Robust Speaker Adaptation by Weighted Model Averaging Based on the Minimum Description Length Criterion

Xiaodong Cui, *Member, IEEE*, and Abeer Alwan, *Senior Member, IEEE*

***Abstract*—The maximum likelihood linear regression (MLLR) technique is widely used in speaker adaptation due to its effectiveness and computational advantages. When the adaptation data are sparse, MLLR performance degrades because of unreliable parameter estimation. In this paper, a robust MLLR speaker adaptation approach via weighted model averaging is investigated. A variety of transformation structures is first chosen and a general form of maximum likelihood (ML) estimation of the structures is given. The minimum description length (MDL) principle is applied to account for the compromise between transformation granularity and descriptive ability regarding the tying patterns of structured transformations with a regression tree. Weighted model averaging across the candidate structures is then performed based on the normalized MDL scores. Experimental results show that this kind of model averaging in combination with regression tree tying gives robust and consistent performance across various amounts of adaptation data.**

***Index Terms*—Maximum likelihood linear regression (MLLR), minimum description length (MDL), model averaging, speaker adaptation.**

I. Introduction

SPEAKER adaptation is a crucial technique for speech recognition systems which modifies the original acoustic models towards a specific speaker given the speaker's acoustic characteristics. It can yield significant improvements over "unadapted" recognizers and therefore plays an important role in real-world applications. In the past two decades, speaker adaptation has become one of the most active research areas in the speech recognition field with many important contributions. Generally speaking, speaker-adaptation techniques fall into two categories: transformation-based approaches and model-based approaches. Transformation-based approaches relate the original and adapted model parameters by either a linear [1] or a nonlinear [2] transformation. The model-based approaches adapt model parameters directly without an assumption of transformation [3]–[5].

Manuscript received May 24, 2005; revised January 18, 2006. This work was supported in part by the National Science Foundation under Grant 0326214. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Geoffery Zweig.

X. Cui was with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA. He is now with the the IBM T. J. Watson Research Center, Yorktown Heights, NY 10523 USA. (e-mail: xdcui@icsl.ucla.edu).

A. Alwan is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: alwan@ee.ucla.edu).

Digital Object Identifier 10.1109/TASL.2006.876773

If a large amount of adaptation data is available, both transformation-based and model-based approaches yield satisfactory performance. In real-world applications, however, situations often occur when only a limited amount of data is available for adaptation. This may be due to difficulty in collecting the data or a requirement of rapid speaker adaptation. In this situation, data sparseness will affect performance. To deal with this issue, a variety of methods has been proposed for both approaches. For the transformation-based approach, a regression class tree is adopted in [1] to dynamically tie the transformation parameters, while dependencies between acoustic units are studied in [6] and [7] to make effective usage of the data. In the model-based approach, a structural maximum *a posteriori* (MAP) adaptation algorithm is proposed in [4] and [5] utilizing hierarchical priors resulting in good performance. These techniques can yield reliable adaptation for observed or unobserved acoustic units by smoothing the adaptation parameters across the sparse data.

Another interesting way to address the sparse-data problem is the eigenvoice method investigated in [8], where the acoustic models are obtained via a linear combination of representative speaker independent models in the eigenvoice (principal components) space. In case of adaptation, only the linear combination coefficients need to be estimated, which makes it a good choice for rapid speaker adaptation. The eigenvoice method has been extensively studied in the past few years and encouraging performance has been reported (e.g., [9]–[11]).

Among speaker adaptation techniques, the maximum likelihood linear regression (MLLR) [1] is one of the most well-known and widely used approaches due to its effectiveness and computational advantages. In this paper, we investigate a robust speaker adaptation scheme using MLLR with a structured transformation matrix. The scheme yields consistent performance across various amounts of adaptation data-sparse or adequate. Structured MLLR transformations are clustered through a regression tree [1] and their ML estimation is provided. Given a certain amount of adaptation data, a variety of transformation structures is chosen and their tying patterns with the regression tree are described by the minimum description length (MDL) [12] to account for the tradeoff between transformation granularity and descriptive ability. Based on the normalized MDL scores, the final transformation is obtained by a weighted average across the candidate structures.

Previous work on applications of the MDL principle in acoustic model selection involves the use of MDL to automatically determine the regression tree depth, applied to mean shifting [13] and structural MAP adaptation [14]. In this paper,

we investigate the use of the MDL principle to determine the optimal transformation structure from a set of predefined structures.

The remainder of the paper is organized as follows. In Section II, the formulation of estimation of structured MLLR transformations is provided. Structure description via the MDL principle and weighted model averaging based on normalized MDL scores are given in Sections III and IV. Section V discusses the choice of proper structures. Experimental results are presented in Section VI and are followed by a discussion in Section VII. A summary is presented in Section VIII.

II. Structured Transformations

As in [1], the MLLR transformation can be written as

$$\hat{\mu} = A\,\xi \tag{1}$$

where $\xi = [1, \mu_1, \cdots, \mu_N]^T$ is the augmented mean vector with $\mu = [\mu_1, \cdots, \mu_N]^T$ denoting the N-dimensional mean vector of a Gaussian mixture in speaker-independent acoustic models. The adapted Gaussian mixture mean $\hat{\mu}$ is computed from the original augmented mean ξ via a linear transformation matrix A with an $N \times (N+1)$ dimension.

