ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT

Computer Speech and Language xxx (2006) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Prior knowledge guided maximum expected likelihood based model selection and adaptation for nonnative speech recognition

Xiaodong He [a,*,1], Yunxin Zhao [b]

[a] Natural Language Processing Group, Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
[b] Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65211, USA

Received 10 October 2005; received in revised form 27 May 2006; accepted 4 June 2006

Abstract

In this paper, an improved method of model complexity selection for nonnative speech recognition is proposed by using maximum *a posteriori* (MAP) estimation of bias distributions. An algorithm is described for estimating hyper-parameters of the priors of the bias distributions, and an automatic accent classification algorithm is also proposed for integration with dynamic model selection and adaptation. Experiments were performed on the WSJ1 task with American English speech, British accented speech, and mandarin Chinese accented speech. Results show that the use of prior knowledge of accents enabled more reliable estimation of bias distributions with very small amounts of adaptation speech, or without adaptation speech. Recognition results show that the new approach is superior to the previous maximum expected likelihood (MEL) method, especially when adaptation data are very limited.

Keywords: Nonnative speech recognition; Model selection; Maximum expected likelihood; Accent classification; Maximum *a posteriori* estimation

1. Introduction

American English speech recognition systems are commonly trained from speech data of native American English speakers. For native American English speakers, these systems may work very well within constrained task domains, but for speakers with heavy foreign accents, the performances deteriorate dramatically in general. The difficulty in foreign accented speech can be largely attributed to confusions of vowels, especially those not in a speaker's mother tongue (Compernolle, 2001; Wijngaarden, 2001), and such phone variation and substitution change greatly with different types and levels of foreign accents as well as phone contexts (Berkling, 2001). Although foreign accent dependent acoustic models may best capture properties of nonnative speech

[*] Corresponding author. Tel.: +1 425 706 4939; fax: +1 425 936 7329.
E-mail addresses: xiaohe@microsoft.com (X. He), zhaoy@missouri.edu (Y. Zhao).

[1] The work was done when the author was with Department of Computer Science, University of Missouri-Columbia.

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.06.001

and therefore are the most accurate for nonnative speech recognition, it remains difficult to train such models since the required vast amounts of training data that cover different types and degrees of foreign accents do not yet exist. It is therefore important to develop effective techniques to improve robustness of state-of-the-art speech recognition systems for nonnative speech recognition.

Increased research activities on nonnative speech recognition have been reported in recent years. These approaches include multilingual acoustic modeling, nonnative speech oriented lexicon modeling, acoustic modeling, and decoding. Although speaker adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) or Maximum *a posteriori* (MAP) estimation (Gauvain and Lee, 1994) are generally applicable to adapt speaker-independent models to foreign-accented speakers, it is well recognized that a large amount of adaptation data is needed from a foreign-accented speaker to achieve an acceptable level of recognition accuracy (Zavaliakos et al., 1995; Zavaliakos, 1996). In multilingual acoustic modeling, phone sets of several languages are mapped to a universal phone set and speech data of these languages are pooled to train an acoustic model (Uelber and Boros, 1999; Fischer et al., 2001; Kohler, 1996; Weng et al., 1997; Schultz and Waibel, 1998; Byrne et al., 2000). In the reported studies, multilingual models were limited to small tasks. Compared with using acoustic models trained from native speech alone, multilingual models improved nonnative speech recognition at the expense of degrading native speech recognition. In Witt and Young (1999), acoustic models of both native English and native foreign language were built based on known target foreign accent, and phone-state mappings between acoustic models of the two languages were learned for recognition of Spanish and Japanese accented English. In Tomokiyo (2000), data-driven lexical modeling of Japanese-accented English was made by generating pronunciation variants of English words from a large amount of Japanese accented English. In Matsunaga et al. (2003), decoding of Japanese-accented English was performed by using a ‘‘bilingual’’ English pronunciation lexicon and using both English and Japanese acoustic models, where each English word has separate transcriptions in English phonemes and Japanese phonemes. In Minematsu et al. (2002), acoustic models were trained from an English speech corpus of Japanese talkers and English speaking proficiency was estimated for each talker to dynamically select proficiency-level appropriate regression trees for MLLR adaptation. The above methods improved performance of nonnative speech recognition to different degrees. However, the scopes of their applications are limited since in these methods, target accent needs to be known *a priori* and large amounts of target accent speech data are required.

The authors of the current paper proposed a model selection based technique of nonnative speaker adaptation (He and Zhao, 2001; He and Zhao, 2003) that overcomes difficulties of the above discussed methods, i.e., it works for arbitrary type of accented speech, and it requires only a moderate amount of adaptation data from a speaker. This technique was evaluated on the Wall Street Journal task and showed significant word error reductions for a wide variety of foreign accented speakers. This technique was motivated from an empirical observation that between native American English and nonnative speakers, the curves of model complexity versus recognition accuracy, and hence the optimum model complexities, were significantly different, indicating that adaptive selection of model complexity would be desirable for foreign accented talkers, and an effective model complexity selection method is needed.

Although information theoretic criteria, such as BIC, MDL and AIC, are widely used for model complexity selection, these criteria are commonly applied at the stage of model training to counterbalance over-fitted likelihood scores by penalizing on the large number of free parameters of the complex model, where in the training stage, the same data set is used for both model parameter estimation and model selection (Schwarz, 1978; Rissanen, 1984; and Akaike, 1974). For the task of recognizing nonnative speech with acoustic model trained from native American English speech data, the adaptation data used for adaptive model selection is independent of the training data used for model parameter estimation. Therefore, the above discussed information theoretic criteria cannot be directly applied. On the other hand, the well known Kullback–Leibler measure (Kullback, 1959) characterizes distribution mismatch and can therefore be used for model selection. However, there is no known closed-form expression of Kullback–Leibler measure between two Gaussian mixture densities. Moreover, for the current task of adaptive model complexity selection, the amount of adaptation data is too limited to estimate adaptation data distributions for many triphone states, making the Kullback–Leibler measure for these states infeasible.

In He and Zhao (2003), model selection by using a small amount of adaptation speech was accomplished by a maximum expected likelihood (MEL) algorithm (details are described in Section 2). The expected likelihood score is a function of the second moment of biases that describe mismatches between nonnative adaptation speech data and native-speech trained acoustic model at individual nodes of phonetic decision trees. Therefore, the key to the MEL algorithm lies in the estimation of distributions of biases, referred to as bias distributions, which directly impacts the selection of model complexity and thus the accuracy of nonnative speech recognition. When adaptation data is sparse, bias distribution estimation may be unreliable. While reliability can be improved by distribution sharing, details are in general compromised as the cost (He and Zhao, 2003).

The current work aims at enhancing the MEL based model selection technique by further improving the estimation of bias distributions in data sparse conditions. A novel method of integrating prior knowledge of foreign accents into bias distribution estimation is proposed within the framework of maximum a posteriori estimation. When the amount of adaptation data is small, estimation of bias distributions can depend more on prior knowledge than data so that a fine-level sharing of bias distributions is maintained without sacrificing reliability. To apply accent-specific prior knowledge automatically for each speaker, an accent classification algorithm is further developed. The proposed new approach, called P-MEL, has been implemented and evaluated on the WSJ task with speech data of American English speakers, British English speakers and mandarin Chinese accented English speakers. For each of the three ‘‘accents,’’ a small set of speech data was used to estimate the priors of the bias distributions. Recognition evaluation test was performed on the 5000-word WSJ task. Experimental results verified that the P-MEL approach is superior to the MEL approach in model complexity selection that leads to reduced recognition word error rate when the amount of adaptation data is very small.

This paper is organized as follows. In Section 2, the MEL based model selection technique is provided as background material. In Section 3, the proposed methods of prior knowledge based estimation of bias distributions and automatic classification of foreign accents are developed. Experimental results are presented in Section 4, and a conclusion is drawn in Section 5.

2. Background of MEL based model complexity selection

To facilitate understanding of the proposed P-MEL technique, the MEL technique as described in He and Zhao (2003) is summarized in this section, including the concepts and implementation procedures of MEL based model selection and adaptation.

