Bayesian Decision Tree State Tying for Conversational Speech Recognition

Rusheng Hu and Yunxin Zhao

Department of Computer Science
University of Missouri, Columbia, MO 65211, USA
rhe02@mizzou.edu, zhaoy@missouri.edu

Abstract

This paper presents a new method of constructing phonetic decision trees (PDTs) for acoustic model state tying based on implicitly induced prior knowledge. Our hypothesis is that knowledge on pronunciation variation in spontaneous, conversational speech contained in a relatively large corpus can be used for building domain-specific or speaker-dependent PDTs. In the view of tree structure adaptation, this method leads to transformation of tree topology in contrast to keeping fixed tree structure as in traditional methods of speaker adaptation. A Bayesian learning framework is proposed to incorporate prior knowledge on decision rules in a greedy search of new decision trees, where the prior is generated by a decision tree growing process on a large data set. Experimental results on the Telemedicine automatic captioning task demonstrate that the proposed approach results in consistent improvement in model quality and recognition accuracy.

Index Terms: decision tree state tying, approximate Bayesian

1. Introduction

Recently, many efforts have been made to improve PDT state tying based acoustic modeling for continuous speech recognition [1, 2, 3]. Tree-structured adaptation methods were also reported, which attempted to apply hierarchically organized priors in building more accurate acoustic models by speaker adaptation [4, 5]. Researchers tackled the tree construction problem from different perspectives, which can be roughly grouped into two categories, namely the knowledge-based and the data-driven approaches. The knowledge-based approach refers to phonetic decision tree state tying which uses phonetic decision rules for clustering of HMMs. The data-driven approach refers to agglomerative clustering based on a distance measure between Gaussian densities. An earlier work in [6] has shown that the two approaches have similar performances while the knowledge-based method has the advantage of allowing model construction for unseen triphones. Another limitation of the data-driven method is its lack of robustness in dealing with mismatches between acoustic feature spaces caused by pronunciation variations when applied to speaker adaptation.

It is our belief that knowledge-based modeling can generalize better in large pronunciation variation situations. However, without adaptive learning, knowledge-based approach could possibly suffer from mismatches between the knowledge source and the specific task domain for which it needs to be applied. Our hypothesis is that systematic relationships between phonological variations and acoustic realizations can be extracted by a dynamic PDT process growing on a relatively large data source. Such information can in turn be used adaptively for generating domain-specific or speaker-dependent acoustic models.

The common framework of tree growing methods is recursive partitioning of input space by using a one-step lookahead strategy. Research efforts on improving phonetic decision tree modeling have been focused on tree growing strategy [1], model structure selection with information criterion [2], and enrichment of splitting questions [1][2]. However, without using appropriate prior knowledge on favored decision tree structures, uncertainty remains in the resulting phonetic decision trees. For instance, once a wrong decision is made, the split is irreversible and there is no provision for backtracking and choosing an alternative split. This problem is acute when speaker adaptation is carried out based on a mismatched tree structure. To the best knowledge of the authors, adaptive learning of phonetic decision tree structures has not yet been shown in previous literatures.

In this paper, we present a novel acoustic modeling approach using knowledge-based adaptive decision tree state clustering. By adaptive, we mean that the prior knowledge on phonological rules is implicitly represented by a tree-generating process on a large corpus, which is used to select good candidate splitting variables for constructing target PDTs in a specific domain that has limited amount of training data. In contrast to traditional methods which find an optimal tree cut in a single large tree (often a speaker independent tree), the proposed method employs prior knowledge of decision rules in a greedy search for domain-specific PDTs, and thus the resulting tree is not necessarily restricted to be a tree cut of an existing tree. The contributions of this paper are the following three aspects.

A general Bayesian learning framework for PDTs is developed to incorporate prior knowledge of favored tree structures. The probability distribution of a decision tree is decomposed into probabilities on tree structure, which contains the tree topology and the tests carried out at internal nodes, and the observation distributions at leaf nodes. By making appropriate simplifications, our tree priors are mainly composed of prior probabilities of splitting variables at internal nodes.

A Bayesian tree information criterion (BTIC) is defined and used as decision tree model selection criterion. Assuming informative priors on tree structure, BTIC is derived as an extension to the well-known Bayesian information criterion (BIC).

A computationally feasible algorithm for prior probability induction is developed. The priors of splitting questions are implicitly represented by a decision tree growing process on a large corpus. In general, considering the number of possible realizations of a decision tree, a direct computation of priors on tree structures is intractable. We propose a novel solution to this problem by introducing an oracle tree generation process which

provides estimates of prior probabilities of splitting variables recursively in a top-down manner.