When the adaptation data are adequate to perform reliable estimation, a full matrix form of A is preferred. However, most often in practical situations, only limited adaptation data are available. Under this condition, it is interesting to investigate different structures of A which may render fewer free parameters to estimate while still providing a good descriptive ability of the transformation. For a particular structure, only the elements of interest in the transformation matrix are taken into account while the rest are set to zeros. For instance, (2) illustrates a structure of the transformation matrix A with elements of interest located in the first column and along the three principal diagonals in the remaining sub-matrix

$$\mathbf{A} = \begin{bmatrix} \times & \times & \times & & & \\ \times & \times & \times & \times & & \\ \times & & \times & \times & \ddots & \\ \vdots & & & \ddots & \ddots & \times \\ \times & & & & \times & \times \end{bmatrix}_{N\times(N+1)}. \tag{2}$$

Before we derive the ML estimate of the structured transformation, let us first review the derivation of transformation matrix A with no assumption of its structure (see for example [1]). It will become manifest that the ML estimate of the structured transformation is an extension of the estimate of a full transformation matrix in an EM framework.

Suppose there are R adaptation utterances and U^r is the number of frames in the rth utterance. $\gamma_t^r(i,k) = p(s_t^r = i, \kappa_t^r = k|\mathcal{O}^r, \overline{\lambda})$ is the posterior probability of being at state i and Gaussian mixture k at time t given the rth observation sequence $\mathcal{O}^r = \{o_1^r, \cdots, o_{U^r}^r\}$. ξ_{ik} and Σ_{ik} are the augmented mean vector of μ_{ik} and covariance matrix associated with state i and Gaussian mixture k. $\overline{\lambda}$ are the parameters of previous models in the EM iterations. Transformations are tied into Q classes: $\{\omega_1, \cdots, \omega_q, \cdots, \omega_Q\}$. For a specific class ω_q, the transformation matrix A_q is shared across all the Gaussian mixtures $\mathcal{N}(\mathbf{o}_t^r; \mu_{ik}, \Sigma_{ik})$ with $(i,k) \in \omega_q$. The ML esitmation of A_q can be obtained from

$$\sum_{r=1}^{R}\sum_{t=1}^{U^r}\sum_{(i,k)\in\omega_q} \gamma_t^r(i,k)\Sigma_{ik}^{-1}\mathbf{o}_t^r\xi_{ik}^T = \sum_{r=1}^{R}\sum_{t=1}^{U^r}\sum_{(i,k)\in\omega_q} \gamma_t^r(i,k)\Sigma_{ik}^{-1}A_q\xi_{ik}\xi_{ik}^T. \tag{3}$$

Define the terms the same way as in [1]

$$V_{ik}^r = \sum_{t=1}^{U^r}\gamma_t^r(i,k)\Sigma_{ik}^{-1} \tag{4}$$

$$D_{ik} = \xi_{ik}\xi_{ik}^T \tag{5}$$

$$Z_q = \sum_{r=1}^{R}\sum_{t=1}^{U^r}\sum_{(i,k)\in\omega_q} \gamma_t^r(i,k)\Sigma_{ik}^{-1}\mathbf{o}_t^r\xi_{ik}^T. \tag{6}$$

Hence

$$\mathrm{vec}(Z_q) = \left(\sum_{r=1}^{R}\sum_{(i,k)\in\omega_q}(V_{ik}^r \otimes D_{ik})\right)\cdot \mathrm{vec}(A_q) \tag{7}$$

where $\mathrm{vec}(\cdot)$ converts a matrix into a vector in terms of the rows and $\otimes$ is the Kronecker product.

When the covariance matrix Σ_{ik} is diagonal, A_q could be computed row by row from the following linear relationship:

$$z_{qm}^T = \mathbf{G}_{qm}\cdot a_{qm}^T \tag{8}$$

where z_{qm} and a_{qm} are the mth row of Z_q and A_q, and

$$\mathbf{G}_{qm} = \sum_{r=1}^{R}\sum_{(i,k)\in\omega_q} v_{ik(mm)}^r D_{ik} \tag{9}$$

where $v_{ik(mm)}^r$ is the mth element on the diagonal of matrix V_{ik}^r.

In this paper, we are interested in the structure of A and ways of exploiting the structure in robust speaker adaptation. For a **structured transformation**, suppose the mth row of A_q has P_m elements of interest, namely

$$a_{qm} = \left[0, \cdots, 0, a_{qm,l_1}, 0, \cdots, 0, a_{qm,l_{P_m}}, 0, \cdots, 0\right]. \tag{10}$$

Define

$$\widetilde{a}_{qm} = \left[a_{qm,l_1}, a_{qm,l_2}, \cdots, a_{qm,l_{P_m}}\right]$$

and

$$\widetilde{z}_{qm} = \left[z_{qm,l_1}, z_{qm,l_2}, \cdots, z_{qm,l_{P_m}}\right]$$

as being the subvectors consisting of only those elements of interest. Then, $\widetilde{a}_{qm}$ can be solved using the following relationship:

$$\widetilde{z}_{qm}^T = \widetilde{\mathbf{G}}_{qm} \cdot \widetilde{a}_{qm}^T \tag{11}$$

where

$$\widetilde{\mathbf{G}}_{qm} = \begin{bmatrix} g_{l_1 l_1}^{(qm)} & g_{l_1 l_2}^{(qm)} & \cdots & g_{l_1 l_{P_m}}^{(qm)} \\ g_{l_2 l_1}^{(qm)} & g_{l_2 l_2}^{(qm)} & \cdots & g_{l_2 l_{P_m}}^{(qm)} \\ \vdots & \vdots & \ddots & \vdots \\ g_{l_{P_m} l_1}^{(qm)} & g_{l_P l_2}^{(qm)} & \cdots & g_{l_{P_m} l_{P_m}}^{(qm)} \end{bmatrix}. \tag{12}$$

In other words, the matrix $\widetilde{\mathbf{G}}_{qm}$ is generated by eliminating the rows and columns of $\mathbf{G}_{qm}$ which correspond to the zero elements in the structure, and keeping those of interest. The ML estimation of the structured transformation obtained in (11) is a general form for all possible structures.