2.1. Model complexity selection and expected likelihood

In state-of-the-art hidden Markov modeling (HMM) of speech, very sharp distributions are commonly employed to describe narrowly defined acoustic units. Although these models work well for recognition of native speech, less detailed models that are more robust to accent-induced phone variations are better suited for nonnative speech. Optimal model complexity that corresponds to minimum word error rate is very different for native and nonnative speakers (a quantitative evaluation is shown in Fig. 6 of Section 4). Moreover, optimal model complexity may also be different for individual speakers. It is therefore desirable to adaptively determine model complexity for each speaker by using a small amount of adaptation data.

The selection of model complexity can be addressed from the perspective of state tying in phonetic decision trees (PDT). In a PDT, each tree node represents an allophone cluster with data distribution of allophones tied in that node; the collection of distributions over the nodes in a tree cut constitutes an acoustic model for the phone unit state that the PDT represents. As illustrated in Fig. 1, a high-level tree cut corresponds to a less

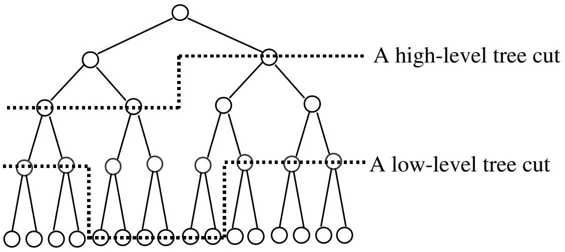

Fig. 1. Model selection in a phonetic decision tree.

detailed model, and a low-level tree cut corresponds to a more detailed model, given fixed distribution complexity at each node (a Gaussian mixture density). A proper tree cut, or equivalently model complexity, should then be selected for each phone state of the acoustic model of a nonnative speaker.

In general, tree cut selection can be based on the likelihood criterion that measures the fit between acoustic model and adaptation data. In speaker adaptation, however, the amount of adaptation data is often limited such that only a small number of PDT nodes may have adaptation data, making the likelihood method unreliable. In order to facilitate model selection with a limited amount of adaptation data, an expected likelihood criterion was proposed in He and Zhao (2003) to convert the problem of computing likelihood of observation data at each tree node into computing expected statistic of data-model mismatch bias. Through a hierarchical sharing of bias distributions, expected bias statistic can be computed for tree nodes with insufficient adaptation data.

Assume that a PDT node is associated with a Gaussian mixture density λ of size K and an adaptation data set $X = \{x_1, x_2, \ldots, x_N\}$, where x_j is a D dimensional vector. The log-likelihood of X given the model λ is then $L(X|\lambda) = \sum_{i=1}^{N} \ln\left[\sum_{k=1}^{K} w_k N(x_i; \mu_k, \Sigma_k)\right]$, where w_k, μ_k and $\Sigma_k = \mathrm{diag}(\sigma_{k,1}^2, \ldots, \sigma_{k,D}^2)$ are the weight, mean vector and covariance matrix of the kth Gaussian component (GC) density, respectively. By applying dominant Gaussian component approximation, i.e, $\sum_{k=1}^{K} w_k N(x_i; \mu_k, \Sigma_k) \approx \max_k w_k N(x_i; \mu_k, \Sigma_k)$, the log-likelihood of X can be approximated as

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

where the set S_k, with size $|S_k| = N_k$, consists of the indices of data vectors that have the kth GC as the dominant component, and $\sum_{k=1}^{K} N_k = N$.

The log likelihood of the data set X can be expressed as a function of data-model mismatch bias. In the dth data dimension, define $b_{k,d} = \bar{x}_{k,d} - \mu_{k,d}$ as the mismatch bias in the kth GC, where $\bar{x}_{k,d}$ is the data sample mean in S_k, and denote the data sample variance in S_k by $v_{k,d}^2$. The log likelihood function can then be expressed as

$$L(X|\lambda) \approx -\frac{1}{2}\sum_{k=1}^{K} N_k \sum_{d=1}^{D}\left[\ln(2\pi\sigma_{k,d}^2) + \frac{v_{k,d}^2}{\sigma_{k,d}^2} + \frac{b_{k,d}^2}{\sigma_{k,d}^2}\right] + \sum_{k=1}^{K} N_k \ln(w_k).$$

Further define the average log likelihood per data sample to be $\mathrm{AL}(X|\lambda) = \frac{1}{N} L(X|\lambda)$. The expectation of $\mathrm{AL}(X|\lambda)$ over the distribution of $b_{k,d}$ is obtained below upon making the mild assumptions of $N_k = N \times w_k$ and $v_{k,d}^2 = \mathrm{const}_d \times \sigma_{k,d}^2$ (He and Zhao, 2003):

$$E[\mathrm{AL}(X|\lambda)] = -\frac{1}{2}\left[\sum_{k=1}^{K} w_k \sum_{d=1}^{D} \frac{E(b_{k,d}^2)}{\sigma_{k,d}^2}\right] + C(\lambda), \tag{2}$$

where $C(\lambda)$ absorbs all the terms independent of the $b_{k,d}$'s. The bias $b_{k,d}$ is assumed to be a Gaussian random variable, and the distribution parameters of $b_{k,d}$ need to be estimated for each PDT node in order to compute $E(b_{k,d}^2)$ for each tree node. Due to the sharing of bias distributions, the expected log likelihood (EL) can be computed for tree nodes without sufficient adaptation data.

2.2. Modeling of bias distributions

As indicated in Eq. (2), the performance of MEL-based model selection depends on the quality of the estimated bias distributions. Estimation of bias distributions and hence model selection could be unreliable when data are very limited. In He and Zhao (2003), a clustering scheme is employed to group similar Gaussian components into allophone clusters with one cluster corresponding to one tree node. In order to reliably estimate bias distributions, a bias distribution is estimated for a tree node only when sufficient samples of bias are accumulated under the tree node, where a bias sample is defined to be a data-model mismatch bias computed with respect to a Gaussian component in a leaf node. As the result, certain tree nodes have bias distributions while others not, and a node without a bias distribution will then use the one from its closest parent node. It is clear that when data is very sparse, only a few large clusters might be generated, and a large cluster may include

GCs with very different properties. In the extreme case, all GCs would be grouped into a single cluster and share a global bias distribution. Since for a foreign accented talker, the degree of data-model mismatch is highly dependent on phones or phone-classes, the accent characteristics could not be adequately modeled by a few, or a global, bias distributions. There is an obvious need for striking a balance between details of the bias distributions and reliability of their parameter estimates in order to achieve good performance in model selection.

2.3. Basic procedure of MEL based model selection

The basic MEL based model selection procedure are summarized in the following three steps. In the first step, an acoustic model as implemented by phonetic decision trees (PDT) for triphone HMMs is trained from native American English speech, where a Gaussian mixture density (GMD) is estimated for each node of a PDT, including tree internal nodes. In the second step, Viterbi alignment is performed on adaptation data and each feature vector is assigned to a dominant Gaussian component density (GC) of a tree leaf node, and for each GC of a tree leaf node having adaptation data, a bias is calculated between the data sample mean and the model mean. For each tree node, the distribution parameters of the biases are estimated based on the assumption that the biases under the node are i.i.d. Gaussian random variables. The expected log-likelihood is then computed for the adaptation data. In the third step, the optimal tree cut, or model complexity, is determined to maximize the expected log-likelihood (EL) over tree cuts by using a bottom-up pruning method similar to Wang and Zhao (2001).

The bottom-up tree pruning algorithm is illustrated in Fig. 2 for an internal tree node p. To determine whether the node p should be made a leaf node (pruning away its children nodes or subtrees) or be kept as an internal node (without pruning), the difference between EL of node p and the weighted sum of its two children's MELs is computed as $\Delta\mathrm{EL}(p,l,r) = [T_l \cdot \mathrm{MEL}_l + T_r \cdot \mathrm{MEL}_r - T_p \cdot \mathrm{EL}_p]$, where T_i is the number of leaf nodes under the node i and $T_p = T_l + T_r$. If $\Delta\mathrm{EL}(p,l,r) \leqslant 0$, then the children nodes of the node p are pruned, otherwise they are kept. The MEL value of the node p is updated as:

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

2.4. Dynamic procedure of MEL model selection and MLLR model adaptation

MEL-based model selection is essentially a method of model complexity adaptation. It can be combined with conventional model parameter adaptation, such as MLLR, to achieve a better performance than either method alone.

