In this paper, an approach of dynamic threshold setting via Bayesian Information Criterion (BIC) in HMM training is described. The BIC threshold setting is applied to two important applications. Firstly, it is used to set the thresholds for decision tree based state tying, in place of the conventional approach of using a heuristic constant threshold. Secondly, it is applied to choosing the number of Gaussian mixture at state mixing-up stage. Experimental results on LVCSR Chinese dictation task indicate that BIC can dynamically set thresholds for cluster splitting according to the underlying complexity of the cluster parameters. Also significant performance improvement is achieved with the dynamic BIC threshold setting.

1. INTRODUCTION

Typical HMM training is to start with a simple set of single Gaussian context-independent phone models and then iteratively refines them by expanding them to include context-dependency and use multiple mixture component Gaussian distributions. Parameter sharing for triphone models cloned from monophone models is conducted according to a decision tree built from phonetic knowledge.

HMM training involves many problems of finding a compact one among a set of candidate models to describe a given data set, which is treated in statistics as the problem of model identification. In decision tree growing, once the best question is found for a tree node, whether or not splitting the node equals to evaluating the modeling of a set of speech samples assigned to this node with double number of parameters. The same is true with decision tree node merger. The state mixing-up iteration is to choose the number of Gaussians for each HMM state according to the state occupancy. In standard HMM training, the above two model identification problems are dealt with by constant threshold methods.

In this paper, the Bayesian information criterion (BIC), a model selection criterion in the statistics literature, is applied to decision tree state tying and Gaussian mixture splitting. The constant thresholds are replaced by dynamic BIC thresholds for both of them.

This paper is organized as follows: section 2 describes Bayesian information criterion in statistics literature; section 3 present BIC thresholds setting for decision tree state tying; section 4 present BIC threshold setting for Gaussian mixture splitting. In section 5, we present some experiment results on LVCSR Chinese dictation task.
whose leaf nodes form a partition of the acoustic phonetic space, and under certain constraints, the log likelihood of the tree is maximized.

The standard one-step greedy growing algorithm is a top-down process. It grows the terminal nodes of the tree one step at a time. At each step, it searches for the best terminal node to grow and the best question to apply so that it leads to a maximum increase of the log likelihood by splitting the node into two children nodes. In other words, it is to find \((t^*,q^*)\) such that

\[
\begin{align*}
(t^*,q^*) = \arg \max_{(t,q)} \left( L_{yes}(t,q) + L_{no}(t,q) - L(t) \right)
\end{align*}
\]

Where \(L_{yes}(t,q)\) and \(L_{no}(t,q)\) are the log likelihood of yes/no split of node \(t\) according to question \(q\). In addition, model identification in decision tree state tying is applied by requiring the node split satisfy the condition:

\[
L_{yes}(t,q) + L_{no}(t,q) - L(t) > \Delta
\]

Where \(\Delta\) is a constant threshold determined by experiments. The constant threshold approach has a relation to the AIC criterion, but it does not depend explicitly on the number of data samples at the tree node.

The tree splitting process in decision tree state tying can be viewed from the statistical hypothesis testing framework for testing the number of components in the mixture. BIC is considered as a more conservative criterion than AIC and leans more than AIC towards lower dimension models. Once the best question has been found for node \(t\), whether or not split tree node \(t\), from the BIC standpoint, becomes

\[
L_{yes}(t,q) + L_{no}(t,q) - L(t) > \alpha \cdot \log(Q_t) \cdot V
\]

Where \(Q_t\) is the state occupancy in node \(t\), \(V\) is the feature vector size.

