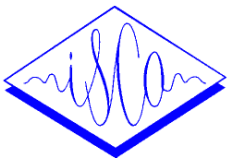


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

DYNAMIC THRESHOLD SETTING VIA BAYESIAN INFORMATION CRITERION (BIC) IN HMM TRAINING

Ying Jia, Yonghong Yan and Baosheng Yuan
Intel China Research Center
Beijing Kerry Center, Beijing, 100020 PRC
Email: {ying.jia, yonghong.yan, baosheng.yuan}@intel.com

ABSTRACT

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.

2. BAYESIAN INFORMATION CRITERION (BIC)

The problem of model identification is to choose a compact one among a set of candidate models to describe a given data set. The candidate models differ from each other in the number of parameter. It is evident that when the number of parameters in the model is increased, the likelihood of the training data is also increased; however, when the number of parameters is too large, this might cause the problem of overtraining. Several criteria for model selection have been introduced in the statistics literature. The Bayesian information criterion (BIC) is a likelihood criterion penalized by the number of parameters in the model. In detail, let $X = \{x_i, i = 1, \Lambda\ , N\}$ be the data set we are modeling; let $\mathrm{M} = \{M_j, j = 1, \Lambda\ , J\}$ be the candidate parameteric models. BIC is to select model j that maximizes

$$\log L_j(x_1, x_2, \Lambda\ , x_N) - \frac{l}{2} k_j \log N$$

Where k_j is the number of parameters in model j and $\log L_j(x_1, \Lambda\ , x_N)$ is the log likelihood of model j given data sample $(x_1, \Lambda\ , x_N)$. BIC criterion is derived from the integration of the marginal likelihood of model parameters with respect to data sample set.

BIC is closely related to other penalized likelihood criterions such as AIC and RIC, and BIC has theoretical advantages because of its connection with Bayesian procedures. Application of BIC to speech recognition have been introduced in IBM ([1]) and Bell Labs ([2]). S. S. Chen investigated the applications of BIC in speaker clustering and Gaussian mixture modeling [1]. Wu Chou proposed a penalized BIC for decision tree state tying [2]. Similar with their work, we use BIC to dynamicly set thresholds for decision tree state tying and state mixing-up.

3. BIC THRESHOLDS SETTING FOR DECISION TREE STATE TYING

In decision tree based state tying, a set of phonetic questions characterizing the phonetic properties of the context is selected. These phonetic questions are related to acoustic phonetic properties of the phonemes, such as a front vowel, nasal, fricative etc. each question divides the acoustic space into two parts depending on the yes/no answer to the question. The phonetic decision tree based state tying is to find a decision tree

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

$$(t^*, q^*) = \arg\max_{(t,q)} \left(L_{yes}(t,q) + L_{no}(t,q) - L(t)\right)$$

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 Δ 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

$$L_{yes}(t,q) + L_{no}(t,q) - L(t) > \alpha \cdot \log(g_t) \cdot V$$

Where g_t is the state occupancy in node t, α is the penalty factor, 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 does 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 Δ will not be up-mixed.

$$g_s > \Delta$$

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 Δ_{BIC} can be expressed as

$$\Delta_{BIC} = \lambda \cdot \left(\log(g_M) \cdot M - \log(g_{M'})M'\right) \cdot (2V + 1)$$

where g_M and $g_{M'}$ 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

Speaker	WER
F01	14.4
F02	5.5
F03	8.4
F04	15.1
F05	9.2
M01	21.4
M02	4.8
M03	7.9
M04	15.2
M05	4.9
M06	9.0
Average	10.7

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

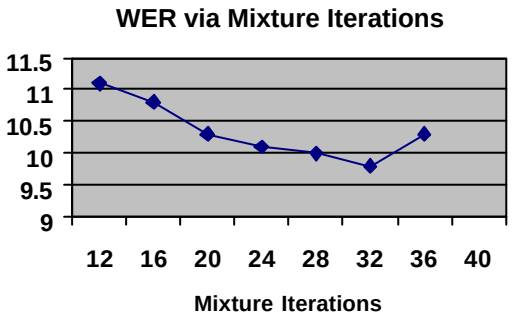
To do a fair comparison with baseline system, we set the BIC splitting penalty to 2.0 and merger penalty to 9.6, such that we got a state number of 6005, nearly same with our baseline system. Also the following clustering training and mixing-up procession are same with the baseline system. On our standard 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.

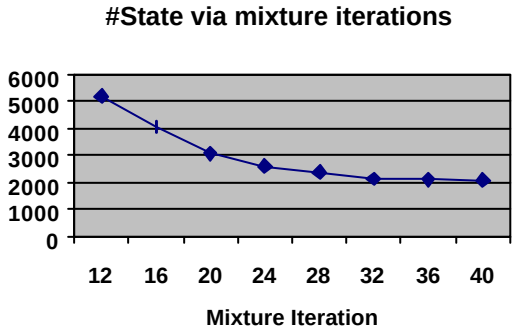
Table 2: Splitting Penalty = 3.0, Merger Penalty = 9.6

Speaker	WER
F01	13.9
F02	3.0
F03	7.2
F04	15.9
F05	5.5
M01	16.0
M02	9.5
M03	8.3
M04	18.3
M05	4.4
M06	8.0
Average	10.1

5.2 BIC THRESHOLD SETTING FOR GAUSSIAN MIXTURE SPLITTING

Our observation with constant thresholds in mixture increase is no performance improvement when the number of mixture beyond 12. From figure 1 and 2, we can see that the WER achieves minimum at mixture 32. After this minimum point, the number of states whose mixture number can be increased further seems stay 2000, and the WER starts to go up while the number of mixture is increasing. The minimum WER is 9.8% at mixture 32. So the BIC mixing-up gives nearly 9% performance improvement.





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 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.

REFERENCE

[1] S. S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, P. Olsen, Recent Improvements to IBM's speech recognition system for Automatic Transcription of Broadcast News. Proc. Of DARPA Speech Recognition Workshop, February 28-March 3, 1999. Herndon, Virginia.

[2]. Wu Chou, Wolfgang Reichl, "Decision tree State tying Based on Penalized Bayesian Information Criterion", ICASSP99, Volume: 1 , 1999 , Page(s): 345 -348.

[3] Scott Shaobing Chen; Gopalakrishnan, P.S. Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on Volume: 2 , 1998 , Page(s): 645 -648 vol.2

[4] Sakai, H. An Application of a BIC-type method to harmonic analysis and a new criterion for order determination of an AR process. Acoustics, Speech and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on Volume: 38 6 , June 1990 , Page(s): 999 -1004.