A CONTEXT ADAPTATION APPROACH
FOR BUILDING CONTEXT DEPENDENT MODELS IN LVCSR

Xiaoxing Liu, Baosheng Yuan and Yonghong Yan

Intel China Research Center
Email:xiao.xing.liu@intel.com
#601, Kerry Center, 1 GuangHua Road, ChaoYang District, Beijing, 100020, PRC
Tel: (86-10) 8529-8800 Ext. 1811 Fax: (86-10) 8529-8717

Abstract
This paper introduces a new context adaptation framework for building context dependent HMM models in LVCSR. In this new framework, all states of each center phone are clustered into groups by the decision tree algorithm. All the tied states of context dependent HMM models were then derived by adapting the parameters of the multiple-mixture context independent model via data dependent MAP (maximum a posteriori probability) method using the training vectors corresponding to the tied state. An advantage of this approach is that it can maintain a high prediction and classification power given limited training data therefore the model trained in this framework is more reliable than in conventional framework. Experimental results on Wall Street Journal corpora demonstrate that the proposed approach leads to a significant improvement in recognition performance.

1. Introduction

Decision tree state tying based context modeling has become increasingly popular for modeling speech variations in large vocabulary speech recognition[1][2]. In the conventional framework the stochastic classifier for each tied state is trained using Baum-Welch algorithm using the training data corresponding to the specific tied state[3]. However, the context dependent classifiers trained using this method are not so reliable for the training data corresponding to each tied state is limited and model parameters are easily to be affected by undesired sources of information such as speaker and channel differences contained in the training data. To attack this problem, We propose a new context adaptation method to estimate the parameters of context dependent models. In this method, a multiple-mixture context independent model is trained firstly. In decision tree clustering, single mixture Gaussian models were used to establish the state tying. After all the context dependent states are clustered into groups, the clustered states of context dependent model were derived by adapting the parameters of the context independent model via data dependent MAP method using training data corresponding to the tied state. We consider the multi-mixture context independent model as covering the space of more broad speaker and environment classes of speech signal, then adaptation is the context dependent tuning of those speaker and environment classes observed in context’s training speech. Mixture parameters for those speaker and environment classes not observed in the training speech of the specific tied state are merely copied from the context independent model. This means that the model has higher prediction and classification power for the test data from speaker and environment classes unseen or rarely seen in the context’s training data. Experimental results on Wall Street Journal corpora demonstrate that the proposed approaches lead to a significant improvement in recognition performance.

2. Context adaptation method

There are three steps to build context dependent HMM model in the context adaptation framework: 1) Context independent model training 2) Decision tree based state clustering 3) Context dependent model training.

2.1 Context independent model training

We train a multiple-mixture monophone model as the context-independent model. The mixture splitting up process is the same as the baseline system. As we know, the speech signal contains information about various sources. ASR systems need to focus on linguistic information while ignoring other undesired sources of information such as speaker and channel information. One method to achieve this is by training stochastic classifiers with data that contain the various sources of information. For example speaker independence is achieved in ASR by training the recognizer using speech data collected from multiple speakers. We consider the multiple-mixture context independent model as covering the space of more broad speaker and environment classes of speech signal. Then high prediction and classification power can be obtained with the context dependent models adapted from this multiple-mixture
context independent model. A problem in training the context independent model is how to determine the number of mixtures for each state of the monophone model. In our experiment, all states still have the same number of mixtures.

2.2 Decision tree based state clustering

Before decision tree clustering, single mixture triphone models are trained first. One decision tree is constructed for each state of each center phone and all the context dependent states of this phone are clustered into groups by the decision tree algorithm. The node splitting and model selection of decision tree based state tying is a data-driven procedure. It integrates both a priori phonetic knowledge and acoustic similarities derived from data[4]. However, because linguistic information and other unwanted information are bound together in an unknown way in the current spectral-based features, training data corresponding to each tied state maybe not only contain similar acoustic information but also contain similar speaker and channel information. While the parameters of the context dependent model are estimated only with these biased data using Baum-Welch algorithm, the model will not be very reliable and robust.

2.3 Context dependent model training

To make the context dependent model more robust, we derive the tied states of context dependent model by adapting the parameters of the context independent model using the tied state’s training data and data dependent MAP method. The adaptation is a two step estimation process. In the first step we compute the sufficient statistic estimates. It is identical to the first step of Baum-Welch algorithm. In the second step these new sufficient statistic estimates are then combined with the old sufficient statistics from the context independent parameters using a data-dependent mixing coefficient. The data-dependent mixing coefficient is designed so that mixtures with high counts of data from the context dependent clustered state rely more on the new sufficient statistics but also contain similar speaker and channel information. While the parameters of the context dependent model are estimated only with these biased data using Baum-Welch algorithm, the model will not be very reliable and robust.

For each mixture \( i \) of the context dependent cluster state \( j \) and each parameter, a data-dependent adaptation coefficient \( \alpha_i^p \), \( \rho \in \{w, m, v\} \), is used in the above equations. This is defined as

\[
\alpha_i^p = \frac{n_i}{n_i + r^p}
\]

where \( r^p \) is a fixed relevance factor for parameter \( \rho \).

Using a data-dependent adaptation coefficient allows a mixture-dependent adaptation of parameters. If a mixture component has a low probabilistic count, \( n_i \), of new data, then \( \alpha_i^p \to 0 \) causing the deemphasis of the new(potentially undertrained) parameters and the emphasis of the old (better trained) parameters. For mixture components with high probabilistic counts, \( \alpha_i^p \to 1 \), causing the use of the new context-dependent parameters. The relevance factor is a way of controlling how much new data should be observed in a mixture before the new parameters begin replacing the old parameters. This approach should thus be robust to limited training data.