III. Description of Structured Transformation Based on MDL

Since it was first proposed in 1978 [12], the MDL has been extensively studied and applied in model selection problems. There are many excellent papers reviewing MDL, e.g., [15]–[17], etc. In speech recognition, the MDL was also used as a means to cluster acoustic units or optimize acoustic models [18], [19]. Rooted in information theory, the MDL principle renders a view to model selection from a coding perspective. It treats a statistical model S with parameter θ as a coding algorithm to compress data X for the estimation. The total length $(L(S))$ to describe the coding of the data via the model includes the length of the compressed data $(-\log p(X|\theta))$ plus the length describing the model itself $(L(\theta))$

$$L(S) = -\log p(X|\theta) + L(\theta) \tag{13}$$

In (13), the first term on the right-hand side accounts for how well the model fits the data and the second term describes the complexity of the model. It is desirable to describe complicated phenomena by a simple model just as the famous Occam's razor states—"One should not increase, beyond what is necessary, the number of entities required to explain anything." Thus, given M competing models, the one with the shortest code length is favored which results in a simple model (or short $L(\theta)$) with a good fit of the data (or short $-\log p(X|\theta)$). In this paper, the MDL is employed to describe the structured MLLR adaptation using a regression tree.

Given an amount of adaptation data and a transformation structure, a regression class tree [1] is a good choice to obtain robust performance by dynamically tying Gaussian mixtures in the acoustic HMMs in terms of spatial similarity. The regression tree is created based on the centroid splitting algorithm using the Euclidean distance between the Gaussian mixture means as described in [20]. During adaptation, the Gaussian mixtures are pooled within their base class leaves or their parent nodes until the occupation counts are satisfactory for reliable estimation.

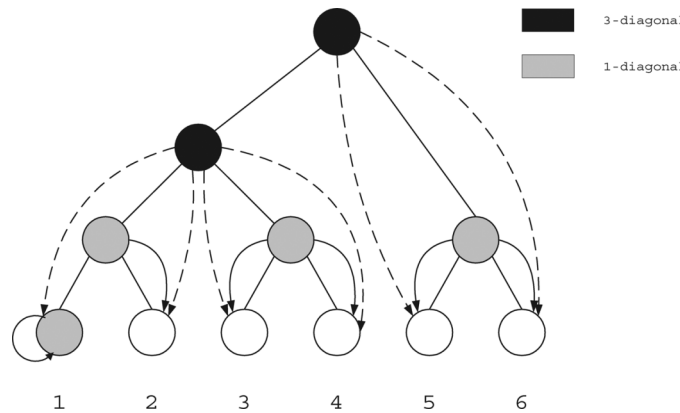

Fig. 1. Comparison of the transformation tying patterns with a regression tree of six base classes using 1-diagonal (gray node), and 3-diagonal (black node) structures.

While different transformation structures have different numbers of parameters, they provide different transformation descriptive ability and require different amounts of data to conduct reliable estimation. For illustration purposes, Fig. 1 compares the tying patterns of a 1-diagonal and 3-diagonal transformations with a six-class regression tree. The six base classes are denoted as the leaves at the bottom of the tree. In the figure, the tying of 1-diagonal structure is represented by grey nodes with solid arrows and 3-diagonal structure by black nodes with dashed arrows. For instance, in the 1-diagonal structure case, Gaussian mixtures from base class 1 share the transformation estimated from their own class while Gaussian mixtures from base class 2 are applied with the transformation estimated from both base classes 1 and 2. On the other hand, base classes 1, 2, 3, and 4 share the same transformation estimated from those classes in the 3-diagonal structure case. There are totally four transformations for the 1-diagonal structure and two transformations for the 3-diagonal structure.

From the figure, since the 1-diagonal structure has fewer parameters than the 3-diagonal case, transformations have been tied at a lower level in the tree which indicates a better granularity. On the other hand, the 3-diagonal structure has more parameters to describe the transformation; this indicates a better descriptive ability. Therefore, a tradeoff has to be made between transformation granularity and descriptive ability.

Suppose there are M competing structures $\{S_1, \cdots, S_M\}$ which result in different regression-tree tying schemes. Typically, complicated structures have transformations tied across more Gaussian mixtures (higher level in the tree toward the root node) and simple structures across less Gaussian mixtures (lower level in the tree toward the leaves). To explore the compromise between transformation granularity and descriptive ability for each transformation structure, the MDL principle is a good criterion.

In particular, suppose the Gaussian mixtures of the original acoustic hidden Markov models (HMMs) are clustered into L base classes with $D_l (l = 1, \cdots, L)$ mixtures in the lth class. For the dynamical tying of the structure $S_m (m = 1, \cdots, M)$

resulting in Q_m transformations with a regression tree over R adaptation utterances $\{\mathcal{O}^1, \mathcal{O}^2, \cdots, \mathcal{O}^R\}$, the description length is composed of three parts

$$L(S_m) = L_1(S_m) + L_2(S_m) + L_3(S_m) \tag{14}$$

where

$$L_1(S_m) = -\log p\left(\mathcal{O}^1, \mathcal{O}^2, \cdots, \mathcal{O}^R | A_1, \cdots, A_{Q_m}; \overline{\lambda}\right) \tag{15}$$

