HIGH RESOLUTION DECISION TREE BASED ACOUSTIC MODELING BEYOND *CART*

Wu Chou and Wolfgang Reichl

Dialogue System Research Department, Bell Laboratories, Lucent Technologies

600 Mountain Avenue, Murray Hill, NJ07974, USA
wuchou@research.bell-labs.com, reichl@research.bell-labs.com

ABSTRACT

In this paper, an m-level optimal subtree based phonetic decision tree clustering algorithm is described. Unlike prior approaches, the m-level optimal subtree in the proposed approach is to generate log likelihood estimates using multiple mixture Gaussians for phonetic decision tree based state tying. It provides a more accurate model of the log likelihood variations in node splitting and it is consistent with the acoustic space partition introduced by the set of phonetic questions applied during the decision tree state tying process. In order to reduce the algorithmic complexity, a caching scheme based on previous search results is also described. It leads to a significant speed up of the m-level optimal subtree construction without degradation of the recognition performance, making the proposed approach suitable for large vocabulary speech recognition tasks. Experimental results on a standard (Wall Street Journal) speech recognition task indicate that the proposed m-level optimal subtree approach outperforms the conventional approach of using single mixture Gaussians in phonetic decision tree based state tying.

1. INTRODUCTION

Decision tree based acoustic modeling has become increasingly popular in speech recognition, in particular for modeling using context dependent model units. In this approach, the acoustic phonetic knowledge of the target language can be effectively incorporated in the model according to a consistent maximum likelihood framework. The statistical framework of decision tree in acoustic modeling provides two major advantages over the previous rule or bottom up based approaches. First, the classification and prediction power of the decision tree allows to synthesis model units or contexts which do not occur in the training data. Second, the splitting procedures of decision tree design provides a way of maintaining the balance between the model complexity and the number of parameters that can be robustly estimated from the limited amount of training data.

In acoustic modeling using decision tree based state tying, it is typical that a phonetic decision tree is constructed for each phoneme or for each state of the phoneme. The phonetic decision tree is a binary tree, not necessarily balanced, in which a yes/no question about the phonetic context is attached to each node of the tree. The structure of the phonetic decision tree leads to a partition of the acoustic space, and acoustic phonetic events are classified and clustered as the leaves of the tree. When questions concerning phonetic properties are applied in decision tree building, the acoustic space partition is constrained. The impurity function used in the phonetic decision tree clustering design is based on the negative log likelihood function. The tree construction is a top down data driven process based on a one-step greedy tree growing algorithm. At each step, a one-step node split is made if it can lead to the largest increase of the log likelihood. This greedy classification and regression tree (CART) algorithm [1] is quite efficient in decision tree based state tying [6].

Recently, there are many efforts to improve the phonetic decision tree based approach in acoustic modeling [2] [3]. Two problems are of particular interest. One is the tree growing and node splitting problem and it concerns the issue of how to find an optimal node split given the particular parametric form of the impurity function. Another one is the parametric modeling problem of the distribution of the cluster during the process of decision tree node splitting. For phonetic decision tree based acoustic modeling, these two problems are closely related. The optimal node splitting problem is a problem of finding the best node split, and the parametric modeling is a problem of providing an appropriate metric which defines the quality of the split. In general, construction of a globally optimal decision tree is a computationally intractable problem and early efforts of using lookahead search in node splitting did not improve the model performance [3]. On the other hand, the parametric forms of distributions used in decision tree node splitting are often based on single mixture Gaussian distributions, although more accurate multiple mixture Gaussian distributions are used in the final acoustic model. This disparity is due in part to the computational complexity in the decision tree clustering process. The multiple mixture Gaussian distribution for each tree node needs to be re-estimated from the data, whereas the parameters of the single mixture Gaussian distribution can be derived from the cluster members without going back to the training data. Earlier results of using multiple mixture Gaussians in decision tree state tying were disappointing [2] and did not indicate any advantageous in terms of recognition performance.

