Fast Model Selection Based Speaker Adaptation for Nonnative Speech

Xiaodong He and Yunxin Zhao, *Senior Member, IEEE*

***Abstract*—In this paper, the problem of adapting acoustic models of native English speech to nonnative speakers is addressed from a perspective of adaptive model complexity selection. The goal is to dynamically select model complexity for each nonnative talker so as to optimize the balance between model robustness to pronunciation variations and model detailedness for discrimination of speech sounds. A maximum expected likelihood (MEL) based technique is proposed to enable reliable complexity selection when adaptation data are sparse, where expectation of log-likelihood (EL) of adaptation data is computed based on distributions of mismatch biases between model and data, and model complexity is selected to maximize EL. The MEL based complexity selection is further combined with MLLR to enable adaptation of both complexity and parameters of acoustic models. Experiments were performed on WSJ1 data of speakers with a wide range of foreign accents. Results show that the MEL based complexity selection was feasible when using as little as one adaptation utterance, and it was able to dynamically select proper model complexity as the adaptation data increased. Compared with the standard MLLR, the MEL + MLLR method led to consistent and significant improvement to recognition accuracy on nonnative speakers, without performance degradation on native speakers.**

***Index Terms*—Maximum expected likelihood, model selection, nonnative speech recognition, speaker adaptation.**

I. Introduction

CURRENT English speech recognition systems are commonly trained from speech data of native English speakers. Although the systems work very well for native talkers, their performance degrades dramatically when recognition is performed on speech with heavy foreign accents. Due to wide varieties of foreign accents, different proficiency levels of English speaking and limited data, it is in general difficult to train a set of acoustic models for each specific accent. Therefore, improving the performance of the state-of-the-art speech recognition systems for nonnative speech remains a challenging task.

Several efforts have been made to improve recognition performance for nonnative speech [1]–[3]. A straightforward approach is to use general speaker adaptation techniques to adapt speaker-independent models to the foreign-accent characteristics of a new speaker. Commonly used adaptation algorithms include Maximum Likelihood Linear Regression (MLLR) [4] and Maximum *a posteriori* (MAP) learning [5]. It has been recognized that although speaker adaptation can improve recognition accuracy for both native and nonnative English speakers, a much larger amount of adaptation speech data is needed from a foreign-accent speaker than a native English speaker to achieve a comparable level of recognition accuracy [1], [2].

Research efforts on using multilingual acoustic modeling techniques for nonnative speech recognition have been reported in recent years [6], [7]. In multilingual acoustic modeling, a universal phone set is used. Phonemes of several languages are mapped to the universal phone set and speech data of these languages are pooled to train the acoustic model in order to capture pronunciation variations of the same phoneme in different languages [6]–[11]. Although promising improvements to nonnative speech recognition were observed for small tasks with this approach, two problems exist. First, much more speech data, including speech of foreign languages, are needed to train a multilingual acoustic model. Second, compared with using acoustic model trained from native speech alone, although multilingual acoustic model improved nonnative speech recognition, it degraded native speech recognition in some cases [6], [7].

A closely related problem to nonnative speaker adaptation is regional dialect speaker adaptation. Digalakis *et al.* investigated adapting acoustic models to fit speakers with dialect accents [12], [13]. In [12], Maximum Likelihood Stochastic Transformation (MLST) was proposed to estimate multiple linear transforms for each model cluster in model adaptation. Although a significant performance improvement was achieved, much more data than that of MLLR were needed, where only one linear transform was estimated for each model cluster in MLLR. In [13], in order to achieve a good performance when adaptation data were sparse, speech data of prototype speakers from target dialect regions were used to generate a set of basis linear transformations and a small amount of new speaker's speech was used to estimate the transform combination weights. In their experiments of Swedish dialect speaker adaptation, the adaptation performance exceeded that of MLLR greatly when the amount of adaptation data was very small. However, a large number of prototype speakers and an adequate amount of data from target dialect speech were needed to form a set of powerful transformation basis.

Compared with recognition of regional accent speech, nonnative speech is much more difficult because it is an interfusion between a speaker's native language and a target nonnative language [14]. Several studies on the properties of nonnative

Manuscript received July 12, 2002; revised February 8, 2003. This work was supported in part by National Institutes of Health under Grant NIH 1 R01 DC04340-01 A2 and the National Science Foundation under Grant NSF EIA 9911095. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

The authors are with the Department of CECS, University of Missouri, Columbia, MO 65211 USA (e-mail: xhb1a@mizzou.edu; zhao@cecs.missouri.edu).

Digital Object Identifier 10.1109/TSA.2003.814379

speech have been reported recently [14]–[16]. In [15], it was found that intelligibility of nonnative speech degraded significantly from that of native speech, corresponding to a difference in signal-to-noise ratio (SNR) of about 3 dB, and the degradation was mainly due to the confusion of vowels, especially those not in the speaker's native language. In [16], the phenomena of phone variation and substitution in nonnative speech were investigated. It was shown that phone variation and substitution properties changed greatly with different foreign accents and phone contexts. It was suggested in [14] that, beside the reduced intelligibility of nonnative speech, the ever-increasing details of the state-of-the-art acoustic models that are tuned for native speakers are not necessarily beneficial to the performance of nonnative speech recognition due to reduced tolerance of the models to variations in nonnative speech.

A speaker adaptation strategy that focuses on adaptively selecting a proper model complexity for each nonnative English speaker has recently been proposed [17], [18]. This approach was motivated from the fact that highly detailed English acoustic models with sharp distributions of very narrow allophone classes do not fit well to speech data with heavy foreign accents, while a certain level of context-dependent modeling needs to be maintained for discrimination among phones [14], [17]. Experimental results of [17] showed that between native speakers and nonnative speakers, the curves of model complexity versus recognition performance were significantly different. Highly detailed acoustic models that produced the best recognition result for native speakers were worst for nonnative speakers. A conjecture is therefore made that an intermediate level of acoustic model complexity determined from adaptation speech may work best for a foreign accent talker.

Among various model complexity selection methods, maximum likelihood (ML) based model selection has been widely used [19]. In the data-rich case, independent "validation data" is employed for model selection. The model that gives maximum likelihood of these data will be selected as the optimal model. However, the requirement of large amount of data by ML-based model selection prevents its application in on-line fast speaker adaptation. In [18], by using a small amount of adaptation data from a nonnative English speaker, a combined ML and pseudo-likelihood (PL) based tree pruning was performed to select complexity of an acoustic model that was trained by native English speech for the nonnative speaker. The spectral mismatch between adaptation speech data and acoustic model was represented by a global bias, which was estimated from adaptation data by using phonetic decision trees with Single Gaussian Densities (SGD) at tree nodes. The global bias was then used to compute a PL value for each SGD of each tree node, and ML/PL based tree pruning was preformed for model selection. In [18], although a significant improvement was resulted from model selection in recognizing speech with heavy foreign accents, there were drawbacks. First, since speech recognition was performed by Gaussian Mixture Density (GMD) based HMM phone models, model selection based on phonetic decision trees of SGDs was not sufficiently precise. Second, single global bias was not adequate in characterizing detailed mismatches between a speaker's speech and the phone models.