Given an amount of adaptation data from a speaker, acoustic model parameters can be first adapted to reduce mismatch between the speaker's speech and the model. As the amount of adaptation data increases, model parameters are better adapted and the mismatch biases become smaller. Consequently, the optimal model structure should change with the amount of adaptation data. To dynamically select the optimal model complexity, it is desirable to perform model selection after an initial model adaptation, and to avoid over-fitting, initial model adaptation and model selection are performed by using separate subsets of adaptation data unless the amount of data is too small to split. Once model selection is performed, the entire set of adaptation data can be used to perform model adaptation on the selected model, and the adapted model is then used by a

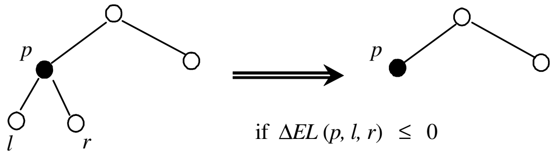

Fig. 2. MEL based tree pruning.

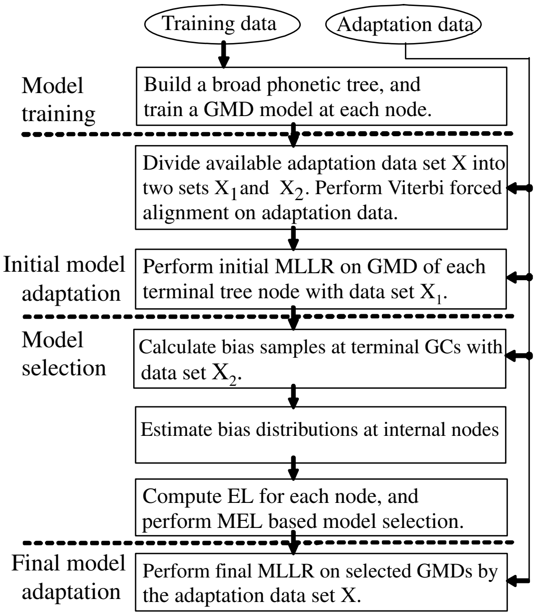
Training data
Adaptation data
Model training
Build a broad phonetic tree, and train a GMD model at each node.
Divide available adaptation data set X into two sets X1 and X2. Perform Viterbi forced alignment on adaptation data.
Initial model adaptation
Perform initial MLLR on GMD of each terminal tree node with data set X1.
Model selection
Calculate bias samples at terminal GCs with data set X2.
Estimate bias distributions at internal nodes
Compute EL for each node, and perform MEL based model selection.
Final model adaptation
Perform final MLLR on selected GMDs by the adaptation data set X.

Fig. 3. Procedure of MEL based dynamic model selection and adaptation.

recognizer for recognition of the speaker's speech. The MEL-based dynamic model selection and adaptation procedure is summarized in the flow chart in Fig. 3 that consists of "Model training", "Initial model adaptation", "Model selection", and "Final model adaptation." In the flow chart, MLLR is shown as the model adaptation method. As will be discussed in Section 4, a cascade MLLR method (Digalakis et al., 1999) is further employed as an alternative model adaptation method to evaluate the effectiveness of the MEL and P-MEL based model selection methods when integrating with different techniques of model parameter adaptation.

3. MAP estimation of bias distributions and accent classification

The proposed enhancements to the MEL-based model selection and adaptation method are covered in this section. The representation of prior knowledge of each accent in the form of prior distributions of bias distribution parameters is described together with the basics of MAP estimation. An accent classification algorithm is developed that enables automatic selection of accent-specific priors of bias distribution parameters for each speaker. The integration of the proposed enhancement technique into the dynamic model selection and adaptation procedure of MEL is discussed last.

3.1. MAP based parameter estimation

An effective approach to acoustic model adaptation with the guide of prior knowledge is maximum *a posteriori* (MAP) estimation (Gauvain and Lee, 1994; Huo et al., 1995; Lee et al., 1991; Zhao, 1996). Given a data set X, while maximum likelihood (ML) estimation obtains an optimal model through

$$\Lambda_{\mathrm{ML}} = \arg\max_{\Lambda} f(X|\Lambda), \tag{3}$$

MAP estimation gives the optimal model by

$$\Lambda_{\mathrm{MAP}} = \arg\max_{\Lambda} f(X|\Lambda)g(\Lambda), \tag{4}$$

where the prior distribution $g(\Lambda)$ characterizes knowledge about the model parameter set Λ. ML and MAP estimations are related through the Bayes' theorem with the posterior distribution $p(\Lambda|X) \propto f(X|\Lambda)g(\Lambda)$. When the size of X is small, $g(\Lambda)$ dominates (4) and the optimal parameter set is estimated mainly base on the prior knowledge; when the size of X is large, $f(X|\Lambda)$ dominates (4) and the optimal parameter set is estimated mainly from the observed data. In MAP estimation, the hyper-parameters of the prior distribution $g(\Lambda)$ need to be obtained in advance.

For the task of modeling a foreign accent, a set of prior distributions can be defined. MAP estimation of Λ for a bias distribution as defined in the MEL method can be solved in different ways, depending on assumptions of $g(\Lambda)$. In the current work, conjugate priors are utilized due to their mathematical property that the posterior and the prior distributions belong to the same distribution family (DeGroot, 1970).

Assume a Gaussian pdf for a bias distribution with the parameters $\Lambda = \{\mu, \theta\}$, where $\theta = 1/\sigma^2$ is the precision parameter, and assume that both mean and variance are random variables. The joint conjugate prior $g(\mu, \theta)$ is then a normal-γ distribution (DeGroot, 1970), where the conditional distribution of μ given θ is a normal distribution with mean v and variance $1/\tau\theta$, and the marginal distribution of θ is a γ distribution with parameters $\alpha > 0$ and $\beta > 0$, i.e.,

$$g(\mu, \theta) = \frac{\sqrt{\tau\theta}}{\sqrt{2\pi}} \exp\left[-\frac{\tau\theta}{2}(\mu - v)^2\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta). \tag{5}$$

Therefore, the joint posterior distribution of $\{\mu, \theta\}$ is also a normal-γ distribution (DeGroot, 1970).

By setting $\alpha = \frac{\tau+1}{2}$, $\beta = \frac{\tau}{2}s^2$, the joint MAP estimate of μ and σ^2 given a set of n sample biases $\{b_i\}$ is solved as (DeGroot, 1970)

$$\hat{\mu}_{\mathrm{MAP}} = \frac{n}{\tau + n} \cdot \bar{b} + \frac{\tau}{\tau + n} \cdot v, \tag{6}$$

$$\hat{\sigma}^2_{\mathrm{MAP}} = \frac{\tau s^2 + nS^2 + \frac{\tau n(\bar{b} - v)^2}{\tau + n}}{\tau + n}, \tag{7}$$

where $\bar{b}$ and S^2 are the sample mean and sample variance of the set, and τ, v and s^2 are the hyper-parameters of $g(\mu, \theta)$.

3.2. Estimation of prior distribution

A proper representation of prior knowledge is important in MAP-based model estimation. For the current task, modeling the priors of bias distributions at the level of phone units appears to be a good choice since phoneme is the basic unit of pronunciation. Although a clustering of allophones at the sub-phone level may characterize more detailed properties of accents, estimating the priors reliably at such a level would require a significant amount of accent-specific training data.

Given speech data of K speakers with a specific accent, hyperparameters of the prior of bias distribution in phone q are estimated by the procedure shown in Fig. 4. The obtained hyper-parameters v_q and s_q^2 (corre-

```
Loop over speakers $k = 1 \sim K$:
    Compute the samples of bias $\{b_{i,k}\}$ in the phone $q$;
    Compute the mean and variance of the sample set, where
        $v_{q,k} = \frac{1}{N_k}\sum_{i=1}^{N_k} b_{i,k}$ , $s_{q,k}^2 = \frac{1}{N_k}\sum_{i=1}^{N_k}(b_{i,k} - e_{q,k})^2$ , and $N_k$ is the size of $\{b_{i,k}\}$;
END Loop

Compute average mean and variance for phone $q$
$v_q = \frac{1}{K}\sum_{k=1}^{K} v_{q,k}$ and $s_q^2 = \frac{1}{K}\sum_{k=1}^{K} s_{q,k}^2$ .

Set the hyper-parameter set of the prior for phone $q$ as $v_q = v_q$, $s_q{}^2 = s_q{}^2$.
```

Fig. 4. Procedure for estimating hyper-parameters of prior distribution in a phone unit q.

sponding to v and s^2 in Eqs. (6) and (spseqn7)) serve as the prior knowledge of the data-model mismatch condition in phone unit q. The hyper-parameter τ is determined empirically to be 15 as in MAP model adaptation (Gauvain and Lee, 1994).