In this paper, we describe a decision tree growing algorithm based on multiple Gaussian mixtures for decision tree state tying based acoustic modeling in large vocabulary speech recognition. It incorporates an *m-level optimal subtree* procedure in classification and regression tree. This leads to a multiple mixture Gaussian distribution parameterization of the cluster distribution, which is consistent with the acoustic space partition of the phonetic questions. Experimental results on large vocabulary speech recognition tasks indicate that the proposed approach is not only computationally feasible but also outperforms the conventional single mixture Gaussian based approach.

2. M-LEVEL OPTIMAL SUBTREE BASED TREE GROWING ALGORITHM

In a typical procedure of phonetic decision tree based state tying, a decision tree for each state of each base phone is constructed from its sample occurrences in the training data. For a tri-phone based system, the phonetic information of each occurrence of that state in the training data is encoded by its left and right phonetic contexts to which it occurs. A set of phonetic questions $\{Q(i) | i = 1,\dots,N\}$ 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 $Q(i)$ divides the acoustic phonetic space into two parts $A_{Q(i)}$ and $\bar{A}_{Q(i)}$ depending on the yes/no answer to the question. The acoustic space partition based on the phonetic questions is formed by the finite intersects of $\{A_{Q(i)}, \breve{A}_{Q(i)} | i = 1,\dots,N\}$. 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 CART 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 $(\bar{t}, \bar{q})$ such that

$$(\bar{t}, \bar{q}) = \operatorname{argmax}_{(t,q)}[L(t,q,y) + L(t,q,n) - L(t)],$$

and

$$L(\bar{t}, \bar{q}, y) + L(\bar{t}, \bar{q}, n) - L(t) \;>\; \Delta ,$$

Where $L(t,q,y)$ and $L(t,q,n)$ are the log likelihood of yes/no split of node t according to question q and Δ is a threshold which controls the node splitting.

2.1. The Tree Growing Algorithm

From the above decision tree formulation, two issues are worth mentioning. First, the quality of the decision tree based state tying depends on the parametric form of the distribution used in evaluating $L(t)$ which should approximate as close as possible to the multiple mixture Gaussian distribution used in the final model. This suggests that using multiple mixture Gaussian distribution, instead of the single mixture Gaussian, to evaluate $L(t)$ might be advantageous. Secondly, the estimated log likelihood of the tree node t should be honest and should not over estimate $L(t)$. When multiple Gaussian distribution is used to split node t, each individual mixture used in $L(t)$ should be supported by the partition of the phonetic questions of the decision tree in order to ensure an honest estimate.

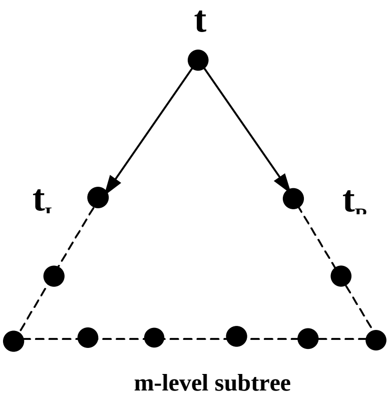

Figure 1: Diagram of an m-level subtree.

In order to solve these two issues, we describe a phonetic decision tree growing algorithm which is based on an honest estimate of multiple mixture Gaussian distributions in node splitting for phonetic decision tree based state tying. The key idea in this approach is to use an m-level optimal subtree during the node split which is an extension of the conventional one-step greedy CART algorithm. Let $T(t,m)$ denote an m-level subtree, with root node t and a maximum level of m. The log likelihood of the m-level tree $T(t,m)$ is defined to be

$$L(\tilde{T}(t,m)) = \sum_{t' \in \tilde{T}} L(t') ,$$

which is obtained by summing the log likelihood over all its leaves. Fig 1. illustrates an m-level subtree from node t .