In the current paper, a maximum expected likelihood (MEL) based algorithm is proposed for effective model complexity selection from a small amount of adaptation data, and comprehensive experimental evaluation results are reported for a wide range of foreign accents. The algorithm consists of three major steps, where the first step is for model training, and the second and third steps are for model selection. In the first step, allophone states are hierarchically clustered through a clustered phonetic decision tree (CPDT), and each node of the tree corresponds to a tied allophone state. A tied allophone state is generated by tying all the allophone states of the terminal nodes of its subtree, and for each tree node, a GMD is estimated. In the second step, given a certain amount of adaptation data, each feature analysis vector is assigned to one dominant Gaussian component (GC) by Viterbi alignment, and a bias between the sample data mean and the model mean is calculated for each GC of each terminal tree node that has adaptation data. Assuming that the biases within an allophone state cluster are i.i.d., the distribution parameters of biases are estimated, and expected log-likelihood is computed at each tree node. In the third step, a bottom-up tree pruning is carried out to select the optimal model complexity that maximizes expected log-likelihood (EL) over the tree nodes.

This paper is organized as follows. The concept of MEL based model selection is discussed in Section II. Several implementation issues, including construction of the clustered phonetic decision tree, hierarchical computation of EL, and adaptive model selection, are presented in detail in Section III. Experiment setup and data are described and results are discussed in Section IV. Finally, a conclusion is drawn in Section V.

II. Rational of MEL Based Model Selection and Adaptation

Most state-of-the-art HMM based acoustic modeling techniques employ very sharp distributions to describe narrowly defined acoustic speech units, and these techniques work very well for recognition of native speech. However, due to diversity of nonnative speech, less detailed models that are more robust to variations are better suited for nonnative speech. Moreover, optimal model complexity may also be different for individual speakers. Therefore, it is desirable to adaptively determine the proper level of model complexity for each specific speaker by using a small amount of adaptation data.

The problem of model complexity selection can be addressed from the perspective of state tying. In conventional state tying, each triphone is modeled by an HMM, and for triphones that have the same centre phoneme, a phonetic decision tree (PDT) is built for the same indexed states of triphone HMMs, and the root node of a PDT corresponds to a context-independent state of that centre phoneme [20]. Each tree node in the PDT represents an allophone cluster and corresponds to a distribution of the allophones tied in that node. As shown in Fig. 1, a tree cut is a collection of nodes that can separate the whole tree into an upper part and a lower part. Data distributions of the nodes in a tree cut of a PDT constitute an acoustic model of a HMM state of a phone unit. Fixing the distribution complexity at each node, a high-level tree cut corresponds to a less detailed model, and

a low-level tree cut corresponds to a more detailed model. A proper tree cut should be selected for each particular speaker.

In certain scenarios of speaker adaptation, only a small amount of adaptation data is available. In such cases, only a small number of PDT nodes have adaptation data, and the matching between acoustic model and adaptation data cannot be reliably measured by direct likelihood calculation. To address this problem, a method of expected likelihood is proposed. To simplify discussions on the concept of EL, the distribution of speech features at each PDT node is assumed as a one-dimensional single Gaussian density, with the understanding that mixtures of multi-dimensional Gaussian densities are used in actual acoustic models, which will be discussed in Section III of this paper.

Consider a node q of a PDT with a data set X_q, where the size of X_q is N_q and the sample data mean is $\bar{x}_q$. Given the node Gaussian pdf $\lambda_q = N\left(\mu_q, \sigma_q^2\right)$, the sample data variance and the bias between the model mean and the data mean are defined as $v_q^2 = (1/N_q)\sum_{i=1}^{N_q}(x_i - \bar{x}_q)^2$ and $b_q = \bar{x}_q - \mu_q$. The model and data distribution parameters are illustrated in Fig. 2. The average log-likelihood per data sample is therefore computed as

$$AL(X_q|\lambda_q) = -\frac{1}{2}\left[\ln(2\pi) + \ln\left(\sigma_q^2\right) + \frac{v_q^2}{\sigma_q^2} + \frac{b_q^2}{\sigma_q^2}\right] \quad (1)$$

and is defined as the log-likelihood at the node q. Assuming that the variance of data is proportional to the variance of the model, i.e., $v_q^2 = \text{const} \cdot \sigma_q^2$, the expectation of the log-likelihood at node q (EL_q) over the distribution of b_q is

$$E[AL(X_q|\lambda_q)] = -\frac{1}{2}\left[\ln(2\pi) + \ln\left(\sigma_q^2\right)\right] - \frac{1}{2}\text{const} - \frac{1}{2\sigma_q^2}E\left(b_q^2\right). \quad (2)$$

If the distribution of the bias b_q can be estimated at each node q, then the above expectation can be readily computed. The expected log-likelihood of a tree cut is defined as a sum of weighted EL_q's of all nodes in that tree cut, and the tree cut that leads to maximum expected log-likelihood can be selected as the optimal tree cut. Through the PDT, allophones are hierarchically clustered. In the data-sparse situation, the bias distributions of nodes that lie in a specified cluster are tied, where biases are computed from the terminal GCs with sufficient adaptation data and the computed biases are taken as samples of the tied bias distribution. As the result, a tied bias distribution can be reliably estimated from the terminal nodes in a local subtree that has enough samples of bias data.

III. MEL Based Model Selection

In system implementation, the distribution of speech features at each tree node is a GMD instead of a SGD, and model selection is performed on a clustered phonetic decision tree instead of PDTs. In this section, we describe the details of building a CPDT, the computation of EL based on GMD at each node, the estimation of tied bias distributions, and the MEL based tree pruning procedure for optimal tree cut selection.

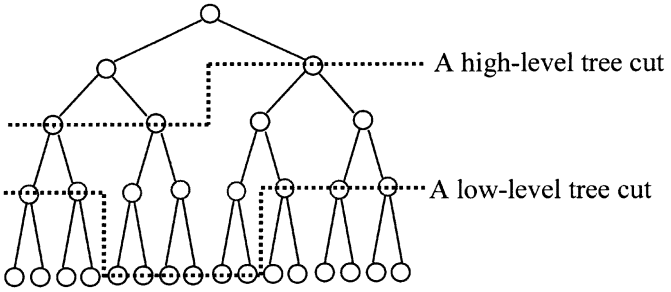

Fig. 1. Model selection based on a phonetic decision tree.

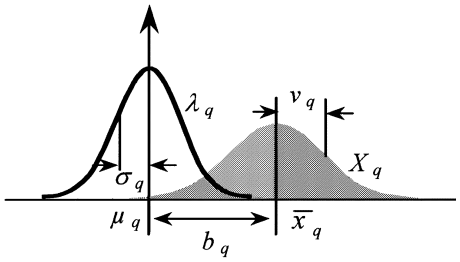

Fig. 2. Parameters and statistics involved in computation of log-likelihood.

A. Construction of Clustered Phonetic Decision Tree

In order to share bias distributions among phone units, a clustered phonetic decision tree is built in two steps as follows.

First, for each phone state, a state-tying binary phonetic decision tree is built as in [20]. Initially each triphone state is modeled by one Gaussian density and the Gaussian densities are placed into a single cluster at the root of the tree. Then a yes/no question about phone context from a pre-defined set is asked to divide the triphones in the cluster into two groups based on their answers to that question, and the question that leads to the maximum likelihood increment is adopted for splitting the node into two children nodes. This process is repeated for each node until the increment in log likelihood due to cluster split falls below a pre-defined threshold, or the number of data samples at each node is less than a threshold. As the result, each node of a PDT corresponds to a tied state of allophones, and the root node corresponds to the state of the phone unit with arbitrary contexts, or context-independent.