In the current task, 42 English phonemes are defined. Due to limited training data as well as their unbalanced sizes over phones, the priors of certain phones cannot be properly estimated. Therefore, for phone units with inadequate training data, prior distribution sharing is employed at the level of phonetic classes as shown in Table 1, where the class definition is based on Rabiner and Juang (1993). In the experimental data sets, about 20 phone units have insufficient data and their prior distributions were backed off to phonetic classes. In Table 2, a listing of the data-deficient phones together with their corresponding phonetic classes is shown for the Chinese accented speech data set (to be described in Section 4.1), where many of the listed phones are seen to be consonants or semi-vowels.

3.3. Automatic accent detection

For different foreign accents, the priors of bias distribution parameters are likely quite different. This requires knowing the accent of each talker in order to use the proper priors in estimating the posterior bias distributions. On the other hand, due to the influence of mother tongue, speakers with the same foreign accent may consistently pronounce certain phonemes well and certain other phonemes poorly. This accent-specific pronunciation pattern of phonemes is reflected in the data-model mismatch over the defined phoneme set and it can be utilized for automatic accent detection. Although other methods were previously proposed in the literature for accent classification, such as Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) based accent detection using conventional speech features employed in speech recognition (Chen

Table 1
Definition of phonetic classes for prior distribution sharing

Phonetic class	Phonemes
Front vowel	iy ih ae eh
Low-back vowel	aa ax ao ah
High-back vowel	er uh uw
Front diphthong	ey ay
Back diphthong	ow aw oy
Liquid semi-vowel	w wh l
Glide semi-vowel	r y
Voiced stop	b d g
Unvoiced stop	p t k
Voiced fricative	v z dh zh jh
Unvoiced fricative	hh f th s sh ch
Nasal	m n en nx

Table 2
Listing of phones that were backed off to phonetic classes, where the list was generated from the Chinese accented speech data set CH1 with 122 utterances

Phonetic class	Phonemes
Low-back vowel	ao
High-back vowel	uh uw
Back diphthong	aw oy
Liquid semi-vowel	w wh
Glide semi-vowel	y
Voiced stop	b g
Voiced fricative	v dh zh jh
Unvoiced fricative	hh th sh ch
Nasal	en nx

et al., 2001; Teixeira et al., 1996), utilizing bias distributions for accent classification comes as a natural choice in the current work since bias distributions are estimated in P-MEL for model complexity selection. A statistical accent classification algorithm is therefore designed based on this rational to enable automatic selection of priors for each speaker.

The accent models are estimated in a similar way as the priors of bias distributions shown in Fig. 4. The Gaussian densities of triphone states belonging to the same center phone are clustered together. Assuming that biases within a cluster are i.i.d. Gaussian random variables, then for each accent Θ and each phone q, a Gaussian distribution $N(v_{\Theta,q}, s^2_{\Theta,q})$ is estimated, and the set of Gaussian distributions over the phone set are taken as the accent model. Once more, for any phone that has insufficient samples of biases, its bias distribution is backed off to that of its phonetic class.

In the testing stage, a set of biases $B = \{b_i\}$ is first computed from adaptation data. The average log-likelihood of B given an accent Θ is then obtained as:

$$\overline{L}(B|\Theta) = \frac{1}{N}\sum_{q=1}^{Q}\sum_{j=1}^{N_q}\log[N(b_{q,j}|v_{\Theta,q}, s^2_{\Theta,q})], \tag{8}$$

where Q is the number of phones, N is the total number of bias samples, N_q is the number of bias samples in phone q, $b_{q,j}$ is the jth bias sample in phone q. The decision rule for accent classification is

$$\Theta^* = \arg\max_{\Theta}[\overline{L}(B|\Theta) + R(\Theta)], \tag{9}$$

where

$$R(\Theta) = \begin{cases} C_{\text{nat}} & \text{if } \Theta = \Theta_{\text{nat}} \\ 0 & \text{otherwise} \end{cases}. \tag{10}$$

where Θ_{nat} denotes native American English speaker and $C_{\text{nat}} > 0$ is a constant. The incorporation of $R(\Theta)$ into the otherwise maximum likelihood decision rule serves the purpose of reducing the risk of classifying a native American English speaker to a foreign accent speaker, which is motivated by the experimental observation that applying an accent-mismatched prior to a native speaker caused severer error than to a nonnative speaker (see Section 4.2 and Table 7).

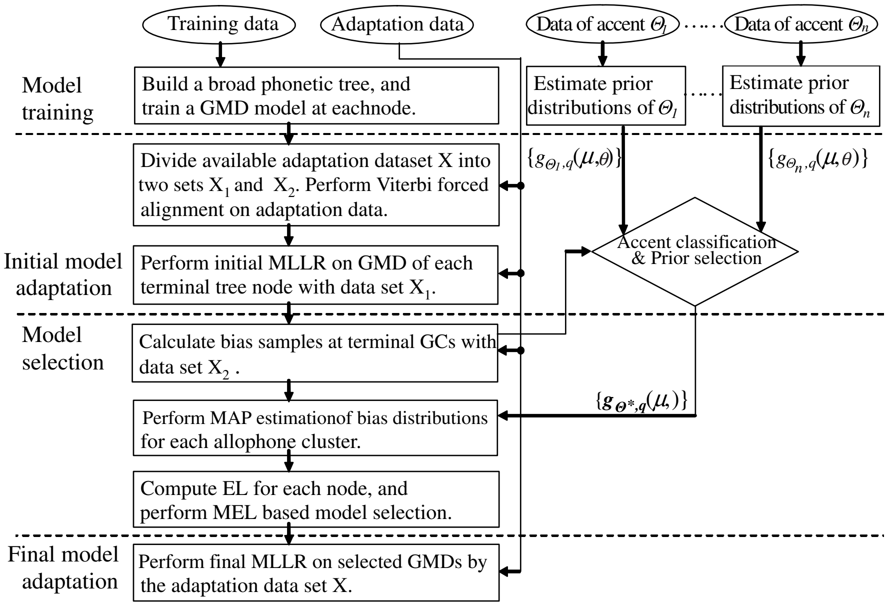
Training data
Adaptation data
Data of accent Θ_1
Data of accent Θ_n
Model training
Build a broad phonetic tree, and train a GMD model at eachnode.
Estimate prior distributions of Θ_1
Estimate prior distributions of Θ_n
$\{g_{\Theta_1,q}(\mu,\theta)\}$
$\{g_{\Theta_n,q}(\mu,\theta)\}$
Divide available adaptation dataset X into two sets X_1 and X_2. Perform Viterbi forced alignment on adaptation data.
Initial model adaptation
Perform initial MLLR on GMD of each terminal tree node with data set X_1.
Accent classification & Prior selection
Model selection
Calculate bias samples at terminal GCs with data set X_2 .
$\{g_{\Theta^*,q}(\mu,)\}$
Perform MAP estimationof bias distributions for each allophone cluster.
Compute EL for each node, and perform MEL based model selection.
Final model adaptation
Perform final MLLR on selected GMDs by the adaptation data set X.

Fig. 5. Procedure of P-MEL based dynamic model selection and adaptation.

3.4. P-MEL based dynamic model selection and adaptation

Fig. 5 shows the overall procedure of the P-MEL based model selection and adaptation method. The prior knowledge related components as proposed in the current work are shown as modules of "Estimate prior distributions" and "Accent classification and Prior selection." The estimated accent-specific priors are first used as accent models in accent classification. From input adaptation data of a speaker, a bias set B is computed and used for classification of the speaker's accent. The set of priors that best match the speaker's accent is selected, and the set of priors is then used as the prior knowledge for MAP estimation of bias distributions. The flow chart consisting of "Model training", "Initial model adaptation", "Model selection" and "Final model adaptation" that lies in the left hand side of Fig. 5 is discussed in Section 2.4.