$$L_2(S_m) = \sum_{q=1}^{Q_m} \frac{|S_m|}{2} \log \Gamma_{mq} \tag{16}$$

$$L_3(S_m) = \sum_{l=1}^{L} D_l I_{ml}. \tag{17}$$

In (14), $L_1(S_m)$ is the code length of the compressed data $\{\mathcal{O}^1, \mathcal{O}^2, \cdots, \mathcal{O}^R\}$ using Q_m distinct transformations with structure S_m, $L_2(S_m)$ is the code length of the Q_m transformations, and $L_3(S_m)$ is the code length identifying one of the Q_m transformations for each Gaussian mixture. The three terms influence the compromise between transformation granularity and descriptive ability in different ways. $L_1(S_m)$ and $L_2(S_m)$ balance the likelihood and the number of transformation parameters by choosing transformation structures that have a higher likelihood with less parameters. In addition, $L_3(S_m)$ introduces penalty for tying patterns with more transformations since they have to employ a longer code to describe the application of the transformations to Gaussian mixtures. In the following, we will provide details on calculation of the three lengths.

Suppose the introduction of the transformation does not alter (a) the initial state probabilities, (b) the state transition probabilities, and (c) the frame/state alignment. Then, the first term in (14) $L_1(S_m)$ could be computed based on the forward–backward procedure [21] using transformed Gaussian mixtures by the transformations $\{A_1, \cdots, A_{Q_m}\}$

$$\begin{aligned}
&\log p(\mathcal{O}^1, \mathcal{O}^2, \cdots, \mathcal{O}^R | A_1, \cdots, A_{Q_m}; \overline{\lambda}) \\
&= \log \prod_{r=1}^{R} p(\mathcal{O}^r | A_1, \cdots, A_{Q_m}; \overline{\lambda}) \\
&= \sum_{r=1}^{R} \log p(\mathcal{O}^r | A_1, \cdots, A_{Q_m}; \overline{\lambda}) \\
&= \sum_{r=1}^{R} \log \left(\sum_{i=1}^{N} \alpha_t^r(i) \beta_t^r(i) \right).
\end{aligned} \tag{18}$$

The forward variable $\alpha_t^r(i)$ and backward variable $\beta_t^r(i)$ are computed using the transformed Gaussian mixture $\mathcal{N}(\mathbf{o}_t^r; A_q \xi_{ik}, \Sigma_{ik})$.

In the second term, $L_2(S_m)$, $|S_m|$ is the number of free parameters in the transformation with the structure S_m. Γ_{mq} is the occupation counts of transformation A_q with structure S_m and can be computed as

$$\Gamma_{mq} = \sum_{r=1}^{R} \sum_{t=1}^{U^r} \sum_{(i,k) \in \omega_q} \gamma_t^r(i,k) \quad \text{given} \quad S_m \tag{19}$$

which denotes the adaptation data's total contribution to the transformation A_q with structure S_m.

The third item, $L_3(S_m)$, is the length of the code to locate a particular transformation in the tree. Each of the Q_m transformations with structure S_m is labeled by an integer from $\{1, \cdots, Q_m\}$ for identification. To specify a transformation with structure S_m for each Gaussian mixture from base class l, the labelling integer $j(m,l)$ of the transformations, which is a function of structure S_m and base class l, is identified and coded. In light of the coding literature such as [22] and [23], the approximate universal code length for a nonzero integer $j(m,l)$ is

$$I_{ml} = \log 2 \cdot \left(2 + \log_2^+ |j(m,l)| + 2 \log_2^+ \log_2^+ |j(m,l)|\right) \tag{20}$$

where $\log_2^+ |\cdot|$ is the positive part of the logarithm function. Substituting (20) into (17), we can compute the total bits needed to identify the transformations for all the Gaussian mixtures in the acoustic models. Note that $L_1(S_m)$ and $L_2(S_m)$ are computed in nats while $L_3(S_m)$ in bits; therefore, scaling factor log 2 is needed to change the different logarithm bases in the summation.

Together, the three coding lengths, $L_1(S_m)$, $L_2(S_m)$, and $L_3(S_m)$, give the description length of a transformation with structure S_m.

IV. Weighted Model Averaging

Given the MDL scores for all the competing transformation structures with a regression tree, the structure with the shortest coding length is preferred and may be considered as the best candidate among all the competing structures. However, problems may occur if only the "best" structure is adopted. First, the MDL is asymptotically accurate when applied to a large amount of data. In case of limited data, the MDL choice may vary from one data set to another and give unsatisfactory results. Moreover, when the MDL scores are close, there is no one structure that is clearly superior to the others. In this situation, weighted model averaging could provide a more stable and robust performance than a single structure.

Suppose the MDL scores for the M competing structures $\{S_1, \cdots, S_M\}$ are $\{\zeta_1, \cdots, \zeta_M\}$ with $\zeta_{\min}$ and $\zeta_{\max}$ being the minimum and maximum scores, respectively. A normalized score of the mth candidate structure S_m is defined as

$$\Delta_m = \eta \cdot \frac{\zeta_m - \zeta_{\min}}{\zeta_{\max} - \zeta_{\min}} \tag{21}$$

where η is empirically determined, and the weight for the structure S_m is computed as

$$\pi_m = \frac{e^{-\Delta_m}}{\sum_{m=1}^{M} e^{-\Delta_m}} \tag{22}$$

Assume the transformation applied to base class l $(l = 1, \cdots, L)$ with structure S_m is $A_{q(m,l)}$, the final transformation for this base class is calculated as