Second, these PDTs are grouped in a clustered phonetic decision tree through a binary "super" tree. Each of the PDT root nodes is represented by one Gaussian density. At the beginning, the root nodes of the PDTs are assigned to a single set at the root of the "super" tree. Then binary-split K-means clustering is performed to split the root nodes of the PDTs into two new sets, or equivalently two children nodes. This procedure continues until each node has only one PDT root node, where the corresponding PDT obtained in the first step is then attached. In the K-means clustering of SGDs, the Mahalanobis distance measure is used for each Gaussian density pair $\lambda_a = N(\mu_a, \Sigma_a)$ and $\lambda_b = N(\mu_b, \Sigma_b)$, i.e., $d(\lambda_a, \lambda_b) = (\mu_a - \mu_b)^T[(\Sigma_a + \Sigma_b)/2]^{-1}(\mu_a - \mu_b)$. As the result of clustering, a binary "super" tree is built on top of the PDTs and the overall tree structure is referred to as a CPDT.

For each PDT node, a GMD can be estimated based on the segmental expectation-maximization (EM) algorithm. The details are discussed in Section IV-A.

B. Expectation of Log-Likelihood Based on GMD

Given a Gaussian mixture density λ with K Gaussian components (GC) and an arbitrary data set $X = \{x_1, x_2, \ldots, x_N\}$, the log-likelihood of X is computed as

$$L(X|\lambda) = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} w_k \cdot N(x_i; \mu_k, \Sigma_k) \right] \tag{3}$$

where w_k, μ_k, Σ_k are the weight, mean vector and covariance matrix of the kth GC, respectively. For each x_i, if the log-likelihood value can be approximated by a dominant GC, we can get $L(X|\lambda) \approx \sum_{k=1}^{K} \sum_{j \in S_k} \ln [w_k N(x_j; \mu_k, \Sigma_k)]$, where S_k is the index set of feature data that are assigned to the kth GC, with $N_k = |S_k|$, and $\sum_{k=1}^{K} N_k = N$. Assuming $\mu_k = [\mu_{k,1}, \ldots, \mu_{k,D}]^T$, and $\Sigma_k = \mathrm{diag}[\sigma^2_{k,1}, \ldots, \sigma^2_{k,D}]$, then

$$L(X|\lambda) \approx -\frac{1}{2} \sum_{k=1}^{K} \sum_{d=1}^{D} \left[N_k \ln \left(2\pi\sigma^2_{k,d}\right) + \frac{1}{\sigma^2_{k,d}} \sum_{j \in S_k} (x_{j,d} - \mu_{k,d})^2 \right] + \sum_{k=1}^{K} N_k \ln(w_k). \tag{4}$$

In the dth feature dimension, from the data that are assigned to the kth GC, the sample mean and sample variance are computed by $\bar{x}_{k,d} = (1/N_k) \sum_{j \in S_k} x_{j,d}$ and $v^2_{k,d} = (1/N_k) \sum_{j \in S_k} (x_{j,d} - \bar{x}_{k,d})^2$, and the bias between the model mean and the sample mean is computed by $b_{k,d} = \bar{x}_{k,d} - \mu_{k,d}$. Then

$$L(X|\lambda) \approx -\frac{1}{2} \sum_{k=1}^{K} N_k \sum_{d=1}^{D} \left[\ln(2\pi\sigma^2_{k,d}) + \frac{v^2_{k,d}}{\sigma^2_{k,d}} + \frac{b^2_{k,d}}{\sigma^2_{k,d}} \right] + \sum_{k=1}^{K} N_k \ln(w_k). \tag{5}$$

Assume that the variance of the data is proportional to the variance of the model, i.e., $v^2_{k,d} = \mathrm{const}_d \cdot \sigma^2_{k,d}, \forall\, k, d$, and the number of feature data assigned to the kth GC is proportional to the weight of that GC, i.e., $N_k = N \cdot w_k$. Then the average log-likelihood per data sample becomes

$$AL(X|\lambda) \approx -\frac{1}{2} \sum_{k=1}^{K} w_k \sum_{d=1}^{D} \left[\ln(2\pi\sigma^2_{k,d}) + \mathrm{const}_d + \frac{b^2_{k,d}}{\sigma^2_{k,d}} \right] + \sum_{k=1}^{K} w_k \ln(w_k) \tag{6}$$

and the expectation of log-likelihood over the distribution of $b_{k,d}$ is computed by

$$E[AL(X|\lambda)] = -\frac{1}{2} \left[\sum_{k=1}^{K} w_k \sum_{d=1}^{D} \ln\left(2\pi\sigma^2_{k,d}\right) + \sum_{d=1}^{D} \mathrm{const}_d + \sum_{k=1}^{K} w_k \sum_{d=1}^{D} \frac{E\left(b^2_{k,d}\right)}{\sigma^2_{k,d}} \right] + \sum_{k=1}^{K} w_k \ln(w_k). \tag{7}$$

C. Properties of Bias Distribution

In order to compute (7), the expectation term $E\left(b^2_{k,d}\right)$ needs to be estimated. Viewing the bias $b_{k,d}$ as a Gaussian random variable, the estimation of the distribution parameters for $b_{k,d}$ is discussed below. For simplicity of notations, the feature dimension index d is omitted in the subsequent discussions.

A subtree that is rooted at the node q of a CPDT is shown in Fig. 3. The term "terminal GC" denotes a Gaussian component of a GMD at a terminal node. The term "full terminal GC" denotes a terminal GC that has been assigned more than a specified amount of adaptation data by Viterbi forced alignment, and the term "nonempty terminal GC" denotes a terminal GC that has been assigned some adaptation data but the amount is less than the specified threshold. The term "full internal node" denotes a node that covered more than a specified number of "samples of bias data" under its subtree.

Relation of Bias Distributions Based on SGD and GMD: In the training stage, GMD and SGD are both estimated for each tree node. Consider GMD and SGD of the same node q, where $\mathrm{GMD}_q = \sum_{k=1}^{K} w_{q,k} N\left(\mu_{q,k}, \sigma^2_{q,k}\right)$, $\mathrm{SGD}_q = N\left(\mu_q, \sigma^2_q\right)$, and denote the respective biases by $b^{(\mathrm{GMD})}_{q,k} = \bar{x}_{q,k} - \mu_{q,k}$ and $b^{(\mathrm{SGD})}_q = \bar{x}_q - \mu_q$. Based on the previous assumption that the number of feature data assigned to the kth GC is proportional to the weight of that GC, $b^{(\mathrm{SGD})}_q$ is approximated as a linear combination of $b^{(\mathrm{GMD})}_{q,k}$ as $b^{(\mathrm{SGD})}_q \approx \sum_{k=1}^{K} w_{q,k} b^{(\mathrm{GMD})}_{q,k}$. Assuming that $b^{(\mathrm{GMD})}_{q,k}$, $k = 1, \ldots, K$, are i.i.d. and follow a Gaussian distribution $N\left(e^{(\mathrm{GMD})}_q, s^{(\mathrm{GMD})^2}_q\right)$, then $b^{(\mathrm{SGD})}_q$ also follows a Gaussian distribution $N\left(e^{(\mathrm{SGD})}_q, s^{(\mathrm{SGD})^2}_q\right)$, with $e^{(\mathrm{SGD})}_q \approx e^{(\mathrm{GMD})}_q$, and $s^{(\mathrm{SGD})^2}_q \approx s^{(\mathrm{GMD})^2}_q \sum_{k=1}^{K} w^2_{q,k}$.