4. Experimental results

Joint MAP estimation of mean and variance of bias distributions as well as accent classification were implemented and evaluated. In order to evaluate stackable gains of MEL and P-MEL based model complexity adaptation on top of model parameter adaptation, two alternative methods of model parameter adaptation, MLLR (Leggetter and Woodland, 1995) and cascade MLLR (Digalakis et al., 1999), were used in the model adaptation steps in Figs. 3 and 5. The cascade MLLR accomplishes model transformation in two stages. In the first stage, a small number of MLLR transformation matrices (full or diagonal) are employed to perform model transformation in a somewhat global manner, since each transform matrix covers many states of PDTs. In the second stage, a large number of bias transforms are employed to perform transformations in a somewhat local manner on models that have been transformed in the first stage, since each transform vector covers fewer states of PDTs. The cascade MLLR method was shown previously to be superior to MLLR. Here it is expected that the overall performance of combined model complexity selection with model parameter adaptation will improve with better method of model parameter adaptation.

4.1. Experimental condition

The baseline acoustic model was the same as used in He and Zhao (2003) and was trained from the entire set of speaker-independent short-term training data (SI_TR_S, 200 speakers) of WSJ1, where within-word triphone HMM model each 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. Speech features consisted of 39 components of 12 MFCCs, energy, and their delta and acceleration derivatives. Cepstral mean normalization 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 (Young et al., 1999). The acoustic model complexity was optimized to have 6473 tied states for recognition of native speech (He and Zhao, 2003). In testing, the language model score scale and word insertion penalty were tuned for recognizing native American English speech.

Given the native American English speech trained acoustic model, recognition word errors versus model complexity (measured in number of Gaussian densities) were first evaluated on the 5K Wall Street Journal standard data sets ET-H2 and ET-S3, where the former consisted of 10 native American English speakers and the latter consisted of 10 nonnative English speakers with different mother tongues. The curves for the two sets of data are shown in Fig. 6. It is seen that while on ET-H2 the lowest word error rate was achieved at the Gaussian density count of over 100 K, on ET-S3 the lowest word error rate was achieved at the Gaussian density count below 20 K. The optimal Gaussian density count, or model complexity, for the trained acoustic model is indeed significantly different for native and nonnative American English speakers.

For estimating the priors of the bias distributions and performing recognition tests, speech data sets of native American English speech, British accented speech and Chinese accented speech were used. Native American English speech data consisted of WSJ1 set SI_DT_05, referred to as NT1, and WSJ1 set SI_ET_H2, referred to as NT2. NT1 included 10 speakers, named as 4k0–4k4 and 4k6–4ka, with each speaker providing about 90 utterances. NT2 also included 10 speakers, 4oa–4oj, with each speaker providing about 60 utterances. The British accented speech data came from WSJCAM0 (www.ldc.upenn.edu), where 20 speakers were

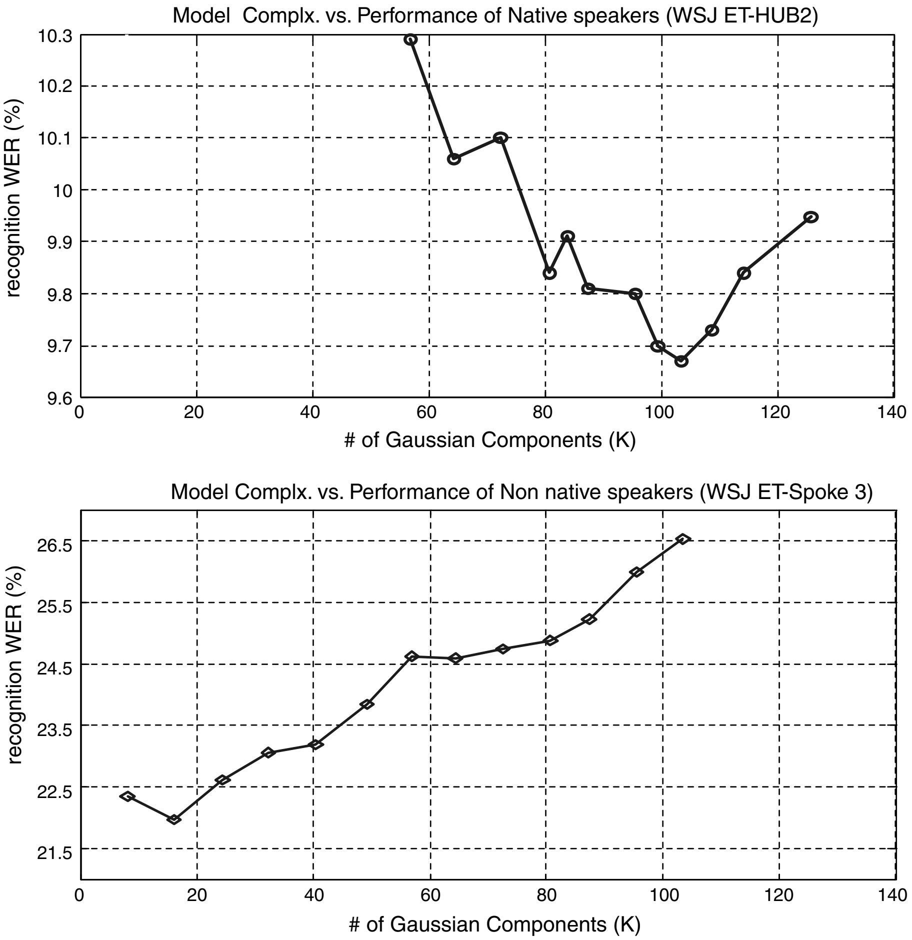

Fig. 6. Comparison on curves of recognition word errors versus model complexity between native American English and nonnative speakers. Evaluation was made on the 5K Wall Street Journal task, where the test sets of native American English and nonnative speech were ET-H2 and ET-S3, respectively.

included in the experiments, each providing about 85 utterances. The speakers were divided into two groups, BR1 and BR2, shown in Table 3. The Chinese accented speech data set CH1 included six talkers, with three males and three females. The speech data were collected locally under a similar acoustic condition (microphone and SNR) and with the same prompting texts as WSJ1, with each talker providing about 120 utterances. In Table 4, the baseline recognition performance on the five groups of speakers is provided. The Chinese accent speaker group was the most difficult, with an average word error rate of 64.55%. The British speaker group also had significant difficulties, with more than 20% higher word error rate (in absolute) in comparison with the native American English speaker group. Further details for the Chinese accent group is shown for individual speakers in Table 5, with high word error rates observed across the board.

In the experimental evaluation of the P-MEL method, NT1 and BR1 were used to estimate the hyperparameters of the priors of the native speakers and British accent speakers, respectively, and NT2 and BR2

Table 3
Definition of two groups of British accent English speakers

Group	Speaker list
BR1	C31, C34, C35, C38,C3C, C3D, C3F, C3J, C3K, C3L
BR2	C32, C37, C39, C3B, C3O, C3R, C3Y, C46, C48, C4A

Table 4
Baseline recognition word error rates on the five data sets: NT1, NT2, BR1, BR2, CH1

Speaker ID	Baseline word error rate (%)
NT1	10.86
NT2	9.67
BR1	31.41
BR2	35.62
CH1	64.55

Table 5
Word error rates of the six speakers in the Chinese accent speech data set CH1

Speaker	WER by baseline model (%)	Gender
ch1	71.97	Male
ch2	62.14	Male
ch3	63.87	Male
ch4	51.16	Female
ch5	73.99	Female
ch6	64.16	Female

were used as the testing sets of native speakers and British accent speakers, respectively. Due to the small number of speakers in CH1, the leave-one-out scheme was deployed. In each round, one of the six speakers in CH1 was held out as the test speaker and the speech data of the rest five speakers were used to estimate the priors. This procedure looped over the six speakers so that each speaker was used as the held-out test speaker once, and the average result of the six test speakers was taken as the result of the CH1 set. The experiments employed several empirically chosen parameters, where system performances were found not very sensitive to these parameters. In estimating the hyper-parameters, the lower bound on the number of feature frames for estimating a bias sample was set to be 35, and at least 30 samples of biases were used to estimate a prior distribution, where the concern was on obtaining reliable and proper number of bias distributions.

4.2. Analysis of prior distributions

The estimated hyper-parameters of the prior distributions, v_q and s_q^2, for native American English speech, British accented speech and Chinese accented speech were compared. In Table 6, the components of v_q and s_q^2 that correspond to the first MFCC coefficient, i.e., v_{q1} and s_{q1}^2, are shown.

