

FAST MODEL ADAPTATION AND COMPLEXITY SELECTION FOR NONNATIVE ENGLISH SPEAKERS

Xiaodong He and Yunxin Zhao

Dept. of Computer Engineering and Computer Science
University of Missouri, Columbia, MO 65211, USA

ABSTRACT

In this paper, the problem of fast model adaptation and complexity selection for nonnative speaker is investigated. The key challenge lies in reliable complexity selection based on a small amount of adaptation data. A novel technique of combining MDL with pseudo likelihood-based state-tying is proposed to enable model complexity selection from using as little as three adaptation speech sentences. In MDL/PL, MDL is performed on nodes with sufficient adaptation data, and pseudo-likelihood based state tying is performed on nodes with insufficient adaptation data. Experiments were performed on WSJ data of six nonnative speakers. The combined model adaptation and complexity selection method led to consistent and significant improvement on recognition accuracy over MLLR, with an average error reduction of 10% when a varying number of adaptation speech sentences were taken from each speaker.

1. INTRODUCTION

Currently, almost all American English speech recognition systems are trained from speech data of native American English speakers. Although the systems work very well for native speakers, 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 and limited data, in general it is difficult to train a set of acoustic models for each specific accent.

Many efforts have been made to improve recognition performance for nonnative speakers [1,2]. A straightforward approach is to apply general speaker adaptation techniques of model parameter estimation or transformation on speaker-independent (SI) models to fit the characteristics of a foreign accent. Commonly used adaptation algorithms include Maximum Likelihood Linear Regression (MLLR) and Maximum a posteriori (MAP) estimation. It was shown previously [2] that although speaker adaptation can reduce recognition errors for both native and nonnative speakers, much more adaptation data are needed from the latter than from the former to achieve a similar level of performance.

In [3], the problem of model adaptation for foreign accent speakers was investigated from a different perspective of model complexity selection. The approach was motivated from the fact

that highly detailed English acoustic models do not fit well to speech with heavy foreign accents, while a certain level of context-dependent modeling is needed for discrimination among allophones. Therefore, an intermediate level of acoustic model complexity determined from adaptation speech may work best for a foreign accent talker. In [3], model complexity selection was accomplished by empirically choosing state-tying thresholds for phonetic decision trees and by applying the principle of MDL. Experimental results showed that for native speakers and nonnative speakers, the curves of model complexity vs. performance are significantly different. Highly detailed acoustic models that produced the best recognition result for native speakers worked worst for nonnative speakers. Moreover, by using MDL to automatically control model complexity, much improved performance was achieved for nonnative speakers.

In MDL-based model complexity selection, detailed phonetic decision trees are first built to organize allophone states hierarchically, and MDL is applied to perform tree pruning and the optimal tree cuts representing optimal model complexity are obtained. In general, using MDL for model complexity selection requires a large amount of data. In [3], 276 sentences of Chinese-accented speech were used in an unsupervised mode for MDL-based model selection. This drawback of MDL prevents its application in on-line fast speaker adaptation. In the current work, we investigate the problem of determining proper model complexity on acoustic models of native English for nonnative English speakers by using a small amount of adaptation data. An approach that combines MDL with likelihood-based state tying is proposed. From the tree nodes with sufficient adaptation data, an average bias can be estimated between the sample means of a speaker's feature data and the model means of the native English training data. The average bias reflects the degree of mismatch between the speaker's speech characteristics and the trained model and is used to compute a measure of pseudo likelihood for every tree node. In the meanwhile, with a small amount of data, MDL can be reliably performed at several nodes. From these nodes, a threshold of pseudo-likelihood increment can be estimated for tree pruning on nodes that have insufficient data.

This paper is organized as four sections. In section 2 the proposed method is described. In section 3, experimental results are provided. A conclusion is made in section 4.

2. MODEL SELECTION USING MDL/PL

2.1. Conventional approach for model complexity selection

In phonetic decision tree based state tying, a context-dependent decision tree is built for each phone state based on phone context. At the root node of a tree, all allophone states are tied and modeled by a single Gaussian density (GD). At each node, the state set is tentatively split into two subsets by asking a phonetic context question, the likelihood increment due to the split is measured for each question, and the question that leads to the maximum likelihood gains is chosen and the state split is determined accordingly. This procedure is carried out top-down recursively until likelihood increment drops below a pre-defined threshold. Allophone states that fall into the same terminal nodes are tied together as a single state. Model complexity is selected by empirically defined thresholds [4].