$$A_l = \sum_{m=1}^{M} \pi_m A_{q(m,l)} \tag{23}$$

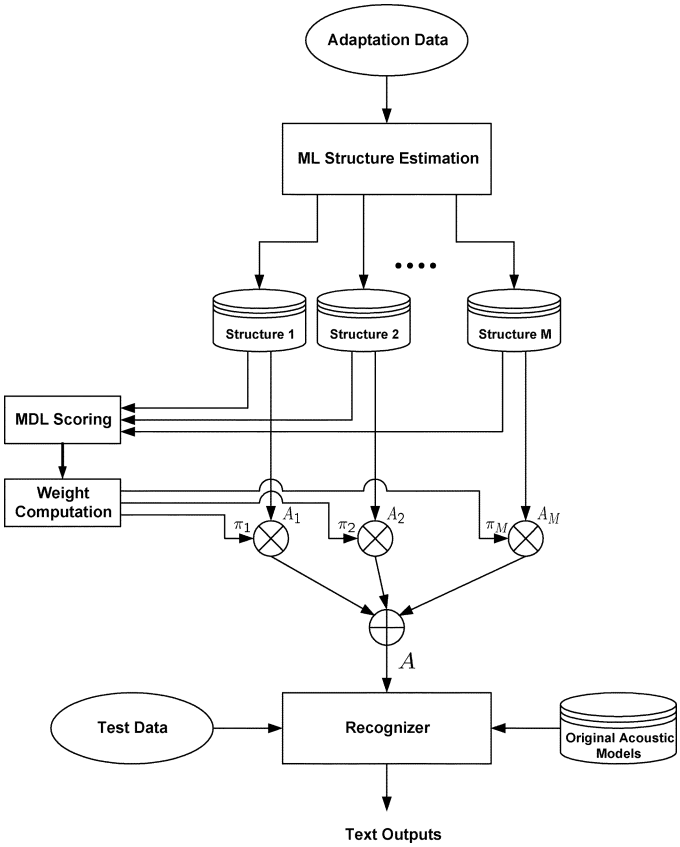

Fig. 2. Flowchart of implementation of weighted model averaging with structured MLLR transformations. The structures are appropriately tied with a regression tree.

Equation (23) represents a final transformation by the weighted average of different structured transformations.

V. Choice of Structure Form

Ideally, for the given amount of adaptation data, all possible transformation structures should be considered and their corresponding MDL scores be calculated. Suppose the transformation matrix A is $N \times (N+1)$ in dimension, then there are $2^{N\times(N+1)}$ possible structures to investigate, which is computationally prohibitive in practical situations. However, earlier research could shed light on the appropriate choice of transformation structures. For instance, [24]–[26] show that vocal tract length normalization (VTLN) in the linear spectral domain can translate into a linear transformation in the cepstral domain, which could be considered as a special case of linear regression. The transformation obtained this way has a special structure: dominant components are located along the several principal diagonals of the matrix. Fig. 3 visualizes two transformation matrices associated with two scaling factors using the approach investigated in [26] for the mel-frequency cepstral coefficients (MFCC) feature. Similar structures could also be found in [24].

The aforementioned VTLN results provide an interesting acoustic motivation on the choice for the transformation structure. Taking into account a reasonable coverage of structures and computational considerations, we choose four structures for our experiments: 3-diagonal (3D), 7-diagonal (7D), 3-block (3B), and full matrix (full).[1] The 3-block structure with sub-full-matrix for the static, first- and second-order derivatives of the transformation is widely used in MLLR speaker adaptation [20]. Table I shows the number of free parameters for the four structured transformation matrices.

TABLE I
Number of Parameters of MLLR Transformation With Different Structures

matrix structure	3-diag	7-diag	3-block	full
number of parameters	154	300	546	1560

VI. Experimental Results

Fig. 2 elaborates the implementation of the proposed weighted model averaging approach with structured transformations. Experiments are performed on the TIDIGITS and resource management (RM) databases. TIDIGITS consists of connected digit string composed of one to seven digits and RM is a continuous speech corpus where the sentences pertain to a naval resource management task. The speech data are sampled at 16 kHz. MFCC features are computed with a 25-ms frame length and a 10-ms frame shift. The feature is 39 in dimension consisting of 13 static MFCCs (including C0) and their first and second order derivatives. TIDIGITS experiments use phoneme-specific HMMs adopting a left-to-right topology with three to five states for each phoneme. A three-state silence model and one-state short pause model are also used. There are six mixtures in each state. All the Gaussian mixtures have diagonal covariance matrices. RM experiments use triphone HMMs with three states for each triphone and six Gaussian mixtures in each state.

Four sets of experiments are designed for TIDIGITS testing: male-trained-female-tested, female-trained-male-tested, adult-trained-adult-tested, and adult-trained-child-tested. The male speaker independent acoustic models are trained with 55 males and the female models with 55 females. The adult models are trained by pooling together the 55 males and 55 females. In the testing set, there are ten males, ten females and ten children. In both training and testing sets, each speaker provides 77 utterances. Before recognition, data from each speaker are extracted to adapt the speaker-independent models by MLLR. The adaptation is performed with 2, 5, 10, 15, 20, 25, 30, and 35 digits. For the RM database, speaker-independent models are trained by 72 speakers with 40 utterances from each speaker. The test set contains ten speakers with 300 utterances from each speaker. The adaptation is performed with 1, 3, 10, 50, and 100 utterances.

An MLLR regression tree with 128 base classes is created for the TIDIGITS task and 512 base classes from the RM task. To ensure matrix invertiblity during the transformation tree-tying, a minimum number of Gaussian mixtures is required at the tying nodes which is 3, 7, 13, and 39 for 3D, 7D, 3B, and full matrix, respectively. Furthermore, for reliable estimation,

[1]The structures discussed here refer to the submatrix after the first column in A in (1). For simplicity, we refer to them as the structure of A.