Relation of Biases Based on Internal Node and Terminal Node SGDs: Refer to Fig. 3 and consider the SGD of an internal node q and the SGDs of the terminal nodes $i = 1$, 2, 3, 4 below the node q. We have $b^{(\mathrm{SGD})}_i = \bar{x}_i - \mu_i$, and $b^{(\mathrm{SGD})}_q = \sum_{i=1}^{4} \alpha_i b^{(\mathrm{SGD})}_i$, where α_i is the contribution weight of bias $b^{(\mathrm{SGD})}_i$ of node i to the bias $b^{(\mathrm{SGD})}_q$ of node q, with $\sum_{i=1}^{4} \alpha_i = 1$.

Relation of Bias Distributions Based on Internal Node and Terminal Node GMDs: Based on the relation of biases between GMD and SGD of the same node drawn above and refer to Fig. 3, we have $b^{(\mathrm{SGD})}_q = \sum_{i=1}^{4} \alpha_i b^{(\mathrm{SGD})}_i \approx \sum_{i=1}^{4} \alpha_i \sum_{k=1}^{K} w_{i,k} b^{(\mathrm{GMD})}_{i,k}$.

For a local subtree with root q, it is assumed that the bias distributions of the GCs of its terminal nodes are tied and the tied distribution is a Gaussian density $N\left(e^{(\mathrm{GMD})}_{\mathrm{termi}}, s^{(\mathrm{GMD})^2}_{\mathrm{termi}}\right)$. Therefore $e^{(\mathrm{SGD})}_q \approx e^{(\mathrm{GMD})}_{\mathrm{termi}}$, and $s^{(\mathrm{SGD})^2}_q \approx s^{(\mathrm{GMD})^2}_{\mathrm{termi}} \sum_{i=1}^{4} \sum_{k=1}^{K} (\alpha_i w_{i,k})^2$. Then

$$E\left(b^2_{q,k}\right) = E\left(b^{(\mathrm{GMD})^2}_{q,k}\right) = e^{(\mathrm{GMD})^2}_q + s^{(\mathrm{GMD})^2}_q, \quad k = 1, \ldots, K \tag{8}$$

where $e^{(\mathrm{GMD})}_q \approx e^{(\mathrm{SGD})}_q \approx e^{(\mathrm{GMD})}_{\mathrm{termi}}$, and $s^{(\mathrm{GMD})^2}_q \approx s^{(\mathrm{SGD})^2}_q / \sum_{k=1}^{K} w^2_{q,k} \approx s^{(\mathrm{GMD})^2}_{\mathrm{termi}} \cdot \sum_{i=1}^{4} \sum_{k=1}^{K} (\alpha_i w_{i,k})^2 / \sum_{k=1}^{K} w^2_{q,k}$.

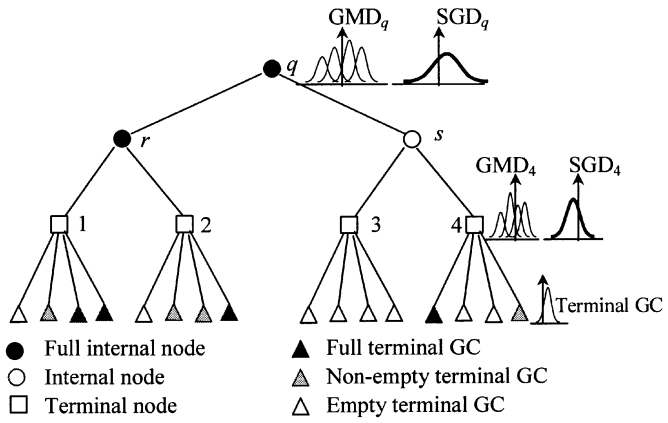

Fig. 3. Illustration of the types of nodes and the GCs in a CPDT.

The parameters $\left(e_{\text{termi}}^{(\text{GMD})}, s_{\text{termi}}^{(\text{GMD})^2}\right)$ can be estimated from the biases of GCs of terminal nodes computed from adaptation data. Denote T_q as the number of terminal nodes under the node q and assume that the contributions of biases $b_i^{(\text{SGD})}$ $i = 1, 2, \ldots, T_q$ to the bias $b_q^{(\text{SGD})}$ are approximately equal, i.e., $\alpha_i = 1/T_q$. Based on the above derivations and incorporating the dimension variable d of feature components into (7) and (8) these equations become

$$\begin{aligned} E[AL(X_q|\lambda_q)] = &-\frac{1}{2}\left[\sum_{k=1}^{K} w_{q,k} \sum_{d=1}^{D} \frac{E(b_{q,k,d}^2)}{\sigma_{q,k,d}^2} \right. \\ &\left. + \sum_{k=1}^{K} w_{q,k} \sum_{d=1}^{D} \ln\left(2\pi\sigma_{q,k,d}^2\right)\right] \\ &- \frac{1}{2} \cdot \sum_{d=1}^{D} \text{const}_d \\ &+ \sum_{k=1}^{K} w_{q,k} \ln(w_{q,k}) \end{aligned} \tag{9}$$

$$\begin{aligned} E\left(b_{q,k,d}^2\right) = &e_{q,d}^{(\text{GMD})^2} + s_{q,d}^{(\text{GMD})^2}, \\ &k = 1, \ldots, K \;\; d = 1, \ldots, D \end{aligned} \tag{10}$$

where $e_{q,d}^{(\text{GMD})} \approx e_{\text{termi},d}^{(\text{GMD})}$, and $s_{q,d}^{(\text{GMD})^2} \approx s_{\text{termi},d}^{(\text{GMD})^2} \cdot \sum_{i=1}^{T_q}\sum_{k=1}^{K}(w_{i,k}/T_q)^2 / \sum_{k=1}^{K} w_{q,k}^2$.

The value of the 'const_d' in (9) does not affect model selection result, and can be set to any real number. Details are discussed in Section III-E.

D. Parameter Estimation of the Bias Distribution

Referring to Fig. 3, for a subtree rooted at an internal node q that represents a tied allophone state, the biases corresponding to the terminal GCs are assumed i.i.d. Gaussian r.v.'s, the computed biases are samples of the distribution, and a bias distribution can therefore be estimated for a full internal node defined in Section III-C. In computing the expected log-likelihood for an internal node q, if it is a full internal node, then (9) and (10) are applied directly to obtain the EL_q; otherwise, the bias distribution under the node is approximated by that of its nearest ancestor full node.

In order to generate reliable bias samples, two cases are considered. In the first case, a terminal GC has been assigned sufficient adaptation data, and a bias sample can be directly computed for the GC. In the second case, a nonempty GC has a data amount below the specified threshold, and a bias sample is computed as a weighted average of biases of several nearby nonempty GCs and the weights are made proportional to the data amount in each of these GCs.

E. MEL Approach for Model Selection

Denote the expectation of log-likelihood for a tree cut F by $\text{EL}(F)$. MEL based model selection attempts to determine an optimal tree cut F^* that maximizes $\text{EL}(F)$, i.e., $F^* = \arg\max_F[\text{EL}(F)]$. $\text{EL}(F)$ can be defined as a weighted summation of expected log-likelihood of all nodes in the tree cut F, i.e., $\text{EL}(F) = \sum_{i\in F} T_i \cdot \text{EL}_i$, where T_i is the number of terminal nodes under node i, and for any tree cut Q, we have $\sum_{i\in Q} T_i = T_{\text{root}}$ which is the total number of terminal nodes in the tree.