The proposed m-level optimal subtree based decision tree growing algorithm consists of the following steps:

(1) If t is the root node, grow an m-level optimal subtree $T(t,m)$, not necessarily balanced, using the phonetic questions. Split the node t into nodes t_L and t_R .

(2) Update the log likelihood of t_L and t_R to be

$$L(t_L) = L(\tilde{T}_L(t,m)) \quad \text{and} \quad L(t_R) = L(\tilde{T}_R(t,m)) ,$$

where $\tilde{T}_L$ and $\tilde{T}_R$ are left and right branches of the m-level optimal subtree $\tilde{T}(t,m)$. The updated log likelihood of $L(t_L)$ and $L(t_R)$ is modeled by 2^{m-1}-mixture Gaussians because they are the sum of single mixture Gaussians from the corresponding m-level optimal subtree leaves.

(3) For each terminal tree node t with cluster sample count greater than the minimum sample count threshold, grow an m-level optimal subtree $\tilde{T}(t,m)$ and split the node t into nodes t_L and t_R provided that

$$L(\tilde{T}(t,m)) - L(t) \;>\; \Delta_m ,$$

Update the log likelihood $L(t_L)$, $L(t_R)$ by performing step (2).

(4) The algorithm stops if there is no terminal node that satisfies step (3) and the minimum sample count constraint.

It should be noted that both $L(\tilde{T}(t,m))$ and $L(t)$ are based on multiple mixture Gaussian distributions when t is not the root node. This is because after step (1), the likelihoods of the nodes are updated by its likelihood from the m-level optimal subtree which is a combination of Gaussians from the corresponding tree leaves. The proposed approach utilizes an m-level optimal subtree to obtain an honest estimate of the multiple Gaussian distribution for node splitting. Although the m-level optimal subtree $T(t,m)$ is derived from the phonetic questions and using single mixture Gaussians, the leaves of the m-level subtree $T(t,m)$ introduce a multiple mixture Gaussian parameterization of the log likelihood of the tree node t. In addition, the multiple mixture Gaussian parameterization of $L(t)$ obtained from the proposed approach is honest in the sense that all its mixtures are supported on the partition of the phonetic questions of the decision tree and it will not give an over estimate of $L(t)$. The conventional one-level greedy tree growing algorithm is a special case in the proposed approach when optimal subtree level $m=1$.

However, there is a fundamental difference between the proposed approach and the look-ahead search technique used in decision tree based state tying. The look-ahead search is to find a more accurate estimate of the log likelihood increase when split node t. In other words, it uses a refined estimate of $L(t_L)$ and $L(t_R)$ but does not change the parametric distribution of t nor the value of $L(t)$. In the proposed approach, a m-level subtree is used as a mean to introduce honest multiple mixture Gaussian parametric distribution for node t, which is used consistently for all related log likelihood estimates in node splitting, $L(t)$, $L(t_L)$ and $L(t_R)$.

2.2. A Scheme for Algorithmic Complexity Reduction

Although the greedy tree splitting algorithm based on single mixture Gaussian distribution may not be accurate enough, it is computationally efficient. For single mixture Gaussian, the log likelihood of a cluster at tree node t in segmental based approach [4] is given by

$$L(t) = -\frac{n(t)}{2}(\log|\Sigma(t)| + D(\log(2\pi)+1)),$$

where $n(t)$ is the number of samples in the cluster, $\Sigma(t)$ is the sample covariance matrix and D is the dimensionality of the data vector. The cluster log likelihood $L(t)$ can be calculated by using the already available information from the untied state clusters without additional access of the data. As a consequence, the phonetic decision tree based state tying only constitutes a small portion of the computation in acoustic model building [4]. This may not be the case when multiple mixture Gaussian distributions are used in node splitting. Although the proposed approach does not make a direct estimation of the multiple mixture Gaussian distribution in decision tree state tying, step (2) of growing an m-level optimal subtree can become expensive. Given a set of N phonetic questions $\{Q(i)|i=1,...,N\}$, finding a 2-level optimal subtree $T(t,m)$ involves in an order of $N \times N$ operations of node splitting. The algorithmic complexity grows exponentially with the subtree level m, making it unfeasible for application in large vocabulary speech recognition.