The rest of the paper is organized as follows. In section 2, a theoretical background on Bayesian trees is introduced. Formulation for Bayesian learning of phonetic decision trees and the proposed BTIC are presented in section 3. Section 4 describes the knowledge-based adaptive PDT algorithm with the use of a dynamic decision tree process for obtaining priors on the splitting questions. Experimental results are presented in Section 5. Finally, findings and future research questions are summarized in Section 6.

2. Background on Bayesian Decision Tree

The theory and algorithms on Bayesian learning of decision trees were first studied in [7], where probability distribution of a decision tree was decomposed into probabilities of a tree structure, which contains the tree topology and the tests at each splitting node, and the observation distribution densities at each leaf node. Subsequently, effective Bayesian stochastic search algorithms using Markov Chain Monte Carlo (MCMC) simulation were developed for Bayesian inference of trees [8]. In introducing the framework of Bayesian decision tree, we will follow the notations as used in [8].

2.1. Bayesian Decision Tree

A binary decision tree with *k* terminal nodes is uniquely identified by a set of variables $T=\left(s_i^{pos},s_i^{\text{var}},s_i^{rule}\right)$, $i=1,...,k-1$, where s_i^{pos}, s_i^{var} and s_i^{rule} denote the position, variable and the point where the variable is split for each splitting node *i*. Let $C=\{c_1,...,c_k\}$ be the set of *k* terminal nodes, and define an associated parameter set as $\Theta=\left(\theta_1,...,\theta_k\right)$, where θ_j is the parameter of the observation distribution density at the j^{th} terminal node. A training data set is defined as $(Y,X)=\{y_t,x_t\}$, $t=1,...,n$, where $y=\left(y_1,...,y_d\right)^T$ is the *d-dimensional* observation variable and $x=\left(x_1,...,x_p\right)^T$ is the *p-dimensional* splitting variable. Assume that conditioned on (Θ,T), the observations are independent across terminal nodes, and are i.i.d. within terminal nodes. The joint distribution of observations is of the form

$$p\left(Y\mid X,\Theta,T\right)=\prod_{i=1}^{k}\prod_{j=1}^{n_i}p\left(y_{ij}\mid\theta_i\right) \tag{1}$$

where $Y_i=\left\{y_{ij},\ j=1,...,n_i\right\}$ denote data points in the terminal node c_i. The posterior distribution of *T* is given by

$$\begin{aligned}p(T\mid X,Y)&\propto p(Y\mid X,T)p(T)\\&=p(T)\int p(Y\mid X,\Theta,T)p(\Theta\mid T)d\Theta\end{aligned} \tag{2}$$

up to a normalizing constant. Analytical forms of the integral $p\left(Y\mid X,T\right)=\int p\left(Y\mid X,\Theta,T\right)p\left(\Theta\mid T\right)d\Theta$ can be obtained by using conjugate priors or *Laplace* approximation [8][9].

2.2. Non-Informative Prior

The prior on tree $T=\left(s_i^{pos},s_i^{\text{var}},s_i^{rule}\right),i=1,...,k-1$ can be specified as follows. First, a discrete distribution $p\left(s_i^{\text{var}}\right)$ is defined over the domain $s_i^{\text{var}}\in\{1,...,p\}$ that corresponds to indices of the *p* splitting variables in $x=\left(x_1,...,x_p\right)^T$. Second, a conditional distribution $p\left(s_i^{rule}\mid s_i^{\text{var}}\right)$ is specified with s_i^{rule} taking a total of $n\left(s_i^{\text{var}}\right)$ possible values for the splitting variable s_i^{var}. Finally, an upper bound of splits allowed in one path down the tree, $S_{\max}$, is set to ensure a finite number of possible trees, i.e., $s_i^{pos}\in\left\{1,...,2^{S_{\max}+1}-1\right\}$.