Minimum Description Length (MDL) [5] is an information theoretical criterion that has been proven effective in the selection of optimal model complexity based on a certain amount of observation data. Several MDL based speaker adaptation techniques were proposed previously [6, 7]. The MDL procedure for model complexity selection is illustrated in Figure 1. For a node G with a single Gaussian model $\lambda_G = (\mu_G, \Sigma_G)$, its description length (DL) for a given set of data X is computed as:

$$DL(X; G) = -\sum_{i=1}^N \log(N(x_i; \mu_G, \Sigma_G)) + \frac{Q}{2} \log(N)$$

where $X = \{x_1, x_2, \dots, x_N\}$ is the feature set aligned to this node, μ_G and Σ_G are mean vector and covariance matrix of the Gaussian density, and Q is the number of free parameters of the Gaussian density. For the left-most branch in Figure 1, we can compute the term

$\Delta DL(G_P, G_L, G_R) = DL(X_P; G_P) - [DL(X_L; G_L) + DL(X_R; G_R)]$ where X_P , X_L , and X_R are the feature sets assigned to the nodes G_P , G_L , and G_R , respectively, and $X_L \cup X_R = X_P$ and $X_L \cap X_R = 0$. If $\Delta DL(G_P) < 0$, then the two children nodes will be discarded, otherwise the DL of parent node G_P is assigned as: $DL(X_P; G_P) = [DL(X_L; G_L) + DL(X_R; G_R)]$. This procedure is carried out bottom-up over all the nodes of a tree.

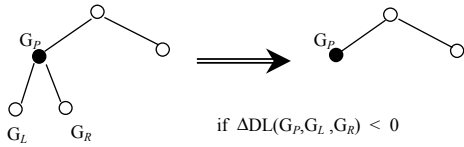


Figure 1. MDL-based tree pruning

2.2. MDL/PL approach for fast adaptation

Assume that the speech feature dimension is D and the covariance matrix of all Gaussian densities are diagonal. Then for the Gaussian density at the node G , $\mu_G = [\mu_{G,1}, \dots, \mu_{G,D}]^T$ and $\Sigma_G = \text{diag}[\sigma_{G,1}^2, \dots, \sigma_{G,D}^2]$, and the log likelihood of X becomes

$$L(X; G) = -\frac{N}{2} \sum_{d=1}^D \log(2\pi\sigma_{G,d}^2) - \sum_{d=1}^D \frac{1}{2\sigma_{G,d}^2} \sum_{i=1}^N (x_{i,d} - \mu_{G,d})^2$$

Denote the sample mean of X by $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$. Then $L(X; G)$ can be expressed as

$$-\frac{N}{2} \sum_{d=1}^D \log(2\pi\sigma_{G,d}^2) - \sum_{d=1}^D \frac{1}{2\sigma_{G,d}^2} \left[\sum_{i=1}^N (x_{i,d} - \bar{x}_d)^2 + \sum_{i=1}^N (\bar{x}_d - \mu_{G,d})^2 \right]$$

$$= -\frac{N}{2} \sum_{d=1}^D \log(2\pi\sigma_{G,d}^2) - \frac{N}{2} \sum_{d=1}^D \frac{\hat{\sigma}_{G,d}^2}{\sigma_{G,d}^2} - \frac{N}{2} \sum_{d=1}^D \frac{E_{G,d}^2}{\sigma_{G,d}^2}$$

where $\hat{\sigma}_{G,d}^2 = \frac{1}{N} \sum_{i=1}^N (x_{i,d} - \bar{x}_d)^2$ is the sample variance of the d -th feature component at the node G and $E_{G,d} = \bar{x}_d - \mu_{G,d}$ is the bias between the sample mean of the d -th feature component and the model mean of the d -th feature component at the node G . Averaging over N , the average likelihood becomes

$$AL(X; G) = -\frac{1}{2} \left[\sum_{d=1}^D \log(2\pi\sigma_{G,d}^2) + \sum_{d=1}^D \frac{\hat{\sigma}_{G,d}^2}{\sigma_{G,d}^2} + \sum_{d=1}^D \frac{E_{G,d}^2}{\sigma_{G,d}^2} \right]$$

From all the nodes G that have sufficient adaptation data, an average bias can be computed as $E_d = \bar{E}_{G,d}$. For each node G' that does not have sufficient data, an approximation is made that $E_{G',d} = E_d$ and $\hat{\sigma}_{G',d}^2 = \sigma_{G',d}^2$, and a pseudo-likelihood score is computed as:

$$PL(G') = -\frac{1}{2} \left[\sum_{d=1}^D \log(2\pi\sigma_{G',d}^2) + D + \sum_{d=1}^D \frac{E_d^2}{\sigma_{G',d}^2} \right]$$