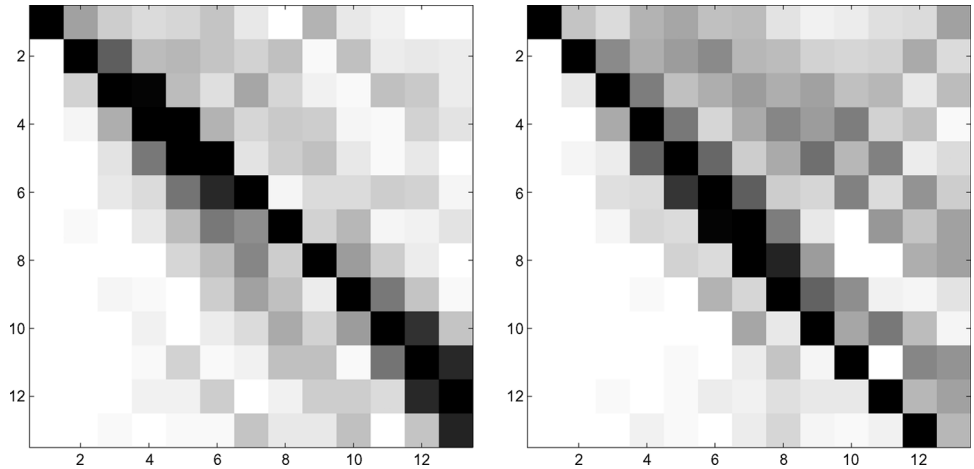

Fig. 3. Transformation matrices generated based on vocal tract length normalization with scaling factor equal to 1.2 (left) and 0.8 (right). The darker the color, the more significant the element is.

TABLE II
WORD ERROR RATE (%) OF MLLR WITH DIFFERENT STRUCTURED TRANSFORMATIONS ON TIDIGITS DATABASE. ACOUSTIC MODELS ARE TRAINED WITH MALE SPEECH AND TESTED ON FEMALE SPEECH. THE PERFORMANCE IS THE AVERAGE OVER THE TEN FEMALE SPEAKERS IN THE TEST SET. 3D, 7D, 3B, AND FULL DENOTE 3-DIAGONAL, 7-DIAGONAL, 3-BLOCK, AND FULL TRANSFORMATION MATRICES, RESPECTIVELY

	Number of adaptation digits							
	2	5	10	15	20	25	30	35
No adaptation	13.0	13.0	13.0	13.0	13.0	13.0	13.0	13.0
3D	2.7	1.4	1.3	1.1	0.8	0.8	0.8	0.6
7D	7.0	1.7	1.3	0.9	0.7	0.7	0.6	0.6
3B	-	2.9	1.2	0.7	0.7	0.6	0.6	0.5
full	-	-	4.8	1.8	0.8	0.6	0.6	0.5
MDL	3.0	1.7	1.2	0.8	0.6	0.7	0.6	0.6
MDL-Ave	2.1	1.3	0.9	0.7	0.6	0.6	0.6	0.5

TABLE III
WORD ERROR RATE (%) OF MLLR WITH DIFFERENT STRUCTURED TRANSFORMATIONS ON TIDIGITS DATABASE. ACOUSTIC MODELS ARE TRAINED WITH FEMALE SPEECH AND TESTED ON MALE SPEECH. THE PERFORMANCE IS THE AVERAGE OVER THE TEN MALE SPEAKERS IN THE TEST SET. 3D, 7D, 3B, AND FULL DENOTE 3-DIAGONAL, 7-DIAGONAL, 3-BLOCK, AND FULL TRANSFORMATION MATRICES, RESPECTIVELY

	Number of adaptation digits							
	2	5	10	15	20	25	30	35
No adaptation	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4
3D	1.7	0.9	0.9	0.9	0.6	0.6	0.6	0.6
7D	3.6	1.9	1.2	1.0	0.8	0.7	0.6	0.6
3B	-	5.6	1.1	0.8	0.8	0.8	0.7	0.6
full	-	-	4.0	1.5	0.9	0.8	0.8	0.6
MDL	1.7	1.3	1.2	0.9	0.7	0.7	0.7	0.6
MDL-Ave	1.0	1.0	0.8	0.8	0.6	0.6	0.6	0.6

a threshold has to be set for each transformation structure depending on its number of free parameters. In this paper, we choose the threshold to be approximately equal to the number of parameters for each structure. That is, 150, 300, 550, and 1500 are the occupation counts for a valid transformation estimation with 3D, 7D, 3B, and full matrix, respectively. The scaling factor η in (21) is set to 2.0. η is tuned to give reasonable weights for the structures. Values around 2.0 will give the best performance according to our experiments and they are consistent across recognition tasks.

Tables II–V show the TIDIGITS experimental results with four transformation structures using different amounts of adaptation data. Baseline results are without adaptation. Adaptation results using the "single" best structure based on MDL, and using the averaged transformation across all four structures weighted using the MDL scores are denoted as "MDL" and "MDL-Ave," respectively. The 3-block and full matrix structure results with very limited data (e.g., two digits for 3-block matrix structure and two and five digits for full matrix structure) are not shown in the tables since even the global tying for the transformation cannot meet the occupation threshold requirement, and the results are thus not meaningful. Table VI shows the experimental results for the RM database.