The optimal tree cut selection can be efficiently accomplished by a bottom-up tree pruning algorithm. The algorithm is illustrated in Fig. 4. For an internal node p, the difference between EL_p and the sum of its two children's MELs is defined as

$$\Delta\text{EL}(p, l, r) = [T_l \cdot \text{MEL}_l + T_r \cdot \text{MEL}_r - T_p \cdot \text{EL}_p] \tag{11}$$

where $T_p = T_l + T_r$, and the MEL value of the node p is assigned as (see (12) at the bottom of the next page).

If $\Delta\text{EL}(p, l, r) \leq 0$, then the children nodes of the node p are pruned, otherwise they are kept. From (9), (11), and (12), we can see that the constant term "$-(1/2) \cdot \sum_{d=1}^{D} \text{const}_d$" in (9) is eliminated in (11) and it does not affect the model selection result.

This procedure is carried out bottom-up over all the nodes of a clustered phonetic decision tree, similar to the method of [21]. After tree pruning, the collection of terminal nodes constitute the optimal tree cut. In implementation, the tree pruning procedure is constrained so that for each state of each phone, at least the root node of the PDT is maintained. Moreover, although in a data sparse situation it is possible to estimate node specific *const*'s at high level nodes of the CPDT, such as estimating a specific *const* vector at the root node of each PDT, the model selection results would remain the same as the case of using a global *const* vector since tree pruning is performed at the lower level tree nodes, in general.

F. Dynamic Model Selection

Given an amount of adaptation data from a speaker, model parameters can be first adapted to reduce the mismatch between the speaker's speech and the adapted model. As the amount of adaptation data increases, model parameters are better adapted and the mismatch biases in general become smaller. Consequently, the optimal model structure should change with the amount of adaptation data. To dynamically select the optimal model, it is desirable to perform model selection after an initial model adaptation. Note that from (9) and (10), the mean parameters of GCs of internal nodes are not involved in model selec-

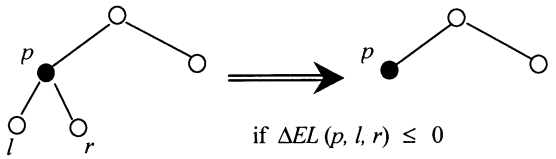

Fig. 4. MEL based tree pruning.

tion, and therefore in initial model adaptation by MLLR, only the mean parameters of terminal GCs need to be adapted.

In order to avoid the effect of over-fitting, the "validation data" used for model selection should be independent of the "training data" used in initial model adaptation. Ideally, if sufficient data are available, then we can divide them to two disjoint sets, one set for model adaptation (*training*), and the other set for model selection (*validation*). This is feasible here since the MEL based model selection can perform well with a relatively small amount of data, and the deduction of this amount of data from initial model adaptation would not drastically change the adaptation performance if the amount of adaptation data is adequate. On the other hand, if the amount of adaptation data is very small, then the data for initial model adaptation and model selection are allowed to overlap. The details of such data partition are discussed in Section IV.

Once model selection is performed, the entire set of adaptation data can be used to perform model adaptation on the selected model. Therefore MLLR adaptation is performed twice, once before model selection and once after it. The complete MEL based dynamic model selection/adaptation algorithm is implemented in seven steps as shown in Fig. 5. The overall procedure is referred to as the MEL based method in Section IV of experiments.

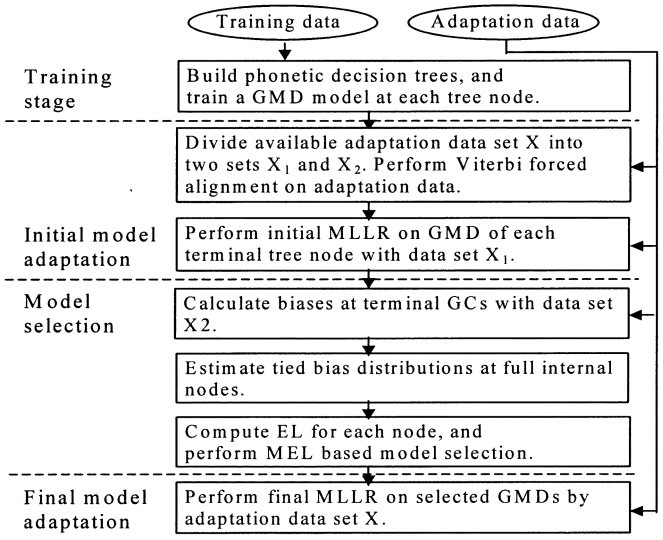

Fig. 5. Procedure of model selection and adaptation.

IV. Experiments

The proposed method was evaluated on the LDC WSJ1 database. The entire set of speaker-independent short-term training data (SI_TR_S, 200 speakers) of WSJ1 was used for acoustic model training. Each triphone HMM model had three emitting states (except for a "short-pause" model, which had a single state), and each state had a mixture of 16 Gaussian densities. Based on the consideration that cross-word transitions in nonnative speech is in general not as smooth as native speech, and based on our previous experience that simpler context-dependent phone models worked better for Chinese accent English speech recognition [17], only within-word triphones were used. An additional benefit of using within-word triphones is the simplification of search at the decoding stage. Speech features consisted of 39 components of 12 MFCCs, energy, and their delta and acceleration derivatives. Cepstral Mean Normalization (CMN) as implemented in HTK was applied to both training and test data. In testing, the standard 5K-vocabulary bigram language model provided by WSJ1 was used, and the decoder was provided by HTK v2.2 [22]. The silence model was not adapted.

A. Baseline System

A baseline system was built as follows. The acoustic model complexity was determined to provide best performance for recognition of native speech. First, a very large clustered phonetic decision tree was built as described in Section III-A. The CPDT was considered as a basic tree with sufficient details, and any reasonable model candidate would correspond to a tree cut in this basic tree. An existing speech recognition system was used to segment the speech data of the entire training set to the state level of phone units by a forced Viterbi alignment. The segmented data were assigned to each node of the CPDT according to the corresponding states of triphones, and EM algorithm was employed to estimate parameters of GMD at each tree node. As such, once a tree cut is determined, GMDs corresponding to that tree cut can be used to form an acoustic model for speech recognition.

Bayesian Information Criterion (BIC) [23] was used to select the optimal baseline model, which is defined as: $\mathrm{BIC}(X,\lambda) = ln\, p(X|\lambda) - \gamma/2M\, ln(N)$, where λ is a model, X is the training data set, M is the total number of free parameters in the model, N is the size of X, and γ is a tuning parameter that can be adjusted to balance likelihood of observations with model complexity [24]. For a GMD with m D-dimensional diagonal Gaussian components, $M = (2D+1)\cdot m$.

