i = \mathbf{U} \times \mathbf{S}_i \times \mathbf{V}^T$, where $\mathbf{S}_i$ is the rank i version of $\mathbf{S}$ obtained by setting the last $D - i$ singular values to zero. Also, $\mathbf{A}_i$ is the best rank i approximation to $\mathbf{A}$ in the sense of Euclidean distance.

Intuitively, the magnitude of the singular values can be used to highlight which dimensions are most affected, and in some sense more important to the transform. The distortion caused by $\mathbf{A}$ can be decomposed to a major part plus a minor part as:

$$\mathbf{A} = \sum_{i=1}^{K} s_i \mathbf{u}_i \mathbf{v}_i^T + \sum_{j=K+1}^{D} s_j \mathbf{u}_j \mathbf{v}_j^T = \mathbf{U}\mathbf{S}_K \mathbf{V}^T + \mathbf{U}(\mathbf{S} - \mathbf{S}_K)\mathbf{V}^T \quad (5)$$

Hold the rotations dominated by $\mathbf{U}$ and $\mathbf{V}$, the transform could be approximately constrained into two lower rank subspaces corresponding to $\mathbf{S}_K$ and $\mathbf{S} - \mathbf{S}_K$. Stretches caused by $\mathbf{S}_K$ have more contributions to the distortion than $\mathbf{S} - \mathbf{S}_K$

2.2. Geometric Constrained MLLR

The geometric constrained MLLR is a gradually improving solution. It consists of three steps. In the first step, we compute the C MLLR full transforms according to the data. Secondly, SVD is performed on these transforms to explore the underlying geometric variations of the acoustic space. This information is employed as a prior knowledge to approximate the transform into lower rank subspaces. Then another EM process is carried on to find refiner transforms for subclasses at lower level along the regression tree. Since geometric structure of the transform is constrained, the number of parameters is effectively decreased for the refinement. This lead to a quicker adaptation, while the full transform advantage is maintained. Additionally, besides current data, no other prior assumptions are needed for GC-MLLR. So with more data, GC- MLLR could consistently improve resolution.

Denote the seed and the target mean as $\mathbf{m} = [m_1; m_2; \cdots; m_D]$ and $\hat{\mathbf{m}} = [\hat{m}_1; \hat{m}_2; \cdots; \hat{m}_D]$ respectively. Formulation of GC-MLLR is based on maximizing the likelihood of observations, in which the variations in subspace of $\mathbf{S} - \mathbf{S}_K$ and all rotations ($\mathbf{U}$ and $\mathbf{V}$) are fixed, only those stretchers in subspace of $\mathbf{S}_K$ and vector $\mathbf{B}$ are considered to be adaptable. The target is:

$$p(\mathbf{O}, \mathbf{W} | \mathbf{B}, \mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{m}, \mathbf{\Sigma}) \quad (6)$$

$\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_T\}$ is the D-dimensional adaptation data, and $\mathbf{W} = \{w_1, \cdots, w_T\}$ is the set of possible underlying state sequences.

The Expectation-Maximization (EM) algorithm could address this ML estimation. The EM auxiliary function is defined as:

$$\begin{aligned} Q(\mathbf{B}', \mathbf{S}' | \mathbf{B}, \mathbf{S}) &= E\{\log p(\mathbf{O}, \mathbf{W} | \mathbf{B}', \mathbf{S}', \mathbf{U}, \mathbf{V}, \mathbf{m}, \mathbf{\Sigma}) | \mathbf{B}, \mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{m}, \mathbf{\Sigma}\} \quad (7) \\ &= \sum_{t=1}^{T} E\{\log p(\mathbf{o}_t, w_t | \mathbf{B}', \mathbf{S}', \mathbf{U}, \mathbf{V}, \mathbf{m}, \mathbf{\Sigma}) | \mathbf{B}, \mathbf{S}, \mathbf{U}, \mathbf{V}, \mathbf{m}, \mathbf{\Sigma}\} \\ &= \sum_{t=1}^{T} \sum_{g=1}^{G} \gamma(t, g) \cdot \log\{p((\mathbf{o}_t, g) | \mathbf{B}', \mathbf{S}')\} \quad where \; \mathbf{S}' = \mathbf{S}'_K + (\mathbf{S} - \mathbf{S}_K) \end{aligned}$$