Table 6
Comparison of the hyper-parameters in the priors of the native American English speech, Chinese accented speech, and British accented speech, where v_{q1} and s_{q1}^2 are averaged within each phonetic class and correspond to the first MFCC coefficient

Phonetic class	Average $\mathbf{s}_{q1}^2$			Average $\mathbf{v}_{q1}$		
	NT	CH	BR	NT	CH	BR
Front vowel	1.66	4.36	2.49	0.20	0.90	0.01
Low-back vowel	1.99	3.99	6.77	0.30	−0.28	1.02
High-back vowel	1.80	3.86	4.51	−0.17	0.51	1.27
Front diphthong	1.67	3.77	3.63	0.29	0.82	0.37
Back diphthong	1.98	7.45	5.27	0.62	0.74	0.40
Liquid semi-vowel	1.56	6.39	4.97	0.31	−1.17	−0.58
Glide semi-vowel	1.73	3.29	7.45	−0.25	0.94	0.53
Voiced stop	1.46	4.70	3.46	0.02	0.00	−0.06
Unvoiced stop	1.43	4.00	4.23	−0.06	−0.08	0.01
Voiced fricative	2.21	4.34	2.19	0.19	−0.21	−0.72
Unvoiced fricative	2.38	4.09	3.18	0.00	0.17	−1.17
Nasal	1.24	2.41	2.04	−0.58	0.43	0.00

It is observed that the parameter s_q^2 is better than v_q in representing data-model mismatch in phone classes, since in computing v_q, biases are averaged in each phonetic class such that positive and negative biases neutralize. For British and Chinese accented speech, the s_q^2's are in general much larger than that of native American English speech, verifying that the level of data-model mismatch is much higher in foreign accented speakers than native speakers. Moreover, the dynamic range of s_q^2 across the phonetic classes of the foreign accented speakers is much larger than that of native speakers, where the largest s_q^2 is about three to four times of the smallest s_q^2 for British and Chinese accented speakers, compared with a factor of less than two for native speakers. This result indicates that for British and Chinese speakers, certain phones are pronounced better than others, and the pronunciation accuracy varies largely across the phone classes. For example, British speakers tend to have a large s_{q1}^2 in the low-back vowels, while the Chinese speakers tend to have large s_{q1}^2's in the back-diphthongs and liquid semi-vowels. The latter case is interesting since back-diphthongs do not exist in Mandarin Chinese, and liquid semi-vowels of "wh" and syllable-ending "l" are nonexistent in Chinese either.

Evaluations were also made on model selection by using priors alone and the effect of model selection on recognition accuracy. Nine cases were studied, which were generated by combinations of the three test speaker groups and the three sets of priors as described in Section 4.1. Recognition results of the nine cases are summarized in Table 7, and the selected model complexity resulting from using the accent-matched priors are shown in Table 8.

Compared with the baseline word error rates in Table 4, there were large performance improvements to British and Chinese accented speech BR2 and CH1 due to accent-matched priors, and only slight performance degradation was observed for the native American English speaker set NT2. As shown in Table 4, the baseline word error rate of NT2 was smaller than that of NT1, indicating that the priors estimated from NT1 might have led to somewhat coarse model selection for NT2. It can also be observed from Table 7 that applying accent-mismatched priors to native American English speakers caused a significant performance degradation, whereas performance improvements were achieved for the British and Chinese accented speakers even with mismatched priors. This result motivated the use of $R(\Theta)$ in Eq. (9). By encouraging the accent classifier to classify talkers as native speakers the risk of increasing word error rate due to a mismatch in priors would be reduced.

4.3. Accent classification

Based on the estimated priors of phones and phonetic classes, the proposed method of automatic accent detection as described in Section 3.3 was evaluated on the test sets NT2, BR2, and CH1. Each speaker provided 40 adaptation utterances. For each test speaker, the first N utterances were selected from his or her adaptation data set for use in accent classification, where the number N was set to be 1, 3, 5, 10, 20, and 40, respectively. In the cases of $N > 1$, the lower bound on the number of feature frames for computing a bias

Table 7
Recognition word error rates (%) after P-MEL model selection, where the selection was made by using the prior distributions only

Test set	Prior		
	CH1	NAT1	BR1
CH1	54.05	57.75	54.19
NAT2	11.15	10.40	11.51
BR2	32.34	32.66	32.07

Table 8
Selected model complexity by P-MEL using accent-matched priors

Speaker set	# states after model selection
CH1	1752
NAT1	3525
BR1	1776

Table 9
Error counts in accent detection

# utterances	1	3	5	10	20	40
BR	0	0	0	0	0	0
CH	1	1	0	0	0	0
NT	1	0	0	0	0	0

sample was set to be 35, and in the case of $N = 1$ the bound was set to 25. The constant C_{nat} in (10) was set to be 2.5. This value was empirically determined based on the development sets NT1 and BR1. It was chosen to reduce the risk of classifying a native American English speaker to a foreign accent speaker. The classification error counts as a function of the used number of utterances is shown in Table 9.

It is observed that the accent classifier worked very well when more than three utterances were used. The errors in CH1 were from confusion with the native American English speaker group, while the error in NT2 was confusion with the Chinese accented speaker group.

4.4. Prior knowledge guided dynamic model selection and adaptation

Recognition experiments were conducted on the proposed P-MEL based method of dynamic model selection and adaptation (as described in Fig 5). In order to avoid unreliable hyper-parameter S^2 in Eq. (7) of priors, it was determined that at least six samples of biases be accumulated in each phone or phonetic class, or else the bias distribution be estimated directly from the prior knowledge, i.e., set $n = 0$ in Eq. (7). The MEL method as proposed in He and Zhao (2003) was also implemented under a similar condition, with the threshold on the number of biases for a full node set as 25, and the threshold on the number of data frames for a full terminal GC set as 30. A clustered phonetic decision tree similar to the one in He and Zhao (2003) was used and the bias distributions were tied to have 42 clusters that corresponded to the 42 phone units.

For each test speaker, N utterances were selected from the adaptation data set for model adaptation and selection, where N was set to be 1, 3, 5, 10, 20, and 40, respectively. The partition of the selected N adaptation utterances into two subsets, one for initial model adaptation and one for mode 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 number of utterances. When the adaptation data were less than 20 utterances, all utterances were used in model selection, and a subset of them was used in initial model adaptation. In Table 10, the assignment of data for initial model adaptation and the chosen forms of transformations for MLLR and cascade MLLR are summarized for different values of N's.

Recognition results on the three speaker groups CH1, BR2, and NT2 are shown in Figs. 7 and 8. In Fig. 7, three curves are drawn for each type of accent: MLLR alone, MEL combined with MLLR, and P-MEL combined with MLLR. In Fig. 8, again three curves are drawn for each type of accent: cascade MLLR alone, MEL combined with cascade MLLR, and P-MEL combined with cascade MLLR. The following observations can be made from the two figures: (1) both MEL + MLLR and P-MEL + MLLR outperformed MLLR alone, and similarly both MEL + cascade MLLR and P-MEL + cascade MLLR outperformed the cascade MLLR alone;

Table 10
Detailed account of data assignment for initial model adaptation and choice of transforms for MLLR and cascade MLLR, where in the case of $N = 20$ or 40, the threshold for a full transform was set as 500 samples, and the threshold for a bias transform was set as 100

N	Number of sentences for initial model adaptation	Transforms in MLLR	Transforms in cascade MLLR
1	0	NA	NA
3	1	One global bias transform	One global bias transform and multiple bias transforms (if any based on the threshold setting)
5	3	One global diagonal transform	One global diagonal transform and multiple bias transforms
10	5	One global full transform	One global full transform and multiple bias transforms
20	10	Multiple full transforms	Multiple full transforms and multiple bias transforms
40	20	Multiple full transforms	Multiple full transforms and multiple bias transforms

(2) P-MEL + MLLR outperformed the MEL + MLLR; (3) generally, cascade MLLR performed better than MLLR, and the advantage of cascade MLLR was maintained in the combined model selection and adaptation methods and a better overall performance was resulted. It is worth noting that if by certain means the accent