In order to reduce the algorithmic complexity, we propose a scheme that is based on caching the top K best second level questions of the previous search in a shortlist table. The shortlist of the top K second level phonetic questions associated with left and right branches used to construct m-level optimal subtree $T(t,m)$ is attached to the new children nodes t_L and t_R. In the future split, the m-level subtree constructed for t_L and t_R will be restricted to questions in the shortlist. For two-level optimal subtree, this reduces the algorithmic complexity of doing node splitting from $N \times N$ to $K \times N$, where K is the depth of the shortlist. This approximation is reasonable in two senses. First, the m-level subtree constructed with this cache scheme is always superior than the subtree constructed from one-step greedy algorithm. Second, the top K best first-level questions for t_L and t_R derived from the m-level optimal subtree of their parent node t contain at least K m-1 level questions and provides a good coverage of the first-level question used for the m-level optimal subtrees of t_L and t_R. The use of the caching scheme makes it practical to use the proposed m-level optimal subtree approach for phonetic decision tree based state tying in large vocabulary speech recognition tasks. In addition, other more aggressive caching schemes can also be used which will lead to further reduction of the algorithmic complexity. In our speech recognition experiments, we observe significant speed up without recognition performance degradation.

3. EXPERIMENTAL RESULTS

The proposed m-level optimal subtree based decision tree clustering algorithm was evaluated and compared on large vocabulary Wall Street Journal (WSJ) tasks. In our speech recognition system, 12 mel-cepstral coefficients and the normalized energy plus their 1st and 2nd order time derivatives were used as speech recognition features. The cepstral mean for each sentence was calculated and removed. All phone models have three emitting states and a left-to-right topology. In our acoustic model training procedure, training data are first aligned into segments corresponding to the state of the hidden Markov models (HMMs). The untied system was constructed based on a two level robust clustering approach [4]. First the parameters of untied models with number of training samples exceeding a threshold were estimated. We used a minimum of 10 samples in our experiments. A robust clustering algorithm [4] was performed on rare triphones whose occurrences in the training data were below the minimum sample counts. The proposed m-level subtree algorithm was then used to form the decision tree state tying. It clusters the equivalent sets of context dependent states in the untied system and constructs the mapping table for unseen triphones. The final models were built by using the segmental k-means based parameter estimation method for all tied states. The number of mixture Gaussians in the final model for each of the tied states varies from 4 to 12 according to the amount of training data. Decoding is based on a one-pass N-gram decoder without adaptation, in which the search was conducted on a layered self-adjusting decoding graph [5].

For the WSJ task, both SI-84 and SI-284 training copra were used. The lexicon was generated automatically using a general English text-to-speech system with 41 phones. The language models used in the experiments were the standard trigram language models provided in the WSJ corpus. The SI-84 training data (7200 sentences) contains about 8600 triphones with more than 10 examples and about 8000 triphones with a frequency count of less than 10 occurrences. The full WSJ training corpus (SI-284) contains 38,700 sentences. Even with more training data, there are about 10,000 of the 24,000 observed triphones occurring less than 10 times. These rare triphones are grouped into 1029 triphone clusters to ensure that these triphone clusters have enough training samples in order to make the estimation of the parameters for state clustering robust.

Model		**NOV92**			
		5k-closed		20k-open	
		std.	two-level	std.	two-level
SI-84	**GI**	5.0%	4.5%	12.8%	12.2%
SI-84	**GD**	4.5%	4.4%	12.1%	11.8%
SI-284	**GI**	3.0%	2.9%	9.8%	9.5%
SI-284	**GD**	3.0%	2.9%	9.8%	9.4%

Table 1: Word error rates for two-level optimal decision trees (trigram LM).