G is the number of possible Gaussians at time t, $\gamma(t,g)$ is the occupancy probability for Gaussian g at time t. And

$$
\begin{aligned}
&p((\mathbf{o}_t, g)|\mathbf{B}', \mathbf{S}') \\
&= \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_g|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{o}_t - \hat{\mathbf{m}}_g)^T \boldsymbol{\Sigma}_g^{-1}(\mathbf{o}_t - \hat{\mathbf{m}}_g)\right) \\
&\textit{where } \hat{\mathbf{m}}_g = \mathbf{U} \times \mathbf{S}' \times \mathbf{V} \times \mathbf{m}_g + \mathbf{B}'
\end{aligned}
\tag{8}
$$

If $\boldsymbol{\Sigma}_g$ is simplified as a diagonal matrix, and the non-zero diagonal elements could be denoted as $\{\boldsymbol{\sigma}_g^2(d)\}_{d=1,\cdots,D}$, then the auxiliary function could be simplified as follows:

$$
\begin{aligned}
Q(\mathbf{B}', \mathbf{S}'|\mathbf{B}, \mathbf{S}) &= \sum_{t=1}^{T}\sum_{g=1}^{G}\gamma(t,g)\sum_{d=1}^{D}\left\{K(g,d) - \frac{[\mathbf{o}_t(d) - \hat{\mathbf{m}}_g(d)]^2}{2\boldsymbol{\sigma}_g^2(d)}\right\} \\
&= \sum_{t=1}^{T}\sum_{g=1}^{G}\gamma(t,g)\sum_{d=1}^{D}\left\{K(g,d) - \frac{\left[\mathbf{o}_t(d) - \sum_{i=1}^{D}[s_i'\mathbf{u}_i(d)\mathbf{v}_i^T\mathbf{m}_g] - \mathbf{B}'(d)\right]^2}{2\boldsymbol{\sigma}_g^2(d)}\right\} \\
&\textit{where} \quad K(g,d) = \log\left(1/\sqrt{2\pi}\boldsymbol{\sigma}_g(d)\right)
\end{aligned}
\tag{9}
$$

In order to find the maximum value of Q, we differentiate it with respect to $\mathbf{B}'$ and $\mathbf{S}'_K$, then solve for zeros respectively.

$$
\begin{cases}
\partial Q/\partial s_k' = 0 \\
\partial Q/\partial \mathbf{B}'(j) = 0
\end{cases}
\Rightarrow
\tag{10}
$$

$$
\sum_{t=1}^{T}\sum_{g=1}^{G}\gamma(t,g)\sum_{d=1}^{D}\left\{\frac{\left[\mathbf{o}_t(d) - \sum_{i=1}^{D}[s_i'\mathbf{u}_i(d)\mathbf{v}_i^T\mathbf{m}_g] - \mathbf{B}'(d)\right]}{\boldsymbol{\sigma}_g^2(d)}\mathbf{u}_k(d)\mathbf{v}_k^T\mathbf{m}_g\right\} = 0
$$

$$
\sum_{t=1}^{T}\sum_{g=1}^{G}\gamma(t,g)\left\{\frac{\left[\mathbf{o}_t(j) - \sum_{i=1}^{D}[s_i'\mathbf{u}_i(j)\mathbf{v}_i^T\mathbf{m}_g] - \mathbf{B}'(j)\right]}{\boldsymbol{\sigma}_g^2(j)}\right\} = 0 \quad \textit{for} \quad \begin{matrix} k = 1,\cdots K \\ j = 1,\cdots,D \end{matrix}
$$

The $D+K$ parameters are obtained directly by solving these $D+K$ linear equations. And the full matrix is reconstructed by:

$$
\mathbf{A}' = \mathbf{U} \times \mathbf{S}' \times \mathbf{V}^T = \mathbf{U} \times (\mathbf{S}'_K + (\mathbf{S} - \mathbf{S}_K)) \times \mathbf{V}^T
\tag{11}
$$

GC-MLLR mathematically guarantees to be more precise than MLLR. Figure (3) is the relation between likelihood of GC-MLLR and full MLLR on same data. It shows that GC-MLLR results a higher likelihood, especially when data set is small.

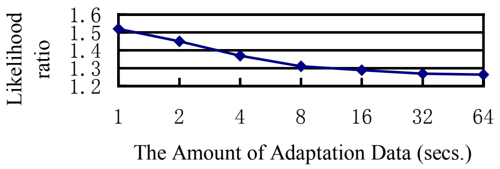

Figure 3:*The likelihood ratio of GC-MLLR to Full-MLLR.*

2.3. Comparison of GC-MLLR and Cascade Transform

To retain the modeling capability of full transform, and at the same time model dependencies among classes, cascade transform has been proposed [10], in which full $\mathbf{A}_c$ and $\mathbf{B}_c$ are estimated for rough classes, then an additional EM process is cascaded to re-estimate bias vector $\mathbf{B}'_c$ for each refiner class.