The use of parameter-dependent relevance factors (and hence parameter-dependent adaptation coefficients \( \alpha_i^p \)) further allows tuning of different adaptation rates for the weights, means, and variances. However, we have not found large gain in using parameter-dependent adaptation coefficients. So we use a single adaptation coefficient for all parameters (\( \alpha_i^w = \alpha_i^m = \alpha_i^v = \alpha_i \) / \((n_i + r)\)) with a same relevance factor \( r=2 \). This means for mixture component with probabilistic counts larger than 2, the new parameters are emphasized otherwise the old parameters are emphasized.

Since the adaptation is data dependent, not all Gaussians in the context independent are adapted during context dependent model training. In our experiment[TABLE 1],
there are usually about 1/10 of the Gaussian mixtures of the final context dependent model not updated from the context independent model and the parameters of these mixtures are merely copied from the context independent model. We can use this factor to reduce model storage requirements.

3. Experimental results

The performance of the proposed context adaptation method was evaluated on the official NOV92 (si_et_05, si_et_20) for the open 20k vocabulary of the Wall Street Journal (WSJ) task. Twelve mel-cepstral coefficients plus their first and second order time derivatives were used as acoustic features. The cepstral mean for each sentence was calculated and removed. All HMMs have three emitting states and a left-to-right topology. We try the number of mixtures for each state from 24 to 56 and get the best performance with 4645 tied states and 48 mixtures for each state (Fig. 1). On NOV92 test set we cut the word error rate by 16.2% from 11.7% to 9.8%. The baseline system was trained using Baum-Welch algorithm after decision tree based state clustering. There are 6080 states in the baseline system and the number of gaussian mixtures for each state is 12. The size of the system using context adaptation method is larger than the baseline system and the decoding is more time consuming correspondingly but the recognition performance is improved largely. The word error rates for all testing speakers are cut down [Table 2]. Some experiments with variable number of mixtures for each state have not gotten significant performance improvement. Number of the tied states varies from 4k to 8k considering the number of mixtures for each state. Decoding was done using a one-pass trigram tree-search decoder, and within word triphone models. The standard SI-284 (60 hour of speech) training data sets were used to train the WSJ models. The language models used in the experiments are the standard trigram language models provided by NIST for the WSJ corpus.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Baseline system</th>
<th>Context Adaptation</th>
<th>Performance Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>11.7%</td>
<td>9.8%</td>
<td>16.2%</td>
</tr>
<tr>
<td>440</td>
<td>10.5%</td>
<td>8.5%</td>
<td>19.0%</td>
</tr>
<tr>
<td>441</td>
<td>19.8%</td>
<td>16.6%</td>
<td>16.1%</td>
</tr>
<tr>
<td>442</td>
<td>11.7%</td>
<td>10.2%</td>
<td>12.8%</td>
</tr>
<tr>
<td>443</td>
<td>12.7%</td>
<td>11.9%</td>
<td>6.3%</td>
</tr>
<tr>
<td>444</td>
<td>13.3%</td>
<td>10.4%</td>
<td>21.8%</td>
</tr>
<tr>
<td>445</td>
<td>6.5%</td>
<td>4.8%</td>
<td>26.15%</td>
</tr>
<tr>
<td>446</td>
<td>10.2%</td>
<td>9.4%</td>
<td>7.84%</td>
</tr>
<tr>
<td>447</td>
<td>9.6%</td>
<td>7.0%</td>
<td>27.08%</td>
</tr>
</tbody>
</table>

**TABLE 2**
WER per speaker comparison with baseline system on NOV92 test set for WSJ task

Table 2 shows the performance improvement using context adaptation method. The number of mixtures for each state is 48 and the number of tied states is 4645. For all speakers, there is significant performance improvement.

![Fig. 1](image.png)

**Fig. 1**
Results with different number of mixtures for each state on NOV92 test set for WSJ task

Fig. 1 illustrates the performance in the context adaptation framework with different number of mixtures per state. As a comparison, the number of mixtures for each state in the baseline system is 12 and WER is 11.7%. There are performance improvement in different extent with these different number of mixtures for each state. We get the best performance with 48 mixtures.

4. Discussion

One of the ideas behind our method is that we consider each tied state as a context class. Every context class contains its own context information. To improve the prediction power, we estimate the distribution parameters of each classifier not only from its own
training data but also the context independent model. So this data-dependent MAP based context adaptation method can be thought as a smooth technique. Data-dependent MAP adaptation method is used widely in speaker recognition. But it is interesting that this adaptation method is also useful in context modeling. The same problem for speaker recognition and context modeling is that different kinds of information sources are bound together in an unknown way in the spectral-based features but only a specific information source is what we wanted. Another problem is that more parameters (in our experiment the size of the best model is almost three times than baseline) are needed in the system using context adaptation method and the decoding is more time consuming.

5. Summary

In this paper we propose a new context adaptation framework for building context dependent HMM models in LVCSR. Unlike the conventional Baum-Welch training method, we derive the tied states of context dependent model by adapting the parameters of the context independent model using training data corresponding to the tied state and MAP method. Using a data-dependent adaptation coefficient allows a mixture-dependent adaptation of parameters. An advantage of this approach is that it can maintain a high prediction and classification power given limited availability of training data and the model trained in this framework is more reliable than in conventional framework. Experimental results on Wall Street Journal corpora demonstrate that the proposed approaches lead to a significant improvement in recognition performance.

6. Reference