Define the increment of pseudo-likelihood due to splitting the node G_P into its children nodes G_L and G_R as:

$$\Delta PL(G_P, G_L, G_R) = \frac{1}{2} [PL(G_L) + PL(G_R)] - PL(G_P)$$

Let Σ_P, Σ_L , and Σ_R be the diagonal covariance matrices corresponding to nodes G_P, G_L and G_R , respectively. Then $\Delta PL(G_P, G_L, G_R) =$

$$\frac{1}{2} \sum_{d=1}^D \left[\log\left(\frac{\sigma_{P,d}^2}{\sigma_{L,d} \cdot \sigma_{R,d}}\right) + E_d^2 \left[\frac{1}{\sigma_{P,d}^2} - \frac{1}{2} \left(\frac{1}{\sigma_{L,d}^2} + \frac{1}{\sigma_{R,d}^2} \right) \right] \right] \quad (1)$$

If we further assume that $\sigma_{L,d} \approx \sigma_{R,d} \approx \bar{\sigma}_{LR,d}$, then ΔPL can be further simplified as:

$$\Delta PL(G_P, G_L, G_R) = \frac{1}{2} \sum_{d=1}^D \left[2 \log\left(\frac{\sigma_{P,d}}{\bar{\sigma}_{LR,d}}\right) + E_d^2 \left(\frac{1}{\sigma_{P,d}^2} - \frac{1}{\bar{\sigma}_{LR,d}^2} \right) \right] \quad (2)$$

Eqs (1) or (2) can be used for tree pruning on the empty nodes $\{G'\}$. If $\Delta PL(G_P) < T$, where T is a threshold, the two children nodes will be pruned, otherwise, they will be kept. The threshold T can be estimated from the nodes with enough data and will be described below.

As show in Figure 3, The proposed method of model selection based on combined MDL and PL is implemented as the following seven steps.

In the first step, a single context-dependent state-tying tree is built for each emitting state of each phone unit HMM. As shown in Figure 2, each terminal node of a tree corresponds to an allophone state that is modeled by a single Gaussian density. Each internal node, which is also modeled by a single Gaussian density, corresponds to a tying state that is tied from all the terminal nodes in its sub-tree. Given a tying threshold, certain terminal nodes will be pruned and certain internal nodes become new terminal nodes. In Figure 2, the black nodes correspond to new terminal nodes after tying, and the corresponding states are referred to as preliminarily tied states. The resulting trees are referred to as pruned state-tying trees.