Obviously, GC-MLLR could be viewed as a cascade scheme, in which there are multiple transform sets within a MLLR class. Gaussians of the same MLLR class are not transformed identically. On the other hand, each GC-MLLR regression class receives its own transform without having to estimate each one independently but jointly with adjacent classes.

However, unlike previous cascade transform, the geometric information is utilized here to simplify the parameterization type for the cascade refinement process. Only the dominant distortions of the geometric span are to be refined in GC-MLLR and the remaining structure is maintained, which guarantees to obtain robust estimates upon refiner classes.

For comparative reason, we perform two types of cascade transform. One is the same as in [10], which only include a bias refinement in the cascaded process; the other is extended to include an additional diagonal matrix in the refinement.

$$
\begin{aligned}
&\textit{Cascade} - I: \quad \hat{\mathbf{m}} = \mathbf{A} * \mathbf{m} + \mathbf{B}' \\
&\textit{Cascade} - II: \quad \hat{\mathbf{m}} = \mathbf{A}' * \mathbf{A} * \mathbf{m} + \mathbf{B}'
\end{aligned}
\tag{12}
$$

Here $\mathbf{A}'$ is a diagonal matrix. There are D and $2 \times D$ free parameters for Cascade-I and Cascade-II respectively to be improved in the cascaded process.

These adaptation methods could be visualized as Figure (4). For standard MLLR, the movements and the distortions of the components in the same class are all the same. For Cascade-I, the movement for each subclass is adapted independently, but geometric distortions maintain the same. For Cascade-II and GC-MLLR, locations, and stretchers along every rotated axe are adapted independently for subclasses, only the rotation directions are maintained. However, since $K < D$, GC-MLLR could obtain a faster adaptation than Casscade-II.

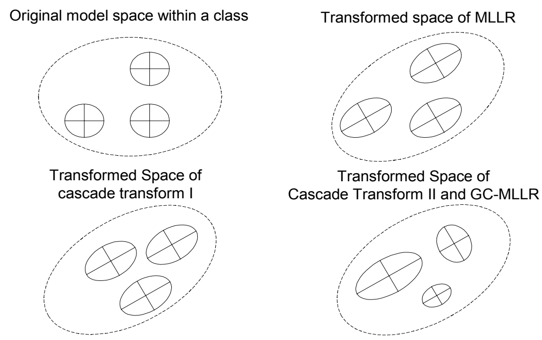

Figure 4:*Transformed acoustic space of different methods.*

3. Experiments

We investigate our proposal on Mandarin dialect adaptation, where our task is to develop a Taiwan-dialect speaker-independent speech recognition system with only limited amount of Taiwan- dialect Mandarin speech corpus available.

Experiments are carried out on a Hidden Marko Model (HMM) based large-vocabulary continuous speech recognizer. The system contains 3463 context-dependent continuous-density phonetic models that share 1395 output distributions, with 16 mixture Gaussian components per distribution. The front-end is configured to extract energy, pitch, 12 MFCC with their delta and delta-delta features.

Our seed model was trained by 863 corpus, which contains more than 70 hours Mandarin speech collected across the mainland of China. The word recognition rate on mainland-dialect Mandarin achieves 95% with vocabulary size of 1000.

The adaptation data consists of 1600 continuous sentences read by 160 speakers from Taiwan, both male and female, and each of them recorded 10 sentences. Each sentence contains 5~15 syllables. The testing set consists of 1200 stock name utterances read by another 12 Taiwanese speakers, both male and female, and one person recorded 10 utterances. Each utterance contains 2~4 syllables. All speech files are recorded through telephone network and 8000Hz sampled.

We compare GC-MLLR with four other adaptation schemes, the diagonal MLLR, the full MLLR and the two cascade transform defined in Equation (5). Different sized data sets are used in experiments. We select the first 1, 3, 5, 10 sentences of each speaker to compose data set of 160, 480, 900, 1600 sentences respectively. For all these five schemes, the number of transforms is optimized and only the best result is shown here. All experiments are conducted in a supervised manner. And the Word Error Rate (WER) statistics are shown in Table. (1). And as a baseline, the WER without adaptation is 29%.