Usually the distributions $p\left(s_i^{\text{var}}\right)$ and $p\left(s_i^{rule}\mid s_i^{\text{var}}\right)$ are chosen as uniform distributions. In such a case, the prior distribution for a complete tree structure becomes

$$\begin{aligned}p(T)&=\left\{\prod_{i=1}^{k-1}p\left(s_i^{rule}\mid s_i^{\text{var}}\right)p\left(s_i^{\text{var}}\right)\right\}p\left(\left\{s_i^{pos}\right\}_1^{k-1}\right)\\&=\left\{\prod_{i=1}^{k-1}\frac{1}{n\left(s_i^{\text{var}}\right)}\frac{1}{p}\right\}\frac{k!}{S_k}\frac{1}{K}\end{aligned} \tag{3}$$

where S_k is the total number of possible ways of choosing $\left\{s_i^{pos}\right\}_1^{k-1}$ to produce a *k*-terminal node tree, and *K* is the maximum number of terminal nodes. For binary decision trees, S_k is given in *graph* theory as the *Catalan number*

$$S_k=\frac{1}{k+1}\binom{2k}{k} \tag{4}$$

The prior on tree topology $p\left(\left\{s_i^{pos}\right\}_1^{k-1}\right)=\frac{k!}{S_k}\frac{1}{K}$ is a function of the number of terminal nodes *k* and is independent of rule assignments in splitting nodes.

3. Bayesian PDT Learning

3.1. Informative Prior on Tree Structure

Note that the prior $p(T)$ defined in (3) is non-informative. When prior knowledge of favored tree structures is available, it is beneficial to consider informative priors on tree structures. In phonetic decision tree based state tying, this knowledge is carried by the splitting variables, i.e., phonetic questions being asked at each splitting node. Since the answers to the phonetic questions only take Boolean values (true/false), we have $p\left(s_i^{rule}\mid s_i^{\text{var}}\right)=1$ conditioned on a given splitting variable. Furthermore, $p\left(\left\{s_i^{pos}\right\}_1^{k-1}\right)$ only depends on tree topology and is assumed uniformly distributed, therefore it is treated as a nuisance factor. By focusing on splitting variables, we use the following form of prior in PDT modeling

$$p(T)\propto\prod_{i=1}^{k-1}p\left(s_i^{\text{var}}\right) \tag{5}$$

The strategy of implicit modeling for $p(s_i^{\text{var}})$ will be given in Section 4.

3.2. Bayesian Tree Information Criterion

The Bayesian model selection criterion chooses the tree structure which has the highest posterior probability. Substituting (1) and (5) into (2) yields

$$\begin{aligned} p(T \mid X,Y) &\propto p(T) \int p(Y \mid X,\Theta,T)\, p(\Theta \mid T)\, d\Theta \\ &\propto \left\{ \prod_{i=1}^{k-1} p(s_i^{\text{var}}) \right\} \times \int \prod_{i=1}^{k} \left\{ \prod_{j=1}^{n_i} p(y_{ij} \mid \theta_i)\, p(\theta_i \mid T) \right\} d\Theta \end{aligned} \quad (6)$$

The Bayesian tree information criterion (BTIC) is defined to be the logarithm of the tree posterior probability

$$BTIC(T) = \log p(T \mid X,Y) \quad (7)$$

A key problem in evaluating BTIC is the computation of the evidence of observations, $p(Y \mid X,T)$, given as,

$$p(Y \mid X,T) = \int \prod_{i=1}^{k} \left\{ \prod_{j=1}^{n_i} p(y_{ij} \mid \theta_i)\, p(\theta_i \mid T) \right\} d\Theta \quad (8)$$

The integral over parameter space Θ is often intractable when considering complex models. The *Laplace* approximation method for exponential family as described in [9] has been extensively used in the literature to evaluate the integral in (8). Assuming that the function $p(Y_i \mid \theta_i)\, p(\theta_i \mid T)$ is strongly peaked at the ML estimate $\hat{\theta}_i$, i.e., $p(Y_i \mid \theta_i)\, p(\theta_i \mid T)$ is dominated by the term $p(Y_i \mid \theta_i)$, a second-order Taylor series expansion of the logarithm of this function around $\hat{\theta}_i$ leads to a tractable form

$$\begin{aligned} &\log \int p(Y_i \mid \theta_i)\, p(\theta_i \mid T)\, d\theta_i \approx \log p(Y_i \mid \hat{\theta}_i) \\ &+ \log p(\hat{\theta}_i \mid T) + \frac{D}{2}\log(2\pi) - \frac{D}{2}\log n_i - \frac{1}{2}\log \left| I_y(\theta_i) \right| \\ &\overset{n_i >> 0}{\approx} \log p(Y_i \mid \hat{\theta}_i) - \frac{D}{2}\log n_i = BIC \end{aligned} \quad (9)$$

where D is the number of free parameters in the model and $I_y(\theta_i)$ is the Fisher information matrix. The resulting value is equivalent to the well known Bayesian information criterion (BIC), also known as Schwarz information criterion (SIC) [9]. After standard analytical simplification, the Bayesian tree information criterion as defined in (8) is derived to be

$$BTIC(T) = BIC(C) + \gamma \sum_{i=1}^{k-1} \log p(s_i^{\text{var}}) \quad (10)$$

where γ is a regularizing parameter, $BIC(C)$ is the Bayesian information criterion for the terminal nodes, given as follows,

$$BIC(C) = \sum_{c_i \in C} BIC(Y_i, c_i) = \sum_{i=1}^{k} \left(\log p(Y_i \mid \hat{\theta}_i) - \frac{D}{2}\log n_i \right) \quad (11)$$