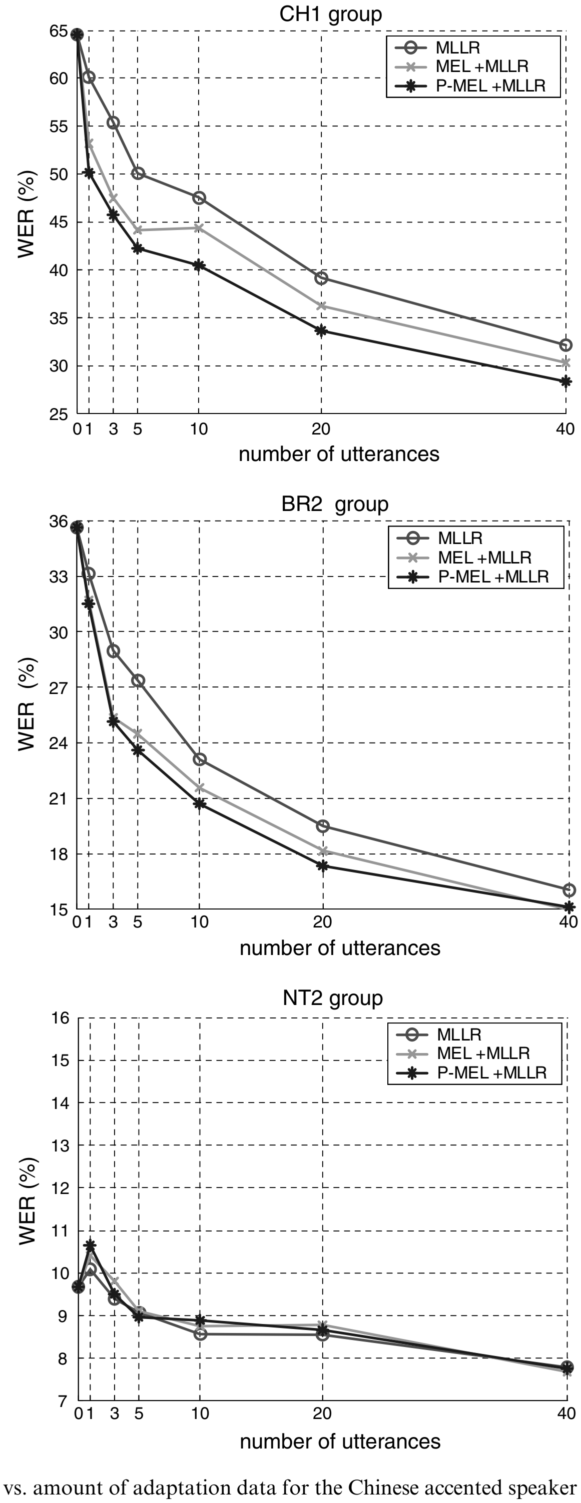

Fig. 7. Recognition word error rate vs. amount of adaptation data for the Chinese accented speaker set CH1, the British accented speaker set BR2, and the native speaker set NT2. The baseline is MLLR.

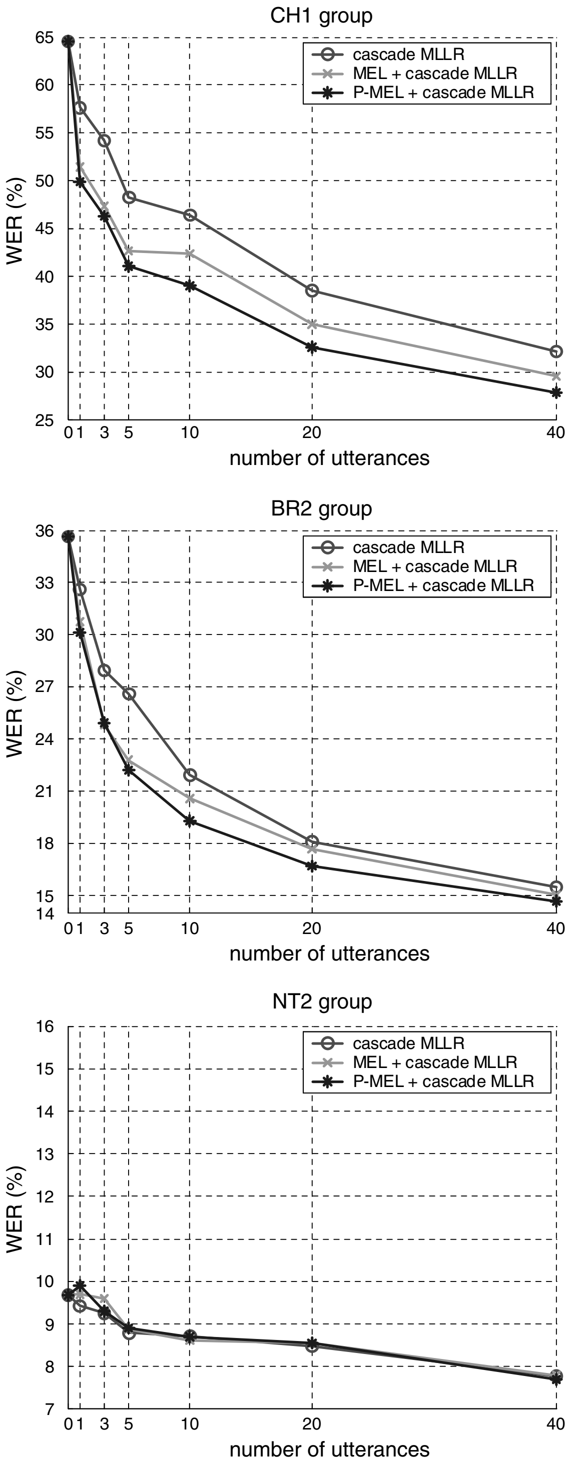

Fig. 8. Recognition word error rate vs. amount of adaptation data for the Chinese accented speaker set CH1, the British accented speaker set BR2, and the native speaker set NT2. The baseline is cascade MLLR. On average, cascade MLLR is about 2% better than conventional MLLR.

types of the foreign accent speakers were given, then the word error rate of the P-MEL method would be significantly reduced at $N = 0$, as given in Table 7 for the matched prior cases, since then the bias distributions can be estimated based on the accent-matched priors without waiting for adaptation data.

Word error rates shown in Figs. 7 and 8 are results of combined model selection and model adaptation. To further identify the effect of model selection, word error rates for the case of using model selection alone are given in Fig. 9. It is clear that for nonnative speakers, model selection alone improved speech recogni-

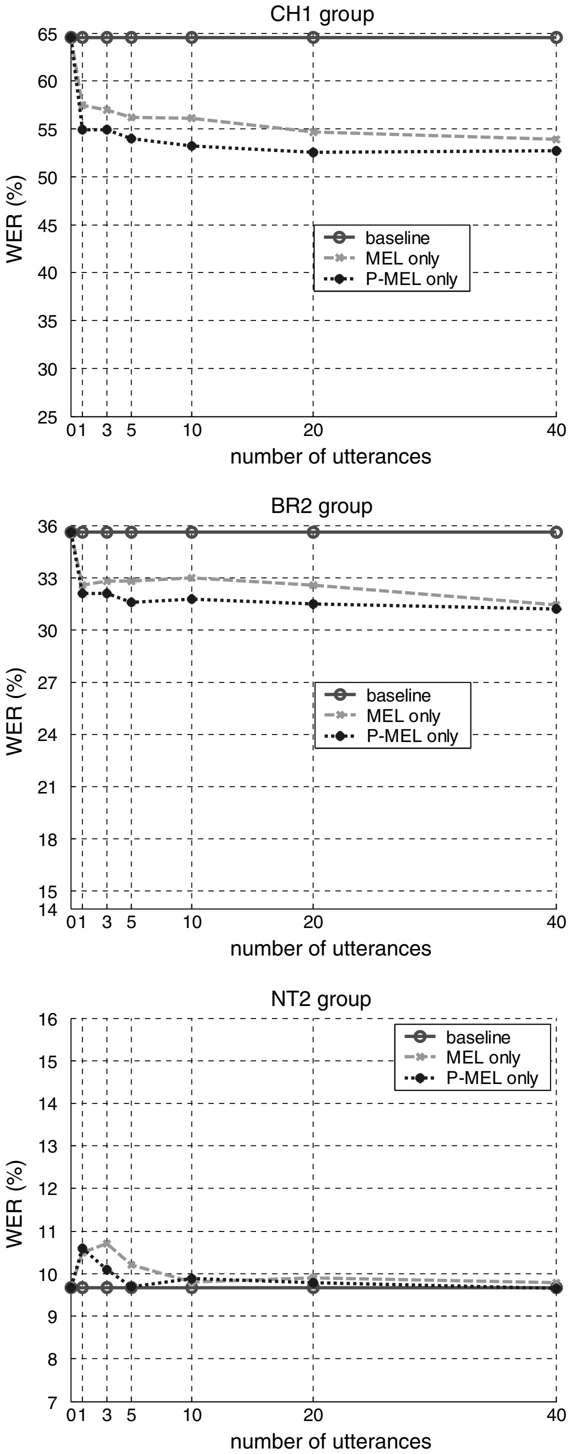

Fig. 9. Recognition word error rate vs. amount of speech data used for model selection for the Chinese accented speaker set CH1, the British accented speaker set BR2, and the native speaker set NT2. The baseline is SI model, and for MEL and P-MEL, parameter adaptation (i.e., MLLR) was skipped after model selection.

tion accuracy significantly, and PMEL outperformed MEL when the numbers of adaptation utterances were very small.