From the tables, structures with less parameters (3D or 7D) tend to give better performance than those with more parameters (3B or full) when the amount of adaptation data is small. When the amount of data increases, however, the situation is reversed. This is mainly due to the tradeoff between transformation granularity and descriptive ability. By choosing the single "best" model with the minimum score, MDL gives a better balanced performance with respect to the amount of adaptation data. Very often, MDL is able to obtain the best performance

TABLE IV
WORD ERROR RATE (%) OF MLLR WITH DIFFERENT STRUCTURED TRANSFORMATIONS ON TIDIGITS DATABASE. ACOUSTIC MODELS ARE TRAINED AND TESTED ON ADULT SPEECH (BOTH MALE AND FEMALE). THE PERFORMANCE IS THE AVERAGE OVER THE 20 ADULT SPEAKERS IN THE TEST SET. 3D, 7D, 3B, AND FULL DENOTE 3-DIAGONAL, 7-DIAGONAL, 3-BLOCK, AND FULL TRANSFORMATION MATRICES, RESPECTIVELY

	Number of adaptation digits							
	2	5	10	15	20	25	30	35
No adaptation	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
3D	1.1	0.7	0.9	0.8	0.7	0.6	0.7	0.7
7D	2.0	0.8	0.9	0.8	0.8	0.7	0.7	0.7
3B	-	2.4	0.8	0.7	0.8	0.7	0.8	0.6
full	-	-	2.6	1.0	0.7	0.7	0.7	0.7
MDL	1.1	0.8	0.8	0.8	0.7	0.7	0.7	0.7
MDL-Ave	1.0	0.7	0.7	0.7	0.7	0.6	0.7	0.6

TABLE V
WORD ERROR RATE (%) OF MLLR WITH DIFFERENT STRUCTURED TRANSFORMATIONS ON TIDIGITS DATABASE. ACOUSTIC MODELS ARE TRAINED ON ADULTS AND CHILDREN. THE PERFORMANCE IS THE AVERAGE OVER THE TEN KID SPEAKERS IN THE TEST SET. 3D, 7D, 3B, AND FULL DENOTE 3-DIAGONAL, 7-DIAGONAL, 3-BLOCK, AND FULL TRANSFORMATION MATRICES, RESPECTIVELY

	Number of adaptation digits							
	2	5	10	15	20	25	30	35
No adaptation	2.9	2.9	2.9	2.9	2.9	2.9	2.9	2.9
3D	2.4	1.4	1.2	1.3	1.2	1.1	1.1	1.0
7D	5.0	1.6	1.2	0.9	1.0	1.1	1.0	0.8
3B	-	9.3	1.3	0.9	0.8	1.0	0.9	0.8
full	-	-	8.2	3.0	1.3	1.1	1.1	1.0
MDL	2.4	1.4	1.2	0.9	0.9	1.0	0.9	0.8
MDL-Ave	1.9	1.4	1.1	0.9	0.9	0.9	0.9	0.6

among the four candidate structures given a certain amount of data. Weighted model averaging across all structures based on the normalized MDL scores gives more consistent and robust performance than MDL alone.

Although MDL-Ave yields the best WER in the TIDIGITS experiments, the statistical significance level of MDL-Ave compared to certain structures (e.g., 3-diag) using the matched-pair test [27] is not high (around 0.2–0.3). This is primary because the test set is not very large and the adaptation baseline for the structures is high. In the experiments on the RM database, MDL-Ave shows a statistically significant difference compared the best-performed structure (full structure) at a significant level of 0.024.

TABLE VI
WORD ERROR RATE (%) OF MLLR WITH DIFFERENT STRUCTURED TRANSFORMATIONS ON RM DATABASE

	Number of adaptation utterances				
	1	3	10	50	100
No adaptation	7.5	7.5	7.5	7.5	7.5
3D	5.9	5.7	5.4	5.0	4.8
7D	5.8	5.5	5.1	4.7	4.5
3B	6.7	5.5	4.9	4.5	4.2
full	7.5	7.0	5.2	4.4	4.1
MDL	5.8	6.2	5.2	4.4	4.1
MDL-Ave	5.9	5.5	4.9	4.2	3.9

VII. DISCUSSION

The tying of the MLLR transformation with competing structures in a regression tree, introduced in this paper, is different from the model selection problems in nested linear regression coefficient selection [15] or the optimal tree cut approaches [18], [19]. This is because the tied transformations can utilize data from overlapped Gaussian mixture sets, which are neither nested nor a partition of the Gaussian mixture space. This makes it a more interesting problem. Furthermore, to locate the transformations, the regression tree has to be traversed to get to certain nodes. In this situation, MDL seems to be a superior choice to Akaike information criterion (AIC) [28] or Bayesian information criterion (BIC)[29] because MDL could be interpreted from the coding point of view, and the traverse of the tree to locate the transformations can be taken into account as a part of the model itself. This cannot be easily dealt with by AIC or BIC.

The coding of the model parameters in MDL is related to the Fisher information matrix which could be directly employed to calculate the MDL score as in [19]. However, the asymptotic form used in this paper may be a good approximation even without a large amount of data under certain Bayesian assumptions [15] and this form also has its computational advantages. As we know, in speech recognition, the connection between a good fit of model based on the ML criterion and good performance in the Viterbi decoding is not strong. Moreover, most of the model selection criteria including MDL, AIC, and BIC are obtained based on large sample theory. Therefore, they may not select the best model in some cases especially when only a limited amount of data is given.

Compared with MDL alone, the weighted model averaging strategy renders more robust performance in most cases. This is because although MDL cannot always choose the best structure, it does give a good "guess" on the goodness-of-fit of the structures. Therefore, a reasonable weight of the structure can

produce better results. The tying pattern of the transformation with the regression tree is decided by (a) the structure and (b) the threshold of reliable estimation for the structure; both can be handled by the MDL weighted model averaging. Different structures other than those investigated in this paper are also possible and the weighted model averaging algorithm can be carried out accordingly. The major computational complexity of the algorithm comes from calculation of individual transformations and computation of MDL scores afterwards. This computational complexity grows linearly with the number of competing structures.