4. Knowledge-Based Adaptive Decision Tree

Our knowledge-based adaptive decision tree (KBA-PDT) is a top-down Bayesian PDT learning approach which utilizes BTIC as model selection criterion. The key in computing BTIC is getting appropriate estimates of the prior probabilities of splitting variables from a large corpus. Considering the huge number of possible realizations of a decision tree, a direct estimation for $p(s_i^{\text{var}})$ would be intractable [8]. In an adaptive learning setting, we propose a novel solution to this problem by recursively defining $p(s_i^{\text{var}})$ based on the beliefs generated by a dynamic decision tree growing process on a large data set, as follows

$$p(s_i^{\text{var}}) \propto \begin{cases} \Delta BTIC, & \text{if } s_i^{\text{var}} \in \text{ top } h \text{ variables} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where

$$\Delta BTIC = \left(BIC(s_{i\_L}) + BIC(s_{i\_R}) \right) - BIC(s_i)$$

is the information gain due to splitting the node s_i to its left and right children nodes s_{i_L} and s_{i_R} according to the splitting variable s_i^{var}, h is the number of splitting variables which give the h-best improvement in BTIC. Note that in splitting the large data set, the prior on splitting variables is assumed uniform and the information gain is equivalent to improvement in BIC. The probability $p(s_i^{\text{var}})$ is defined positive only for the h-best splitting variables, and its value is proportional to the corresponding information gain with the stochastic constraint that the sum of the probabilities equal to one. Forcing the probabilities of ineffective splitting variables to zero is for reducing noise and uncertainty in the tree learning process.

As discussed above, BTIC model selection is performed by two interleaved tree growing processes. The primary tree process is the domain-specific PDT which we are searching for, and hence is called a target tree. The secondary tree process provides beliefs on splitting variables to the primary tree, and is therefore called an oracle tree. The split of oracle tree is governed by the target tree and is in fact an identical copy of the target tree but growing in a different observation space. Each tree is built top-down in a recursive fashion. Initially, all the states to be clustered are pooled at the roots of the oracle tree and target tree, respectively, and the BTICs of the trivial trees (contain only one node) are computed. Next, the oracle tree tries all the splits and get the estimates of $p(s_i^{\text{var}})$, and forward these probability estimates to the target tree. Having received $p(s_i^{\text{var}})$, the target tree node is split into two by finding the question which gives the maximum increase in BTIC. The target tree then sends its splitting information (node split and question used) back to oracle tree. At last, oracle tree follows the same split as that of target tree. This process is repeated until some stopping criterion is met. These stopping criteria include thresholds on occupancy count at leaf nodes, and information gain obtained from a split. To evaluate BTIC, recall that we use the approximated BTIC given by

$$BTIC(T) \approx \log L(T) - \frac{D}{2}\sum_{i=1}^{k} \log n_i + \gamma \sum_{i=1}^{k-1} \log p(s_i^{\text{var}}) \quad (13)$$

where $L(T)$ is the likelihood of the observations on the leaf nodes, γ is an adjustable regularizing factor, and the sample count at the leaf node c_i, n_i, is approximated by accumulated

state occupancies which are estimated from the Baum-Welch algorithm.

5. Experiments

5.1. Experimental Setup

The proposed knowledge-based adaptive decision tree algorithm was evaluated on the Telemedicine automatic captioning task developed at the University of Missouri-Columbia. For a detailed description of this project, please refer to [10]. Speaker dependent acoustic models were trained for 5 speakers, including two females (D1 and D5) and three males (D2, D3, D4). A summary of the data sets is provided in Table 1. The training and test datasets were extracted speech data from the speakers' conversations with clients in mock Telemedicine interviews. Along with speech durations, word counts from transcription texts are also given in Table 1. Speech features consisted of 39 components including 13 MFCCs and their first and second order time derivatives. Feature analysis was made at a 10 ms frame rate with 20 ms window size. Gaussian mixture density based hidden Markov models (GMM-HMM) were used for within-word triphone modeling, where each GMM contained 16 Gaussian components. The task vocabulary is of the size 46,489, with 3.07% of vocabulary words being medical terms.