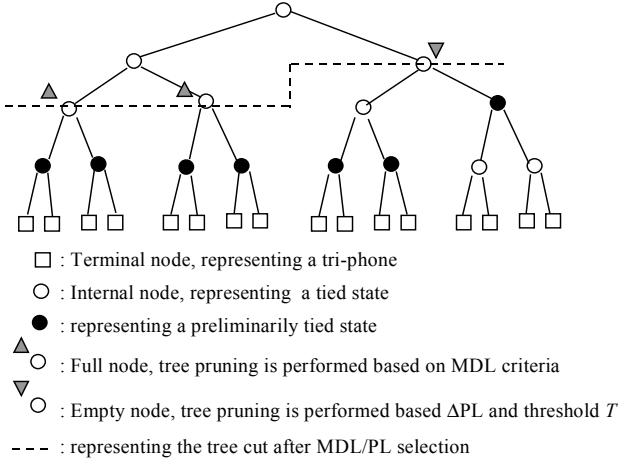


Figure 2. Illustration of a state-tying tree

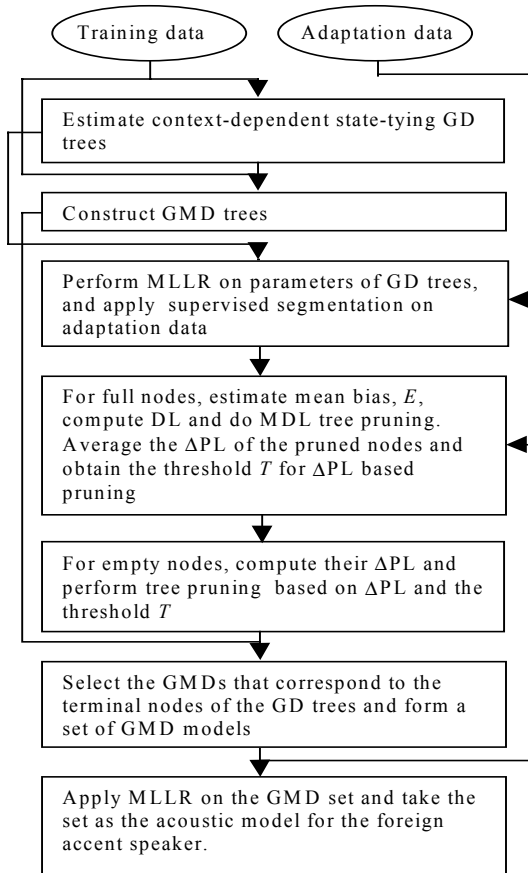


Figure 3. Illustration of a state-tying tree

In the second step, corresponding to each pruned state-tying tree, an identically structured tree copy is constructed where at each node of the tree copy, a Gaussian mixture density (GMD) model is estimated instead of a Gaussian density. The resulting trees are referred to as the GMD trees. The tree structures and corresponding GDs/GMDs will be used in subsequent speaker adaptation.

In the third step, using a given amount of adaptation data, the parameters of the GDs are transformed by MLLR to reduce the mismatch between the GD models and the speech features of the speaker. Using the transformed GD models, a supervised alignment is performed on the adaptation data. The nodes that have sufficient data are referred to as full nodes, and the nodes that do not have sufficient data are referred to as empty nodes.

In the fourth step, for the full nodes, DLs are computed and tree pruning is performed based on the MDL criteria. In addition, the global bias E is estimated, and the Δ PLs of these pruned nodes are computed. The average pseudo-likelihood increment at the pruned nodes is considered to be representative of the proper balance point between model fitting and model complexity for this speaker, and this average increment is taken as the threshold T to be applied to nodes without sufficient data.

In the fifth step, the pseudo likelihood increment (Δ PL) of each empty node is computed. Then tree pruning is performed on each empty node based on Δ PL and the threshold T as determined in the 4-th step.

In the sixth step, the surviving terminal nodes are taken to form a cut of the GD tree as the “optimum” state-tying structure for the foreign accent speaker. The Gaussian mixture density models in the GMD trees that correspond to the terminal nodes of the GD trees are selected as the acoustic models for the speaker.

Finally, in the seventh step, the selected GMDs are further adapted by MLLR using the same set of adaptation data, and as such, a more accurate model set for the foreign accent speaker is obtained.

3. EXPERIMENTAL RESULTS

3.1. Experimental condition

The proposed method was evaluated on the LDC WSJ1.0 Spoke3 task corpus (SI_DT_S3). There are a total of 10 nonnative American English speakers in the Spoke3 data set. Each speaker has 40 adaptation sentences and approximately 40 testing sentences. The first four speakers with heavy foreign accents were included in the test set. In addition, speech data of two other speakers with Mandarin Chinese accent were collected in a similar acoustic condition as WSJ and with the similar prompting texts. For each speaker, The first N adaptation sentences were used as adaptation data, where $N = 3, 5, 10, 20, 30, 40$, and the first 20 testing sentences were used in testing. The recognition results were averaged over the six speakers.

The entire set of speaker-independent short-term training data (SI_TR_S) of WSJ was used for acoustic model training. Each 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. The baseline system was trained by HTK 2.2. It consisted of about 130K Gaussian densities after state tying. The 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 WSJ was used. The baseline system was tested on WSJ HUB2, where an accuracy of 90.10% was achieved by using the models with 8151 states (corresponding to the black nodes in Figure 2).

3.2. Experimental results

The experimental conditions include baseline speaker-independent recognition, the MLLR alone (MLLR), the combination of MDL with PL without assumption of $\sigma_{L,d} = \sigma_{R,d}$ (MDL/PL (1)) and with this assumption (MDL/PL (2)). In the latter condition, $\bar{\sigma}_{LR,d} = \sqrt{\sigma_{L,d} \cdot \sigma_{R,d}}$ was used in the implementation. These results are summarized in Figure 4 for the varying number of adaptation sentences.

As is illustrated by Figure 4, by MDL/PL based model complexity selection, a proper model complexity level and a set of proper models can be chosen by a small amount of speech adaptation speech data. Compared with the baseline system and MLLR alone, a significant error reduction was achieved by using the complexity-controlled models.