Structural MAP and eigenvoice approaches are both effective techniques which have obtained excellent results. The MDL and MDL-Ave algorithms discussed in this paper basically address the sparse data problem within the MLLR framework. In the structural MAP approach, it is difficult to choose the optimal tree structure for adaptation. The eigenvoice approach adapts acoustic models from prior models in the eigen-space which is similar to the MDL-Ave in terms of linear combination. However, eigenvoice is most effective when a limited amount of data is available. When the adaptation data increase, its performance saturates quickly [8]. MDL-Ave does not have this problem. It gives robust performance across various amounts of adaptation data. Despite the difference among the MDL-Ave, structural MAP and eigenvoice, they have similarities from the parameter smoothing perspective. MDL-Ave performs parameter smoothing on predefined structures, structural MAP on prior distributions and eigenvoice on prior eigen-models.

VIII. Summary

In this paper, we investigate a robust maximum likelihood linear regression speaker adaptation approach with weighted model averaging across a variety of transformation structures. A general form of the maximum likelihood estimation of the structured transformation is given. The MDL is adopted to describe the balance between transformation granularity and descriptive ability of the structured transformations tied using a regression tree. Based on the normalized MDL scores, transformations are averaged across all structures. Experimental results show that the proposed approach obtains robust performance with respect to the amount of adaptation data.

Acknowledgment

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The authors would like to thank Prof. Y. Wu and Prof. M. Hansen for discussions on model selection.

References

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

[2] M. Padmanabhan and S. Dharanipragada, "Maximum-likelihood nonlinear transformation for acoustic adaptation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 572–578, Nov. 2004.

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

[4] K. Shinoda and C. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE-SP Workshop Automatic Speech Recognition and Understanding*, 1997, pp. 381–388.

[5] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, May 2001.

[6] V. Digalakis *et al.*, "Rapid speech recognizer adaptation to new speakers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1999, pp. 765–768.

[7] E. Bocchieri *et al.*, "Correlation modeling of MLLR transform biases for rapid HMM adaptation to new speakers," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 1999, pp. 773–776.

[8] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptaion in cigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[9] S. D. Peters, "Hypothesis-driven adaptation (HYDRA: a flexible eigenvoice architecture)," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2001, pp. 349–352.

[10] Y. Tsao *et al.*, "Segmental eigenvoice for rapid speaker adaptation," in *Proc. Eur. Conf. Speech Communication and Technology*, 2001, pp. 1269–1272.

[11] B. Mak *et al.*, "A study of various composite kernels for kernel eigenvoice speaker adaptation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2004, pp. 325–328.

[12] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[13] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 717–720.

[14] I. Illina, "Tree-structured maximum *a posteriori* adaptation for a segment-based speech recognition system," in *Proc. Int. Conf. Spoken Language Process.*, 2002.

[15] M. Hansen and B. Yu, "Model selectin and the principle of minimum description length," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[16] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[17] R. Stine, Model Selection Using Information Theory and the MDL Principle [Online]. Available: http://www-stat.wharton.upenn.edu/~stine

[18] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. Eur. Conf. Speech Communication and Technology*, 1997, pp. 99–102.

[19] S. Wang and Y. Zhao, "Online Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 663–677, Nov. 2001.

[20] S. Young *et al., The HTK Book (Version 3.1)*. Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2001.

[21] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[22] P. Elias, "Universal codeword sets and representation of the intergers," *IEEE Trans. Inf. Theory*, vol. IT-21, no. 2, pp. 194–203, Mar. 1975.

[23] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, pp. 416–431, 1983.

[24] M. Pitz and H. Ney, "Vocal tract normalization as linear transformation of MFCC," in *Proc. Eur. Conf. Speech Communication and Technology*, 2003, pp. 1445–1448.

[25] G. Ding, Y. Zhu, C. Li, and B. Xu, "Implementing vocal tract length normalization in the MLLR framework," in *Proc. Int. Conf. Spoken Language Processing*, 2002, pp. 1389–1392.

[26] X. Cui and A. Alwan, "Adaptation of children's speech with limited data based on formant-like peak alignment," *Comput. Speech Language*, to be published.

[27] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 532–535.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[29] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

Xiaodong Cui (M'05) received the B.S. degree from Shanghai Jiao Tong University (with the highest honor), Shanghai, China, in 1996, the M.S. degree from Tsinghua University, Beijing, China, in 1999 and the Ph.D degree from the University of California, Los Angeles, in 2005, all in electrical engineering.

From 2005 to 2006, he was a Research Staff Member with the DSP Solutions R&D Center, Texas Instruments, Dallas, TX. Since 2006, he has been a Research Staff Member with the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include speech recognition, speech processing, statistical signal processing, machine learning, and pattern recognition.

Abeer Alwan (SM'00) received the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1992.

Since then, she has been with the Electrical Engineering Department at the University of California, Los Angeles (UCLA), as an Assistant Professor (1992–1996), Associate Professor (1996–2000), Professor (2000-present), and Vice Chair of Graduate Affairs (2003-present). She established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (http://www.icsl.ucla.edu/~spapl). Her research interests include modeling human speech production and perception mechanisms and applying these models to improve speech processing applications such as noise-robust automatic speech recognition, compression, and synthesis.

Prof. Alwan was the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). She is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served, as an elected member, on the Acoustical Society of America Technical Committee on Speech Communication (1993–1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996–2000) and on Speech Processing (1996–2001, 2005-present). She is a member of the Editorial Board of Speech Communication and was an Editor-in-Chief of that journal (2000–2003.) She is a Fellow of the Acoustical Society of America.