For each specified γ, a tree cut that led to the maximum BIC value was determined. By varying the value of γ, a series of BIC optimal models were selected. These models were validated on the WSJ1 test set ET_H2, which consisted of 10 native speakers with totally 215 test utterances. The recognition performance

$$\mathrm{MEL}_p = \begin{cases} \frac{1}{T_p}(T_l \cdot \mathrm{MEL}_l + T_r \cdot \mathrm{MEL}_r), & \text{if } \Delta\mathrm{EL}(p,l,r) > 0. \\ \mathrm{EL}_p, & \text{if } p \text{ is a terminal node, or } \Delta\mathrm{EL}(p,l,r) \le 0. \end{cases} \tag{12}$$

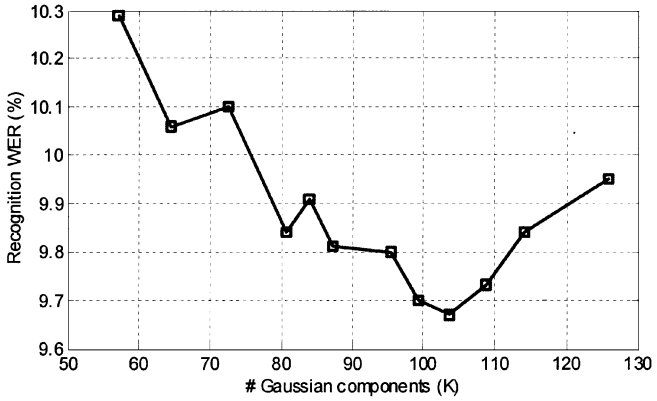

Fig. 6. Recognition WER versus model complexity for native speech.

versus model complexity is shown in Fig. 6. The best performance with a WER of 9.7% was achieved by a model with 103K GCs, corresponding to a tree cut with 6473 nodes. This model was chosen as the baseline model in the subsequent tests.

B. Experimental Condition

Test was conducted on speakers with different foreign accents and different levels of English speaking proficiency. A total of 32 speakers were included in the test set. WSJ1 database provides two groups of nonnative speakers (DT_S3 and ET_S3) and one group of native speakers (ET_H2). Each group has ten speakers. In addition, speech data of two speakers with Mandarin Chinese accent (chn1 and chn2) were collected under a similar acoustic condition and with similar prompting texts as WSJ1. A total of 40 adaptation utterances were available from each test speaker. As show in Table I, these 32 speakers were divided into four groups based on their English speaking proficiency as measured by baseline recognition error rate, where G1 is the nonnative speaker group with the highest recognition error rate, and G4 is the native speaker group with the lowest recognition error rate.

In testing, the decoding parameters, including language model score scale and beam-search pruning thresholds, were optimized for native speaker group ET_H2 and were applied to all the four groups. For each test speaker, N adaptation utterances were randomly selected from his or her adaptation data set for use as adaptation data, where $N = 1$, 3, 5, 10, 20, 40, and the first 20 test utterances were used in testing (except for the ET_H2 group, where each speaker had only about 20–23 test utterances and therefore all the test utterances were used). The adaptation experiments were repeated three times with different selection of adaptation utterances, except that when $N = 40$, the whole set of adaptation data was used once. Recognition results were averaged over each group. In MLLR implementation, the CPDT was used as a regression tree, and only mean vectors of Gaussian components were adapted. The sample size threshold for estimating a MLLR transformation was set to 500, and only mean vectors of Gaussian components were adapted because the performance gain from variance adaptation is usually small compared with that of mean adaptation [25]. In final model adaptation, for estimation reliability, only a diagonal transformation with a bias vector was estimated when $N = 1$. For MEL based model selection, the threshold on the number of biases for a full node was set to 25, and the threshold on the number of feature data for a full terminal GC was set to 30.

TABLE I
SPEAKER GROUPS DEFINED BY BASELINE RECOGNITION ERROR RATE (%)

Group ID	G1	G2	G3	G4
proficiency	worst	bad	good	best
Speakers [a]	4n0,1,3,4, chn1,chn2	ET_S3 (4nd~4nn)	4n5,8, 9,a,b,c	ET_H2 (4oa~4oj)
Avg. WER	60.8	26.5	18.1	9.7
S.D. WER	3.8	5.9	7.0	5.9

[a] 4n0, 1, 3, 4, 5, 8, 9, a, b, c belong to DT_S3

In MEL, the partition of adaptation utterances into two subsets, one for initial model adaptation and one for model selection, was empirically determined for different N. When the adaptation data were 20 utterances or more, two disjoint subsets were generated, each had half the adaptation data. When adaptation data amount was less than 20 utterances, all utterances were used in model selection, and a subset of them was used in initial model adaptation. Specifically, in the $N = 1$ case, initial adaptation was not performed; in the $N = 3$ case, the first utterance was used for initial adaptation to estimate a global bias-only transformation; in the $N = 5$ case, the first two utterances were used for initial adaptation to estimate a global diagonal MLLR transformation; in the $N = 10$ case, the first five utterances were used to estimate a global full MLLR transformation. For $N = 20$ or 40, half amount of adaptation utterances were used to estimate full MLLR transformations with the number of transforms determined by the threshold discussed above.

C. MEL Based Model Selection and Adaptation

The basic CPDT tree used in MEL model selection had a total of 7137 terminal nodes. The evaluation conditions included model adaptation by conventional MLLR alone and by the proposed MEL based method. These results are summarized by recognition word error rate (WER) in Fig. 7(a)–(d) for the four speaker groups G1–G4. Recognition results show that MEL based model selection produced a significant impact on recognition performance for heavy foreign accent speakers. This verified the notion that a detailed model that is optimal for native speakers is not suitable for heavy accent speakers. Instead, a less complex model structure can better tolerate distribution deviations of nonnative speech from native speech. The effect of error reduction due to MEL based model selection is observed to reduce with the increase of English speaking proficiency. The MEL based model selection did not improve recognition for native speakers, due to the fact that the model complexity of the baseline system was optimized for the native speakers.

In Table II, the number of remained states after model selection is shown for each group. It is worth noting that only half of adaptation data were used in model selection when the number of adaptation utterances $N \geq 20$. We can observe that for speakers with heavy foreign accents, a simpler model structure was selected than that for speakers with slight accents. On the other hand, as more adaptation data became available, more complex models were selected for each group of speakers. The