During the process of the proposed m-level optimal subtree based decision tree state tying, a separate minimum sample count for nodes in m-level subtree is used. The minimum sample count used for the m-level subtree should be based on the sample count needed for robust estimation of the single Gaussian, whereas the minimum sample count for the node in decision tree has to control the model complexity. In addition, the second sample count should be much less than the regular node sample count in order to grow m-level optimal subtrees near the final decision tree leaves. In our experiments, a minimum sample count of 100 was used for the regular decision tree node and a minimum sample count of 20 was used for the m-level optimal subtree. Since our training of the phonetic decision tree state tying based acoustic model is a segmental based approach and uses separate phonetic decision trees for different phones and states, the model training, including the decision tree clustering process, is therefore highly parallel. In our experiments, the number of phonetic questions is $N = 208$ and we used $K = 30$ for the depth of the shortlist in the proposed caching scheme. It is only 15% of the computational complexity of the original algorithm. The testing results are tabulated in Table 1. The experiments were performed on the Wall Street Journal task for both 5K close vocabulary (nov_92_5k_close) and 20K open vocabulary (nov_92_20k_open) tests.

The experimental results confirmed that using multiple mixture Gaussian distributions in decision tree based state tying is advantageous over the conventional single mixture Gaussian based approach. In Table 1, GI represents the gender independent model and GD represents gender dependent model. The std column lists recognition results of the model obtained from the conventional single mixture Gaussian based decision tree state tying approach, and the two-level column contains the recognition results obtained from the proposed m-level optimal subtree based approach where the subtree level $m = 2$. Although the performance improvement varies from case to case with the maximum improvement of 10% for 5K GI case, using multiple mixture Gaussian in decision tree based state tying is advantageous in our approach. This is the first time to our knowledge that decision tree state tying using multiple mixture Gaussians is proved to be useful and efficient enough for acoustic modeling in large vocabulary speech recognition.

4. SUMMARY

In this paper, an m-level optimal subtree based phonetic decision tree clustering algorithm was described. Unlike prior approaches, the m-level optimal subtree construction in the proposed approach is to generate honest log likelihood estimate using multiple mixture Gaussians for phonetic decision tree based state tying. It is different from other conventional look-ahead search in that the multiple mixture Gaussians were used consistently for all related log likelihood estimates in decision tree node splitting, In order to reduce the algorithmic complexity, a caching scheme based on previous search results was also described. It led to a significant speed up of the m-level optimal subtree construction without degradation to the recognition performance. Experimental results on Wall Street Journal speech recognition task indicated that the proposed m-level optimal subtree approach is advantageous comparing to the approach using single mixture Gaussian in phonetic decision tree state tying.

5. REFERENCES

1. Breiman, L., Friedman, J., Olshen, R., Stone, C. “Classification and Regression Trees,” Chapman & Hall, 1984

2. Nock, H.J., Gales, M. and Young, S. “A Comparative Study of Methods for Phonetic Decision Tree State Tying”, Eurospeech’97, page 111—114, 1997.

3. Lazarides, A., Normandin, Y. and Kuhn, R. “Improving Decision Trees for Acoustic Modeling,” *Proceedings of ICASSP*’96, page 1053–1057, 1996.

4. Reichl, W. and Chou, W. “A Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling”, *Proceedings of ICASSP’98*, page 801 – 804, *1998.*

5. Chou, W and Zhou, Q. “An Approach of Continuous Speech Recognition based on Self-Adjusting Decoding Graph”, *Proceedings of ICASSP’97*, page 1797 – 1782, 1997.

6. Young, S.J., Odell, J. and Woodland, P.C., “Tree-based State Tying for High Accuracy Acoustic Modeling”, In *ARPA Human Language Technology Workshop*, page 286 – 291, 1994.