Model Complexity Optimization for Nonnative English Speakers

Xiaodong He and Yunxin Zhao

Dept. of Computer Engineering and Computer Science
University of Missouri, Columbia, MO 65211, USA
xhb1a@mizzou.edu zhaoy@missouri.edu

Abstract

In this paper, a study is made on selecting existing acoustic models that are trained from native English speech for improving recognition of nonnative English talkers' speech. The problem is addressed from the perspective that foreign accents prevent detailed tri-phone models that are commonly used in high-performance speech recognition systems to match well with these talkers' speech, and therefore an appropriate level of context-dependent acoustic modeling is needed for foreign accent speakers. In this work, model complexity selection is accomplished by empirically choosing a set of model tying thresholds and by using the principle of MDL. An experiment was performed on the Wall Street Journal task on three nonnative English talkers with Chinese accent (276 sentences). Compared to the result obtained from using the models optimized to native English speakers, the best model tying threshold and MDL yielded similar and significant reduction to recognition word errors by 23%.

1. Introduction

Current English speech recognition systems are commonly trained from speech data of native English speakers. Although these systems can work very well for native English speakers, their performances drop dramatically for nonnative speakers. In general, it is difficult to train speech models for each foreign accent due to wide varieties of accent, different proficiency levels of English and limited amounts of available data.

Many efforts were made in improving recognition accuracy of foreign-accent speech. One way is to use general speaker adaptation techniques to adapt speaker-independent (SI) models to the characteristics of a foreign-accent speaker, for example, Using Maximum Likelihood Linear Regression (MLLR) and Maximum *a posteriori* (MAP) estimation. 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 for a foreign-accent speaker than for a native English speaker [1]. Boulis et al. [2] investigated the problem of adaptation to dialect speakers, where speech data of prototype speakers from a target dialect region 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 were needed to form a set of powerful transformation basis.

In the current work, we investigate a different problem in recognition of foreign or dialect accent speech, i.e., fitting an existing native English acoustic model to foreign accent speech by choosing an appropriate level of model complexity. The basic idea is that due to foreign accents, new speakers' speech may not fit well with the detailed acoustic models of English. On the other hand, a certain level of context-dependent modeling is still necessary for capturing contextual effect on phone classification. These two aspects can be addressed simultaneously by model tying, i.e., an intermediate level of acoustic model complexity may best fit a foreign accent speaker, with mono-phone models representing the lower end of the complexity and tri-phone models representing the higher end of the complexity.

Our current focus is on Chinese-accent speakers. Chinese is a monosyllabic language. Each syllable can be divided into two parts, initial and final, where the initial part consists of consonants and the final part consists of vowels. In [3], different context-dependent acoustic models for Chinese speech recognition were compared. The experimental results indicated that although the acoustic realization of initial and final parts were influenced by both left and right contexts, the influence from right context is significantly stronger than that of left context. Since a speaker's native language has a large effect on his/her pronunciation in a second language, for a nonnative speaker, not only the parameters of model, but also the model structure, such as context-dependency structure and model-tying property, may be different from the requirement in native English speech models. In this work, we investigate the problem of finding a proper model complexity from existing acoustic models of native English for nonnative English speakers with Chinese accent, using both empirically chosen model-tying thresholds and the automatic model selection method of

Minimum Description Length (MDL). In section 2 the MDL method is described. In section 3, experimental results are provided. A conclusion is made in section 4.

2. MDL-based Model Selection for Nonnative Speakers

Minimum Description Length (MDL) [4] is an information 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 [5, 6]. Unlike these efforts in which MDL was used to determine model transformation complexity, we use MDL to optimize model complexity by selecting a set of properly tied models from tri-phone models. The procedure of model selection based on MDL is described as the following three 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 1, each terminal node of a tree corresponds to a tri-phone state that is modeled by a single Gaussian density. Each internal node, which is also modeled by a single Gaussian density, corresponds to a state that is tied from all the terminal nodes in its sub-tree, where tying is jointly determined by data volume and phone model similarity [7]. In figure 1, the black nodes correspond to the tied states for a given tying threshold, and these nodes become new terminal nodes after tying, and the corresponding states are referred to as preliminary tied states. The resulting trees are referred to as pruned state-tying trees.

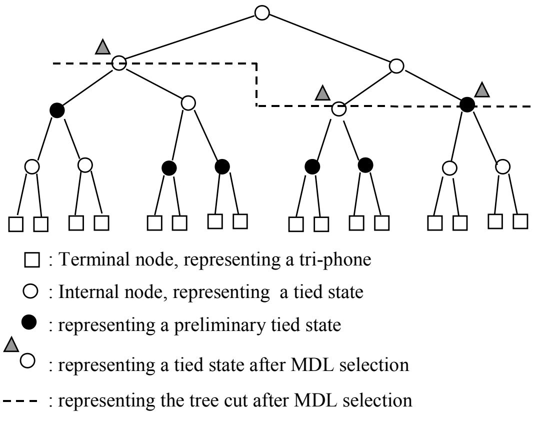
□ : Terminal node, representing a tri-phone
○ : Internal node, representing a tied state
● : representing a preliminary tied state
▲○ : representing a tied state after MDL selection
- - - : representing the tree cut after MDL selection

Figure 1. Illustration of a state-tying tree

In the second step, corresponding to each pruned state-tying tree, a tree copy is constructed. Each of the tree copy nodes is designated a Gaussian mixture density (GMD) model and the parameters are estimated. The resulting trees are referred to as the GMD trees.