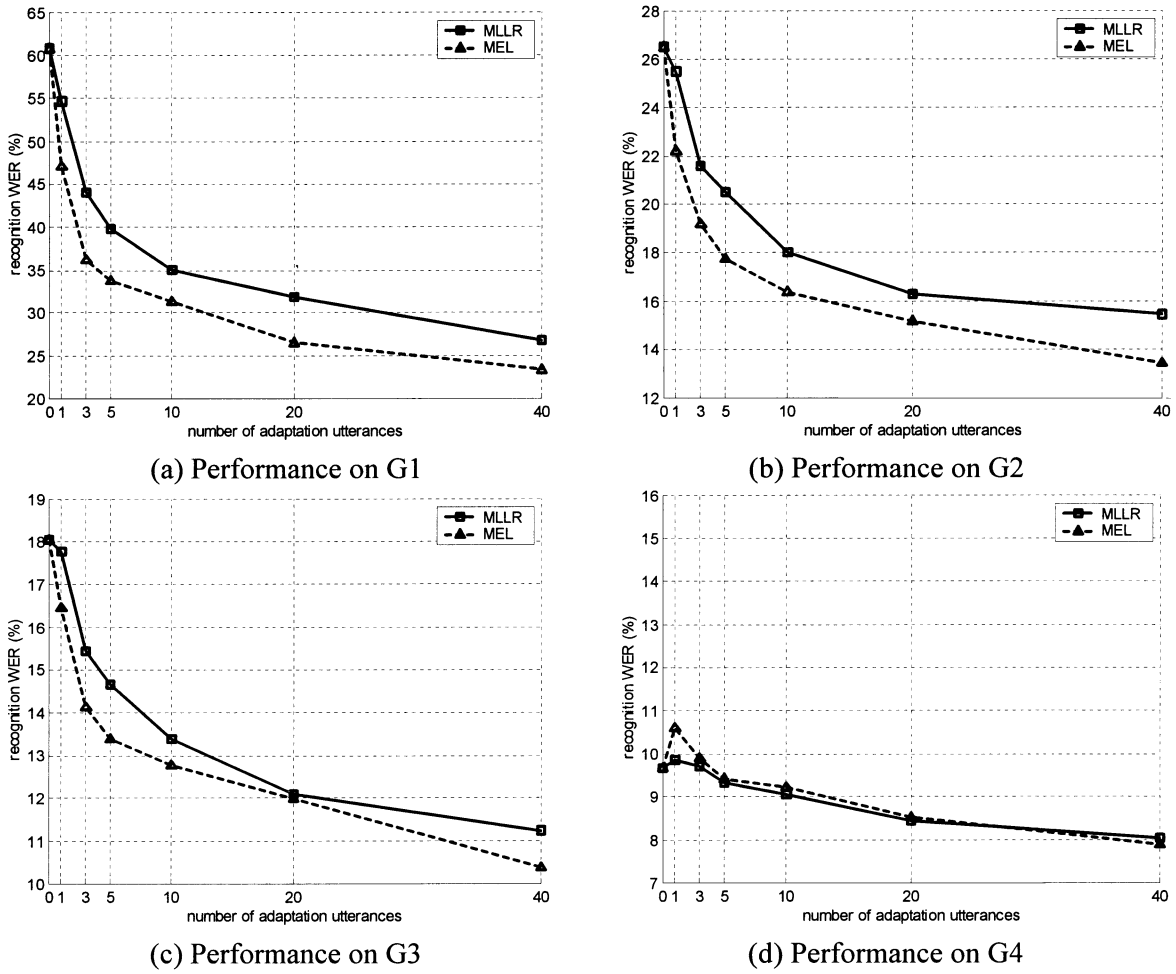

Fig. 7. Recognition WER versus amount of adaptation data for speaker groups defined by English speaking proficiency.

TABLE II
NUMBER OF TIED STATES RESULTING FROM MODEL SELECTION

# adaptation utterances	1	3	5	10	20	40
G1	1994	1793	2101	2688	2966	3221
G2	2581	2048	2395	2978	3074	3397
G3	2462	2137	2457	3310	3409	3569
G4	2616	2701	3043	4051	4213	4410

proposed MEL algorithm was able to capture this information and dynamically select more complex model structures with increasing amounts of adaptation data. It is also worth noting that, even for native speakers, the MEL selected model was less complex than the baseline one without causing performance degradation. This is because that the baseline system was a speaker independent one and it had certain redundancy with regard to individual native speakers.

In implementing MLLR for final model adaptation, two approaches were investigated. In the first approach, the GCs of the nodes in the optimal tree cut were treated as terminal GCs, and MLLR model adaptation was carried out with respect to those GCs. In the second approach, at the model selection stage a node that was to be pruned was first marked to keep the original structure of the CPDT temporarily, then at the final model adaptation stage each MLLR transformation was estimated with respect to terminal GCs in the original CPDT, and the estimated transformation was used to adapt the GCs of the selected optimal tree cut. The marked nodes were pruned after model adaptation, and the adapted GMDs corresponding to the optimal tree cut were used to form a final acoustic model for speech recognition.

Fig. 8 illustrates these two approaches for final model adaptation. A hypothetic optimal tree cut F^* and its constituent nodes r and s are illustrated in Fig. 8. Assuming that adequate amounts of data are accumulated at nodes q and r, then for both approaches of model adaptation, two MLLR transformations, W_q and W_r, are estimated at nodes q and r, with W_q used for the GCs at node s and W_r for the GCs at node r. The difference between these two approaches lies in the estimation of the MLLR transformations. For example, in the first approach, W_q is estimated based on the GCs at nodes r and s and their corresponding data, and as a contrast, in the second approach, W_q is estimated based on the GCs at nodes 1–4 and their corresponding adaptation data.

We found that the latter approach had a small but consistent advantage over the first approach. The results presented here were therefore based on the latter approach.

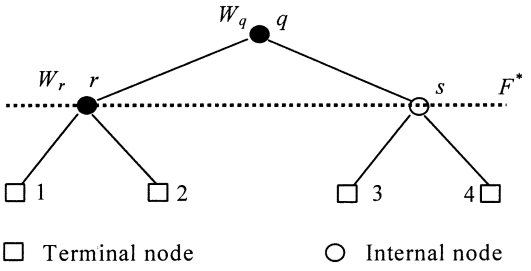

Fig. 8. Illustration of final MLLR model adaptation.

D. Robustness of MEL Based Model Selection

The robustness of MEL based model selection with respect to the amount of adaptation data used in model selection was also evaluated. Same adaptation data sets as used in Section IV-D were used here. To isolate model selection from model adaptation, initial model adaptation was not performed and all data in one adaptation set were used for model selection. Given a fixed number of adaptation utterances, the resulting numbers of tied allophone states for individual speakers in each group were averaged. Model selection was also evaluated on speakers in the training set, where ten speakers (460–469) were selected to form a group G5, and the adaptation set conditions of G5 were the same as those of the other groups.

As shown in Fig. 9, when the data amount was very small, the model selection results were inconsistent, but the ordering among the five groups was still reasonable, and when the data amounts were ten utterances or more, the selected model complexities became stabilized. This indicates that the proposed model selection technique worked well with a small amount of data. The robustness can be attributed to the fact that in the MEL based model selection, only the expectation of the squared bias needs to be estimated for each tied allophone cluster, and this term was effective in capturing the degree of mismatch between the adaptation data and the acoustic model.

Also shown in Fig. 9 is that the selected model complexity increased with the degree of matching between the adaptation data and the model. Among the four test groups G1, G2, G3, G4, for group G1 of the heaviest foreign accents the simplest model was chosen, and for group G4 of native speakers a complex model was chosen. Furthermore, since group G5 had 10 training set native speakers and therefore matched the model best, the corresponding model complexity was the highest, higher than that of the test group G4.

E. Expected Log-Likelihood versus Log-Likelihood

The agreement between the expected log-likelihood value and the log-likelihood value as computed from speech data was evaluated. All 32 speakers of the four groups G1–G4 were included. For each test speaker, the average log-likelihood of 40 adaptation utterances was computed based on the baseline model. From the same set of adaptation data, as in model selection, bias distributions were estimated and the expected log-likelihood corresponding to the selected tree cut of the baseline model was computed and normalized. The results are shown as a scatter plot in Fig. 10, and the correlation coefficient between the normalized EL and the average log-likelihood over the 32 speakers is 0.81, indicating a good agreement between EL and log-likelihood.