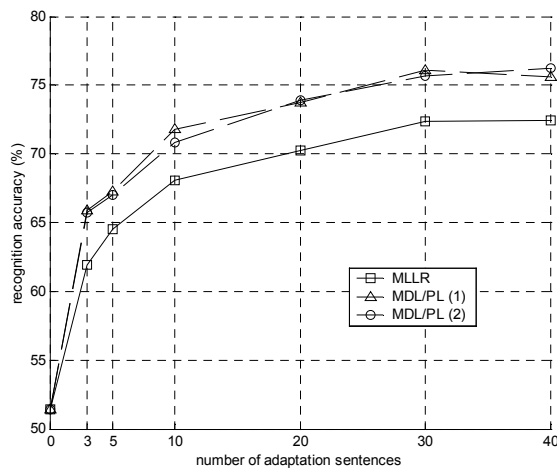


Figure 4. performance vs. adaptation data volume

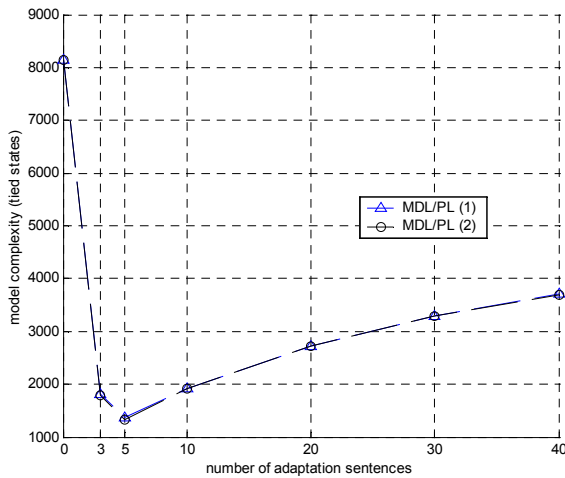


Figure 5. model complexity vs. adaptation data volume

Figure 5 shows that the complexity of the MDL/PL selected models increased as the amount of adaptation data increased. This is intuitively appealing since as more adaptation data became available, the MLLR transformed Gaussians became

more suitable for acoustic match with the speaker's speech, and therefore more detailed models were selected.

On the other hand, with a small amount of adaptation data, highly detailed models could not be adapted well, and it prevented the model to match the speaker's characteristics well. Therefore a less-detailed but more robust model worked better. With more adaptation data, detailed models can be well adapted and therefore the detailed models matched the characteristic of the speaker better than simple models.

We also noted that the results of MDL/PL (1) and MDL/PL (2) are very similar, implying that the approximation made in Eq. (2) is reasonable.

4. CONCLUSION

Model complexity selection methods as applied to detailed acoustic models in general requires large amounts of data in order to reliably compute the fitness of model to data and therefore select a proper level of model complexity. However, the requirement for a large amount of adaptation data is impractical in fast on-line speaker adaptation. In this paper, a novel technique is proposed that combines MDL with PL to accomplish the task of complexity selection from a small amount of adaptation data. Experimental results indicate that on nonnative English speech, model complexity selection led to consistent and significant improvements to MLLR, and the proposed method of MDL/PL worked reasonably well with small amounts of adaptation data.

REFERENCES

- [1] Boulis, C. and Digalakis, V., "Fast Speaker Adaptation of Large Vocabulary Continuous Density HMM Speech Recognizer Using A Basis Transform Approach," *Proc. ICASSP 00*, vol. 2, pp. 989-992.
- [2] Zavaliakos, G. Schwartz, R. and Makhoul, J., "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," *Proc. ICASSP 95*, pp. 676-679.
- [3] He, X. and Zhao, Y., "Model Complexity Optimization for Nonnative English Speakers," *Proc. EUROSPEECH'01*, pp.1461-1464, Scandinavia, Denmark, September 2001.
- [4] Kershaw, D. et al, "The HTK book", <http://htk.eng.cam.ac.uk/docs/docs.shtml>.
- [5] Rissanen, J., "Universal Coding, Information, Prediction, and Estimation", *IEEE Trans. IT*, vol.30, 1984.
- [6] Shinoda, K. and Watanabe, T., "Speaker Adaptation With Autonomous Model Complexity Control By MDL Principle", *Proc. ICASSP 96*, vol. 2, pp. 717-720.
- [7] Wang, S. and Zhao, Y., "Online Bayesian Tree Structure Transformation of HMMs with Optimal Model Selection for Speaker Adaptation," *IEEE Trans. on SAP*, vol. 9, no.6, pp. 663-677, Sept. 2001.