5. Summary

The method of MEL based model complexity selection as proposed in He and Zhao (2003) is effective when a proper amount of adaptation data is available, where recognition accuracy of nonnative English speech can be significantly improved without performance degradations on native speakers. However, when adaptation data is very limited, the distributions of data-model mismatch biases cannot be reliably estimated at a sufficient level of details, making reliable model selection a difficult task.

In the current work, a more robust method for selecting model complexity in the case of sparse data is proposed by utilizing knowledge of foreign accents. Knowledge of different accents is represented as accent-specific priors for bias distributions that model data-model mismatch so that maximum *a posteriori* estimation that is more robust than maximum likelihood when data is sparse is performed for bias distributions. Algorithms are developed for estimating hyper-parameters of the prior distributions and for automatic classification of accents so that priors that match a speaker's accent can be applied. Experiments were performed on three types of speech data, including American English speech, British accented speech, and the mandarin Chinese accented speech. Results show that the estimated prior distributions reflected meaningful data-model mismatch condition for each set of speakers. With the priors serving as knowledge of each accent, more reliable estimation of bias distributions was made in the case of very small amount of adaptation speech, or even without any adaptation speech. Experimental results show that the proposed P-MEL method is superior to the previous MEL method for model selection, especially when adaptation data are extremely limited. Experimental results on combining P-MEL and MEL with MLLR and cascade MLLR also indicate that the P-MEL or MEL based methods scale well with model parameter adaptation methods, where overall recognition accuracy improves with the quality of model parameter adaptation method. As such, the proposed methods of model complexity selection can be combined with varieties of model parameter adaptation techniques, for example, the multi-scale based cascade MLLR method (Kannan and Ostendorf, 1997) and the MLLR + MAP method (Digalakis and Neumeyer, 1996).

A practically important issue is how to handle nonnative speakers without priors trained for their accents. It is therefore desirable to devise a prior modeling method that allows clustered modeling of foreign accents so as to cover those accents that do not have trained priors. A flexible framework for such a purpose is mixture modeling, where available foreign accented speech data can be used to train a mixture prior model, and for a speech utterance with an unknown foreign accent, the mixture prior can be adapted to the speaker by replacing *a priori* mixture weights with the *a posteriori* probabilities of mixture components that are computed from online speech. The adapted mixture prior can then be used for knowledge guided model selection and adaptation. It is noted that a Gaussian mixture prior of conventional acoustic model parameters was previously proposed for acoustic model training from mixed cellular and wire line speech datasets with different regional dialects of American English (Buhrke and Liu, 2000), where mixture weights were fixed empirically. In contrast, adaptive prior approach will utilize posterior mixture weights that are more suitable for untrained foreign accented speech. Our preliminary experiment on this approach showed a promising result. This approach will be fully investigated in a future work, including the design of dataset with a good coverage of representative nonnative speech characteristics, as well as dynamical estimation of phone dependent mixture-prior weights. The proposed methods of prior modeling of bias distributions and accent classification will be evaluated on more varieties of foreign accents, and a further analysis will be made on the phone-class dependent properties of hyper-parameters in representing data-model mismatch of different foreign accents.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. AC-19, 716–723.

Berkling, K., 2001. SCoPE, syllable core and periphery evaluation: Automatic syllabification and foreign accent identification. Speech Commun. 35, 125–138.

Buhrke, E.R., Liu, C., 2000. A generalization of the maximum a posteriori training algorithm for mixture priors. Proc. ICASSP, 993–996.

Byrne, W. et al., 2000. Towards language independent acoustic modeling. Proc. ICASSP, 1029–1032.

Chen, T., Huang, C., Chang, E., Wang, J. 2001. Automatic accent identification using Gaussian mixture models. In: Proceedings of the ASRU.

Compernolle, D., 2001. Recognizing speech of goats, wolves, sheep and non-natives. Speech Commun. 35, 71–79.

DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill, New York.

Digalakis, V., Neumeyer, L., 1996. Speaker adaptation using combined transformation and Bayesian methods. IEEE Trans. Speech Audio Process. 4, 294–300.

Digalakis, V. et al., 1999. Rapid speech recognizer adaptation to new speakers. Proc. ICASSP 2, 765–768.

Fischer, V., Janke, E., Kunzmann, S., Ross, T., 2001. Multilingual acoustic models for the recognition of non-native speech. In: Proceedings of the Automatic Speech Recognition and Understanding workshop.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2, 291–298.

He, X., Zhao, Y., 2001. Model complexity optimization for nonnative English speakers. Proc. EUROSPEECH, 1461–1464.

He, X., Zhao, Y., 2003. Fast model selection based speaker adaptation for nonnative speech. IEEE Trans. Speech Audio Process. 11, 298–307.

Huo, Q., Chan, C., Lee, C.-H., 1995. Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition. IEEE Trans. Speech Audio Process. 3, 334–345.

Kannan, A., Ostendorf, M., 1997. Modeling dependence in adaptation of acoustic models using multiscale tree processes. In: Proceedings of the EUROSPEECH.

Kohler, J., 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. Proc. ICSLP, 2195–2198.

Kullback, S., 1959. Information Theory and Statistics. Wiley, New York.

Lee, C.-H., Lin, C.-H., Juang, B.-H., 1991. A study on speaker adaptation of the parameters of continuous density hidden Markov models. IEEE Trans. Signal Process., 806–814.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech Lang. 9, 171–185.

Matsunaga, S., Ogawa, A., Yamaguchi, Y., Imamura, A., 2003. Non-native English speech recognition using bilingual English lexicon and acoustic models. Proc. ICASSP 1, 340–343.

Minematsu, N., Kurata, G., Hirose, K., 2002. Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition. Proc. ICSLP, 529–531.

Rabiner, L.R., Juang, B-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 333–348.

Rissanen, J., 1984. Universal Coding, Information, Prediction, and Estimation. IEEE Trans. IT 30.

Schultz, T., and Waibel, A., 1998. Multilingual and crosslingual speech recognition. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Teixeira, C., Trancoso, I., Serralheiro, A., 1996. Accent identification. In: Proceedings of the ICSLP.

Tomokiyo, L.M., 2000. Lexical and acoustic modeling of nonnative speech in LVCSR. Proc. ICSLP 4, 346–349.

Uelber, U., Boros, M., 1999. Recognition of non-native German speech with multilingual recognizers. Proc. EUROSPEECH, 911–914.

Wang, S., Zhao, Y., 2001. Online Bayesian tree structure transformation of HMMs with optimal model selection for speaker adaptation. IEEE Trans. Speech Audio Process. 9, 663–677.

Weng, F., et al., 1997. A study of multilingual speech recognition. In: Proceedings of the EUROSPEECH.

Wijngaarden, S.J.V., 2001. Intelligibility of native and non-native Dutch speech. Speech Commun. 35, 103–113.

Witt, S., Young, S., 1999. Offline acoustic modeling of nonnative accents. In: Proceedings of the EUROSPEECH.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1999. The HTK book, Version 2.2. Available from: http://htk.eng.cam.ac.uk/docs/docs.shtml.

Zavaliakos, G., Schwartz, R., Makhoul, J., 1995. Batch, incremental and instantaneous adaptation techniques for speech recognition. Proc. ICASSP, 676–679.

Zavaliakos, G., 1996. Maximum a posteriori adaptation for large scale HMM recognizers. Proc. ICASSP, 725–728.

Zhao, Y., 1996. Self-learning speaker and channel adaptation based on spectral variation source decomposition. Speech Commun. 18 (1), 65–77.