In the third step, using speech data from foreign accent speakers, MDL-based tree pruning is applied to the pruned state-tying tree. The procedure is illustrated in figure 2. For a single Gaussian model, its description length (DL) for a given set of data *X* is computed as:

$$DL(X,\mu,\Sigma) = -\sum_{i=1}^{N} \log(\mathrm{N}(x_i;\mu,\Sigma)) + \frac{D}{2}\log(N)$$

where $X = \{x_1, x_2, ..., x_N\}$ is the feature set assigned to this node, μ and Σ are the parameters of the Gaussian density and D is the number of free parameters of the Gaussian density.

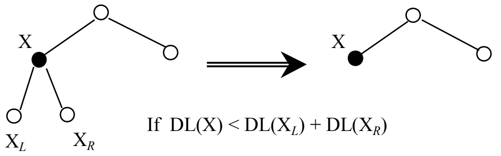

Figure 2. MDL-based tree pruning

For each internal node, if its DL is smaller than the sum of the MDLs of its two children nodes, the two children nodes are pruned off, otherwise, the MDL of this node is assigned as the sum of the MDLs of its two children. This procedure is carried out bottom-up recursively. In the end, the surviving terminal nodes form a cut of the tree that represents an optimum state-tying structure for the foreign accent speakers, and the Gaussian mixture density models in the GMD trees that correspond to these terminal nodes are selected as the acoustic model for these speakers.

In implementing the MDL step, unsupervised decoding is first applied to speech utterances of nonnative speakers to align speech data to each phone unit. Then the phone-labeled speech data are clustered to the terminal nodes of the corresponding pruned state-tying trees.

3. Experiment Result

3.1. Experimental condition

Speech data were taken from the native American English part of WSJ 1.0. The entire set of speaker independent short-term training data was used for acoustic model training by HTK 2.2. The acoustic models were continuous density HMMs. 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 features consisted of 39 components of 12 MFCCs, energy, and their delta and acceleration derivatives. In addition, speech data were collected from three speakers with Chinese-accent and a native English speaker (KEM). In order to match the acoustic environment of WSJ data, the speech data were collected in a quiet office using a close-talking sennheiser microphone. A 92-sentence transcript was selected from WSJ, and each speaker read the transcript once. To further eliminate environment mismatch, Cepstral Mean Normalization (CMN) as implemented in HTK was applied to both training and test data. To verify that a match of acoustic conditions was achieved, the acoustic model trained from WSJ data was applied to the WSJ test set (si_dt_05) and the KEM's speech data and recognition word accuracies of 90.64% and 91.74% were attained, respectively. The similar level of recognition accuracy on the two sets of speech data indicated that environment mismatch between WSJ and our collection condition was basically eliminated.

3.2. Results

3.2.1. *Performance versus model complexity.*

To investigate the effect of model complexity on acoustic modeling for nonnative speakers, several sets of tied tri-phone models with different model complexities were built from the tri-phone models by controlling the state-tying thresholds. In addition, a context-independent mono-phone model set was built that consisted of a total of 130 states. In Table 1, the recognition accuracies of these models are compared for KEM and the nonnative English speech test set, where TP denotes tri-phone and model complexity is measured by the number of states remained after state tying.

Table 1. Word accuracy (%) vs. model complexity

Model set	states	Native speaker (KEM)	Nonnative speakers
Baseline TP	8673	91.74	57.22
Very Low tied TP	3160	91.72	62.16
Low tied TP	2082	91.61	61.97
High tied TP	1172	91.34	63.12
Very high tied TP	809	90.20	64.01
Mono-Phone	130	82.01	58.31

The results indicated that for the native English speaker KEM, the more detailed the model was, the better the performance, but the performance was not very sensitive to the degree of state tying. However, for nonnative speakers, a tri-phone model with an intermediate complexity level appeared to be optimal, and recognition performance degraded significantly when the models were too detailed or too simple.

3.2.2. *Performance versus context dependency.*

Another experiment was conducted to investigate the influence of a speaker's foreign accent on the requirement of context-dependency structure in acoustic modeling. In Table 2, the recognition performances of three acoustic models with different context-dependency structure are compared for the native English talker KEM and the nonnative English speakers. To eliminate the effect of state-tying degree variation on recognition performance, the three models were made to have approximately the same number of states, i.e., 1172, 1181 and 1179 for HTP, HLDP and HRDP, respectively. As expected, for the native English speaker, tri-phone model had the best performance, followed by left and right context-dependent di-phone models, with the latter two having approximately the same recognition accuracy. However, for Chinese accent speakers, the best result was obtained by left context-dependent di-phone model, and it was followed first by the tri-phone model, and then by the right context-dependent di-phone model.