Table 1: *Experiment results using different adaptation scheme.*

DataSets / Methods	160	480	900	1600
Diag-MLLR	24.8	23.7	23.6	23.2
Full-MLLR	25.8	23.4	21.8	20.1
Cascade-I	24.5	22.8	21.6	20.1
Cascade-II	24.6	22.0	20.4	18.9
GC-MLLR	23.1	21.2	20.1	19.0

4. Discussion

We could see that the improvement of Diag-MLLR saturate very quickly and GC-MLLR achieves 9.3%, 7.6%, 5.9%, 3.8% WER relative reduction compared to Full-MLLR with 160, 480, 900, 1600 adaptation sentences respectively. Compared to conventional Cascade-I method, GC-MLLR provides additional 4.8%, 5.5%, 5.2%, 3.8% WER relative reduction with different size of data. With comparatively small data size, GC-MLLR shows to be superiors to Cascade-II. With more data available, these two methods show almost equal performance. Generally, GC-MLLR is the best one with varying size of data set, especially when the data size is small. Moreover, as adaptation sentences increased, GC-MLLR constantly gives slightly better performance over Cascade-II.

The above experiment results with varying amounts of data indicate that our new approach to simplify parameterization for refinement performs better than other simplifications used in previous cascade method. Of cause, constraining the transform to a lower rank K subspace would introduce some limitation for the refinement. But if we control K value, this could be effectively compensated in refinement.

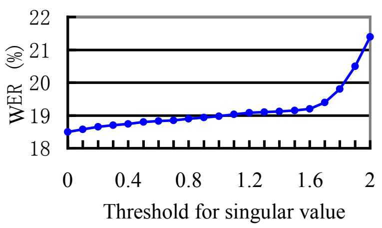

Figure 5:*Adaptation performance with varying* ε value.

The K value is determined empirically. We set a threshold ε for the singular values to determine the rank K subspace. In some sense, ε reflects the precision limitation of our parameterization simplification in the refinement process. Figure (5) shows the experimental results with varying ε. When $\varepsilon \leq 1.5$, this simplification has no obviously affection on performance. It results about $K = 0.5D$ in our system.

It should be noticed that GC-MLLR is suitable for fast adaptations. Bur if more data available, it could also obtain a performance parallel to full MLLR and Cascade-II. Compared with other fast schemes for acoustic adaptation, no additional assumptions outside current data are embedded in GC-MLLR. So its potential capability with more data and highly-mismatch adaptation is unlimited.

5. Conclusion

We proposed a geometric-constrained linear-transformation based fast adaptation algorithm. For this method, we employ a new compact transform parameterization based on a prior geometric structure analysis of the conventional full MLLR matrix. And an iterative resolution is given. Compared to previous cascade transform, this method could reconstruct the transform for refiner classe more quickly and more effectively. And the adaptation resolution is not lost. At the same time, as previous cascade transform, no model selection and no additional prior-knowledge of the transforms are explicitly used in the computation. So its potential capability for highly mismatched adaptation and more data adaptation is unlimited. Our primary experiments on Mandarin dialect adaptation show that new method significantly outperforms conventional MLLR transform and previous cascade transform with varying amounts of adaptation data, especially when data is limited.

6. References

[1] J.L. Gauvain, L. Lamel, *"Large-Vocabulary Continous Speech Recognition: Advances and Applications"*, Proceedings of the IEEE, Vol. 88, No. 8, Aug 2000.

[2] P.C. Woodland, *"Speaker Adaptation: Techniques and Challenges"*, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 1999.

[3] C. J. Leggetter and P. C. Woodland, "*Speaker adaptation of continuous density HMMs using multivariate linear regression*, " in Proc. of ICSLP 94, Japan, Sep.1994,

[4] M. Gales, "*The generation and use of regression class trees for MLLR adaptation"*, Technical Report. TR263, Cambridge University Engineering Department, 1996.

[5] O. Siohan, C. Chesta, and C.-H. Lee. *"Hidden Markov model adaptation using maximum a posteriori linear regression"*. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, 1999.

[6] V. Diakoloukas and V. Digalakis, *"Maximum-likelihood stochastic-transformation adaptation of hidden Markov Models"*, IEEE Trans. on SAP, Vol. 7, No. 2, Mar 1999.

[7] L.Neumeyer et al, *"A Comparative Study of Speaker Adaptation Techniques"*, Proc. Eurospeech'1995.

[8] M. Gales, *"Cluster Adaptive Training for Speech Recognition"*, Proc. ICSLP'1998.

[9] R. Kuhn et al., *"Eigenvoices for speaker adaptation"*, Proc. ICSLP'1998.

[10] *http://www.clsp.jhu.edu/ws98/projects/adapt*