Table 1. *Datasets of 5 Speakers: speech(min.)/text(no. of words)*

	Training set	Test set
D1	210/35,348	29.8/5,105
D2	200/39,398	14.3/2,760
D3	145/28,700	19.3/3,238
D4	180/39,148	27.8/6,492
D5	250/44,967	12.1/3,998
Total	985/187561	103.3/21593

5.2. Experimental Results

HMM states were tied using the proposed BTIC based decision tree procedure (KBA-PDT), where the large corpus for oracle tree construction contained pooled speech from all the speakers, and the small corpus for a target tree contained speech of a single speaker. PDT question set used was the same as the HTK question set [6]. Prior to building the trees, single Gaussian acoustic models were first estimated for untied triphone states and sufficient statistics were accumulated for the oracle and target trees. The resulting speaker dependent PDTs were then used to cluster HMM states and construct unseen triphones. At last, tied single Gaussian models were augmented to 16 components by the HTK splitting procedure. Baseline models were also trained by using the conventional maximum likelihood criterion (ML-PDT). The model complexity and word accuracy results are summarized in Table 2, where the tuning factors of h and γ were optimized for each speaker. The average results were weighted by the relative word counts of the five test datasets. It is shown that KBA-PDT consistently outperformed ML-PDT in increased accuracy (by 0.7% absolute) and reduced model complexity (by 27% relative).

6. Discussion and Conclusions

In this paper, we presented a novel acoustic modeling approach using knowledge-based adaptive decision tree state tying. A Bayesian learning framework for PDT was developed to incorporate prior knowledge on tree structures, and an oracle-tree/target-tree process was devised to efficiently search for optimal splits based on a Bayesian tree information criterion newly proposed in this work.

Table 2. Effectiveness of knowledge-based adaptive PDT

		KBA-PDT	ML-PDT
D1	Number of states	1611	2238
	Word accuracy	***81.75***	*81.17*
D2	Number of states	1119	1569
	Word accuracy	***74.30***	*73.15*
D3	Number of states	799	1156
	Word accuracy	***74.98***	*73.95*
D4	Number of states	1027	1521
	Word accuracy	***78.35***	*77.96*
D5	Number of states	1552	1838
	Word accuracy	***83.55***	*82.80*
W. Avg.	Number of states	1240	1700
	Word accuracy	***79.09***	*78.39*

It has been shown that the proposed method gives consistent improvement over conventional methods in model quality and recognition performance. When tested on the Telemedicine automatic captioning task, it improved the word error rate by 0.7% (absolute) on average with 27% reduced model complexity, using optimal settings of training factors for each speaker.

7. Acknowledgements

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC04340-01A2.

8. References

[1] W. Reichl and W. Chou, "Robust decision tree state tying for continuous speech recognition," IEEE Trans. Speech Audio Proc., vol. 8, no. 5, pp. 555–566, 2000.

[2] J-T. Chien and S. Furui, "Predictive hidden Markov model selection for speech recognition," IEEE Trans. Speech Audio Proc., vol. 13, no. 3, pp. 377–387, 2005.

[3] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," Comput. Speech Lang., vol. 17, no. 4, pp. 311–328, 2003.

[4] S. Wang and Y. Zhao, "Online Bayesian tree-structured transformation of HMMs with optimal model selection for speaker adaptation," IEEE Trans. Speech Audio Proc., vol. 9, no. 6, pp. 663–677, 2001.

[5] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech Audio Proc., vol. 9, no. 3, pp. 276–287, 2001.

[6] S. J. Young, J. J. Odell and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in Proc. ARPA Human Lang. Tech. Workshop, pp. 307–312, 1994.

[7] W. L. Buntine, A Theory of Learning Classification Rules, PhD thesis, School of Comput. Sci., Univ. Tech., Sydney, 1992.

[8] D. Denison, C. Holmes, B. Malick and A. Smith, Bayes. Methods for Nonlinear Classification and Regression, Willey, 2002.

[9] G. Schwarz, "Estimating the dimension of a model," Ann. Statist., vol. 6, no. 2, pp. 465-471, 1978.

[10] Y. Zhao, X. Zhang, R-S. Hu, J. Xue, X. Li, L. Che, R. Hu and L. Schopp, "An automatic captioning system for telemedicine," in Proc. ICASSP06, to appear.