Table 2. Word accuracy (%) vs. context dependency structure

Model set	Context dependency	Native speaker	Nonnative speakers
HTP	Tri-phone	91.34	63.12
HLDP	left dependent di-phone	90.27	66.24
HRDP	right dependent di-phone	89.66	61.67

In order to verify the statistical significance of the results on nonnative English speakers in Table 2, a statistical hypothesis test was made. The difference between the percentages of word accuracy per sentence obtained using model(1) and model(2) is denoted by *c*, and is assumed to be a Gaussian random variable with an unknown variance. The null hypothesis H_0 postulated insignificant superiority of model (1) over model (2), i.e., $E[c] = 0$, and the alternative hypothesis H_1 asserted a positive difference, i.e., E[c] > 0. The test statistic was $q = \bar{c}/(s/\sqrt{n})$, with a *t*-distribution and n-1 degrees of freedom, where $\bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(c_i - \bar{c})^2$, with i indexing the sentences and $n = 276$ [8]. Among the model pairs formed from the three models, HLDP/HTP and HLDP/HRDP led to the rejection of null hypotheses H_0 with the type-I error $\alpha < 0.005$. Therefore, we can state that for Chinese-accent English speakers, the highly tied left context-dependent di-phone models are better than highly tied tri-phone models and highly tied right context-dependent di-phone models.

One possible explanation for the superiority of the highly tied left context-dependent models is that within a CV syllable, left context represents the co-articulation effect of the consonant over the vowel and ignores the effect of the vowel over the consonant, and the Chinese accent speakers may not be able to pronounce the consonant parts as accurately as the vowel parts and hence the left context-dependent model fit well with their pronuniciations. We hope to verify this hypothesis in a later work by studying a large amount Chinese-accented English speech data.

3.2.3. MDL based model selection

By using HTK, 126 state-tying trees for 42 non-silence phones with 3 states each were first built from the WSJ training data and the total number of preliminary tied-state nodes was 8673. The Gaussian mixture density models were then estimated for the GMD trees, with each density have 16 mixture components. A total of 276 sentences from nonnative English speakers were used in MDL-based Gaussian tree pruning, and a total of 987 terminal nodes survived the pruning. Among the 126 MDL-pruned state-tying trees, 23 trees were pruned to have only a single root node. The acoustic models corresponding to the pruned state-tying trees and the models selected by MDL were used for recognition of the same set of nonnative speakers' speech, and a recognition accuracy improvement from 57.22% to 66.90% was obtained, corresponding to an error rate reduction of 23%.

4. Discussion

In this work, we investigated the problem of fitting existing acoustic models that are trained from native English speech to nonnative English talkers' speech without adaptation data. Recognition accuracies versus model complexity and model context dependency structure were studied for nonnative English speech. It was found that for nonnative English speakers, a not-so-detailed acoustic model, which is tolerant of large mismatches in pronunciation while still providing a certain degree of context-dependency information, produced the best result.

We also investigated the problem of unsupervised model complexity selection based on the MDL principle. The result obtained by MDL matches that of the best threshold in state tying, indicating the effectiveness of MDL as a model complexity control method.

An interesting phenomenon we observed from the study is that for speakers with Chinese accent, the best context-dependency structure turned out to be left context-dependent di-phone model. We plan to collect more data from Chinese accent English speakers to analyze in more details this context-dependency phenomenon.

In this study, the acoustic model parameters that were trained from native English speech were not adapted to new speakers. We plan to further investigate the problem of speaker adaptation for nonnative English talkers in conjunction with model complexity optimization.

Acknowledgement

This work was supported in part by the National Science Foundation under the grant NSF EIA 9911095.

References

[1] Zavaliakos, G. Schwartz, R. and Makhoul, J., "Batch, Incremental and Instantaneous Adaptation Techniques for Speech Recognition," *ICASSP 95,* pp. 676-679

[2] Boulis, C. and Digalakis, V., "Fast Speaker Adaptation of Large Vocabulary Continuous Density HMM Speech Recognizer Using A Basis Transform Approach", *ICASSP 00*, vol. 2, pp. 989-992

[3] Ma, B. Huang, T. Xu, B. Zhang, X. and Qu, F., "Context-Dependent Acoustic Models For Chinese Speech Recognition", *ICASSP 96*, vol. 1, pp. 455-458

[4] Rissanen, J., "Universal Coding, Information, Prediction, and Estimation", *IEEE Trans. IT, vol.30, 1984*

[5] Shinoda, K. and Watanabe, T., "Speaker Adaptation With Autonomous Model Complexity Control By MDL Principle", *ICASSP 96*, vol. 2, pp. 717-720

[6] Wang, S. and Zhao, Y., "Optimal On-line Bayesian Model Selection For Speaker Adaptation", *ICSLP 00*, vol. 3, pp. 710-713

[7] Kershaw. D. et al, "The HTK book", http://htk.eng.cam.ac.uk/docs/docs.shtml

[8] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., 3rd edition