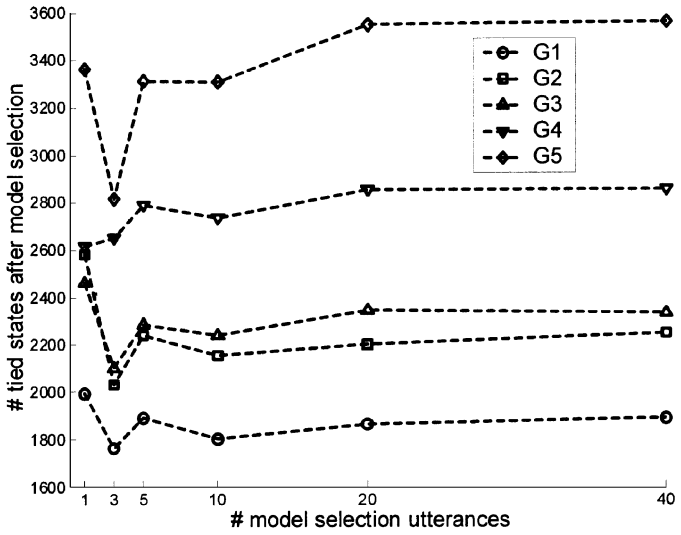

Fig. 9. Model complexity versus amount of data by using MEL based model selection.

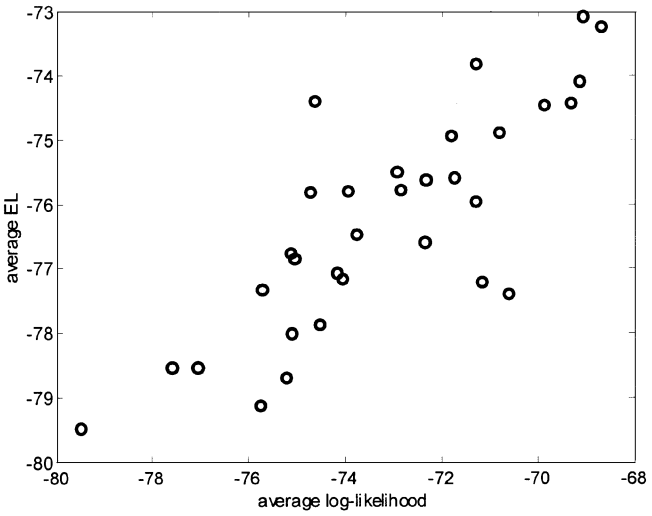

Fig. 10. Scatter plot of expected log-likelihood versus log-likelihood.

V. Conclusion

Although the performance of native speech recognition is not very sensitive to model complexity upon reaching a certain limit, highly detailed acoustic models that are trained from native speech are not suitable for nonnative speech recognition due to various deviation and variation factors in nonnative speech with respect to native speech. An acoustic model with a proper level of complexity is desirable to balance the needs for discrimination of speech sounds and for tolerance of variations in nonnative speech. In this paper, a novel technique of model complexity selection is proposed to select an optimal tree cut based on maximization of expected likelihood, and model adaptation technique of MLLR is integrated with the MEL based model selection to allow dynamic selection of model complexity and full usage of adaptation data. On

nonnative English speech, the proposed model complexity selection method led to consistent and significant improvements to MLLR, while for native English, speech recognition performance similar to MLLR was maintained.

The proposed MEL based model selection technique needs to be further improved to work more reliably when the data amount is small. Additional performance improvements are also possible by integrating the proposed model selection/adaptation framework with Bayesian speaker adaptation techniques. Furthermore, it is of interest to study the relation of optimal model structure resulting from model selection with different types of foreign accents.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable suggestions that helped make this paper better.

REFERENCES

[1] G. Zavaliakos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, 1995, pp. 676–679.
[2] G. Zavaliakos, "Maximum a posteriori adaptation for large scale HMM recognizers," in *Proc. ICASSP*, 1996, pp. 725–728.
[3] S. Witt and S. Young, "Off-line acoustic modeling of nonnative accents," in *Proc. EUROSPEECH*, 1999, pp. 1367–1370.
[4] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.
[5] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
[6] U. Uelber and M. Boros, "Recognition of nonnative german speech with multilingual recognizers," in *Proc. EUROSPEECH*, 1999, pp. 911–914.
[7] V. Fischer, E. Janke, S. Kunzmann, and T. Ross, "Multilingual acoustic models for the recognition of nonnative speech," in *Proc. Automatic Speech Recognition and Understanding Workshop*, 2001.
[8] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP*, 1996, pp. 2195–2198.
[9] F. Weng *et al.*, "A study of multilingual speech recognition," in *Proc. EUROSPEECH*, 1997.
[10] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
[11] W. Byrne *et al.*, "Toward language independent acoustic modeling," in *Proc. ICASSP*, 2000, pp. 1029–1032.
[12] V. Diakoloukas and V. Digalakis, "Maximum likelihood stochastic transformation adaptation of hidden markov models," *IEEE Trans. Speech Audio Processing*, pp. 177–187, March 1999.
[13] C. Boulis and V. Digalakis, "Fast speaker adaptation of large vocabulary continuous density HMM speech recognizer using a basis transform approach," in *Proc. ICASSP*, vol. 2, 2000, pp. 989–992.
[14] D. Compernolle, "Recognizing speech of goats, wolves, sheep and … nonnatives," *Speech Commun.*, vol. 35, pp. 71–79, 2001.
[15] S. J. V. Wijingaarden, "Intelligibility of native and nonnative dutch speech," *Speech Commun.*, vol. 35, pp. 103–113, 2001.
[16] K. Berkling, "SCoPE, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification," *Speech Commun.*, vol. 35, pp. 125–138, 2001.
[17] X. He and Y. Zhao, "Model complexity optimization for nonnative English speakers," in *Proc. EUROSPEECH*, Scandinavia, Denmark, Sept. 2001, pp. 1461–1464.
[18] ——, "Fast model adaptation and complexity selection for nonnative speakers," in *Proc. ICASSP*, vol. 1, 2002, pp. 577–580.
[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.
[20] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree based state tying for high accuracy modeling," in Proc. ARPA Workshop on Human Language Technology, Mar. 1994.
[21] S. Wang and Y. Zhao, "Online bayesian tree structure transformation of HMM's with optimal model selection for speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 663–677, Sept. 2001.
[22] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK Book, Version 2.2. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml
[23] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[24] S. Deligne, E. Eide, R. Gopinath, D. Kanevsky, B. Maison, P. Olsen, H. Printz, and J. Sedivy, "Low-resource speech recognition of 500-word vocabularies," in *Proc. EUROSPEECH*, 2001.
[25] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, pp. 249–264, 1996.

Xiaodong He received the B. Eng. degree in precision instruments from Tsinghua University, Beijing, China, in 1996, and the M.S. degree in signal and information processing from Chinese Academy of Sciences, Beijing, in 1999. He is currently pursuing the Ph.D. degree in the Department of Computer Engineering and Computer Science, University of Missouri-Columbia.

His research interest lies in the field of speech and signal processing, pattern recognition, spoken language processing, large vocabulary speech recognition, acoustic modeling, speaker adaptation, and nonnative speech recognition.

Mr. He is a member of Sigma Xi.

Yunxin Zhao (S'86–M'88–SM'94) received the Ph.D degree in 1988 from the University of Washington, Seattle.

She was Senior Research Staff and Project Leader with the Speech Technology Laboratory, Panasonic Technologies, Inc., from 1988 to 1994. She was Assistant Professor with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign from 1994 to 1998. She is currently Professor with the Department of Computer Engineering and Computer Science, University of Missouri-Columbia. Her research interests are in spoken language processing, automatic speech recognition, multimedia interface, multimodal human–computer interaction, statistical pattern recognition, statistical blind systems identification and estimation, speech and signal processing, and biomedical applications.

Dr. Zhao was Associate Editor of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and a member of IEEE Speech Technical Committee. She received 1995 NSF Career Award, and is listed in American Men and Women of Science, February 1998.