4. BIC THRESHOLD SETTING FOR GAUSSIAN MIXTURE SPLITTING

Once HMM states with single Gaussian mixture have been tied by decision tree, HMM training goes into the clustering training phase to choose the number of Gaussians for each state by iterative increasing the state mixture components by a specified number at a time. It is well known that too few Gaussians do not give sufficient model complexity whereas too many leads to overtraining. A common heuristic solution of this problem is the thresholding method. According to the number of samples belonging to the HMM state in the training data, one chooses the number of Gaussians proportionally. Namely to say, state \(s\) with less occupancy than a constant threshold \(\Delta\) will not be upmixed.

\[
\gamma_s > \Delta
\]

Here the BIC criterion is to adaptively choose the number of Gaussians according to the underlying complexity of the HMM state. Thus, if the log likelihood for state \(s\) at \(M\) mixtures and \(M'\) mixtures does not satisfy condition

\[
\log(L_M) - \log(L_{M'}) > \Delta_{BIC}(s)
\]

, the Gaussian components for state \(s\) can not be increased further more. The BIC threshold \(\Delta_{BIC}\) can be expressed as

\[
\Delta_{BIC} = \lambda \cdot \left( \log(\gamma_M) \cdot M - \log(\gamma_{M'}) \cdot M' \right) \cdot (2V + 1)
\]

where \(\gamma_M\) and \(\gamma_{M'}\) are state occupancy for state \(s\) at \(M\) and \(M'\) mixtures.

5. EXPERIMENTAL RESULTS

The BIC threshold setting was evaluated on the LVCSR Chinese dictation task. 12 mel-cepstral coefficients plus their 1st and 2nd order time derivatives were used as acoustic features. Phonetic decision tree states tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones. The final triphone HMMs were built based on the tied states from the clustering. Decoding was done using a one-pass trigram tree-search decoder, and within word triphone models. The training set contains 204,153 utterances from 516 speaker, totally 65,547,305 frames. The test set contains 110 utterances from 11 speaker (5 female and 6 males). The language model is trigram and the vocabulary is 51k. The baseline system with 9990 HMMs, 6007 tied states and 12 Gaussian mixture per state, totally 72076 mixtures achieves a Word Error Rate of 10.7 on this test set.

Table 1: WER for baseline system

<table>
<thead>
<tr>
<th>Speaker</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01</td>
<td>14.4</td>
</tr>
<tr>
<td>F02</td>
<td>9.5</td>
</tr>
<tr>
<td>F03</td>
<td>8.4</td>
</tr>
<tr>
<td>F04</td>
<td>15.1</td>
</tr>
<tr>
<td>F05</td>
<td>9.2</td>
</tr>
<tr>
<td>M01</td>
<td>21.4</td>
</tr>
<tr>
<td>M02</td>
<td>6.8</td>
</tr>
<tr>
<td>M03</td>
<td>7.9</td>
</tr>
<tr>
<td>M04</td>
<td>15.2</td>
</tr>
<tr>
<td>M05</td>
<td>4.9</td>
</tr>
<tr>
<td>M06</td>
<td>9.0</td>
</tr>
<tr>
<td>Average</td>
<td>10.7</td>
</tr>
</tbody>
</table>

The WER listed in table 1 was gotten with the constant thresholds in decision tree based state tying and Gaussian mixture splitting.

5.1 BIC THRESHOLD SETTING FOR DECISION TREE BUILDING

To do a fair comparison with baseline system, we set the BIC splitting penalty to 2.0 and merger penalty to 9.6, such that we got a state number of 6005, nearly same with our baseline system. Also the following clustering training and mixing-up procession are same with the baseline system. On our standard test set, it gives 5% relative improvement. The real reason for us to use BIC thresholds for decision tree is to build a tree with less number of leaves, such as a state number of 3k, without performance loss. Table 2 is the result for comparison experiment with 6k state number.
6. SUMMARY

In this paper, the BIC threshold setting is described and applied to two applications. Firstly, it is used to set the thresholds for decision tree based state tying, in place of the conventional approach of using a heuristic constant threshold. Secondly, it is applied to choosing the number of Gaussian mixture at state merging stage. Experimental results on LVCSR Chinese dictation task indicate that BIC can dynamically set thresholds for cluster splitting according to the underlying complexity of the cluster parameters. Also significant performance improvement is achieved with the dynamic BIC threshold setting.

REFERENCE


