A BAYESIAN APPROACH FOR PHONETIC DECISION TREE STATE TYING IN 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 of 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 of 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—Acoustic modeling, decision tree state tying, approximate Bayesian, implicit prior

1. INTRODUCTION

In speech recognition literature, the common framework of growing phonetic decision trees is recursive partitioning of input space by using a greedy search 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 our original work [3], we proposed a novel acoustic modeling approach using knowledge-based adaptive decision tree state clustering. By adaptive, we mean that the prior knowledge of linguistic 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 new method can be summarized in 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.

A Bayesian tree information criterion (BTIC) is defined and used as a 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 this paper, we interpret our knowledge-based adaptive phonetic decision tree construction algorithm as a knowledge variation modeling approach, which is part of pronunciation variation modeling in conversational speech recognition. In deriving BTIC, we give an exact solution based on Normal-Wishart prior distributions in addition to the solution obtained from Laplace approximation.

The rest of the paper is organized as follows. In section 2, the theoretical background and formulation for Bayesian learning of phonetic decision trees are introduced. The BTIC-based knowledge variation modeling is presented in section 3. Experimental results are given in Section 4. Finally, findings and future research questions are summarized in Section 5.

2. BAYESIAN PHONETIC DECISION TREE

2.1. Bayesian Decision Tree

The theory and algorithms on Bayesian learning of decision trees were first studied in [4]. Subsequently, effective Bayesian stochastic search algorithms using Markov Chain Monte Carlo (MCMC) simulation were developed for Bayesian inference of trees [5]. In introducing the framework of Bayesian decision tree, we will follow the notations as used in [5].

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC04340-05.

Given a set of splitting variables $x=\left(x_1,...,x_p\right)^T$, 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*. The unique positions s_i^{pos} can be defined by a simple labeling scheme. The root node, which is always in a binary tree, is the first split node and its position is labeled $s_1^{pos}=1$. Any descendant splitting node's position is then uniquely defined by its parent's position, i.e., letting $s_i^{pos}=2parent(s_i^{pos})$ if it is the left child and $s_i^{pos}=2parent(s_i^{pos})+1$ otherwise. The positions of leaf nodes are similarly defined but are not included in the model because they are completely determined by the tree structure given the internal nodes. To illustrate the use of s_i^{var} and s_i^{rule}, suppose the question at split node s_i^{pos} is "$x_1<2?$", then the split variable $s_i^{\text{var}}=1$, the split point is 2 and thus the split rule variable $s_i^{rule}=2$. When the split variable takes binary values (*0/1*), then $s_i^{rule}\equiv 1$, hence the split rule variable can be ignored in the model. 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 $Y_i=\left\{y_{ij},\ j=1,...,n_i\right\}$ denotes 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 [5][6].

2.2. Informative Prior on Tree Structure

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

$$p\left(T\right)\propto\prod_{i=1}^{k-1}p\left(s_i^{\text{var}}\right) \tag{3}$$

The strategy of implicit modeling for $p\left(s_i^{\text{var}}\right)$ will be given in Section 3.

2.3. Bayesian Tree Information Criterion

The Bayesian model selection criterion chooses the tree structure which has the highest posterior probability. Substituting (1) and (3) into (2) yields

$$\begin{aligned}p\left(T\mid X,Y\right)&\propto p\left(T\right)\int p\left(Y\mid X,\Theta,T\right)p\left(\Theta\mid T\right)d\Theta\\&\propto\left\{\prod_{i=1}^{k-1}p\left(s_i^{\text{var}}\right)\right\}\times\int\prod_{i=1}^{k}\left\{\prod_{j=1}^{n_i}p\left(y_{ij}\mid\theta_i\right)p\left(\theta_i\mid T\right)\right\}d\Theta\end{aligned} \tag{4}$$

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

$$BTIC\left(T\right)=\log p\left(T\mid X,Y\right) \tag{5}$$

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

$$p\left(Y\mid X,T\right)=\int\prod_{i=1}^{k}\left\{\prod_{j=1}^{n_i}p\left(y_{ij}\mid\theta_i\right)p\left(\theta_i\mid T\right)\right\}d\Theta \tag{6}$$

The integral over parameter space Θ is often intractable when considering complex models. In PDT literature, two kinds of approaches are commonly employed to tackle this problem, referred to as the exact method and the approximate method, respectively. The exact method makes assumption on the parametric forms of observation distributions and the prior of the distribution parameters at leaf nodes. For multivariate normal observation distributions at leaf nodes, i.e.

$$p\left(y_{ij}\mid\theta_i\right)=N_d\left(y_{ij}\mid m_i,R_i\right) \tag{7}$$

where $N_d\left(y_{ij}\mid m_i,R_i\right)$ is a *d-dimensional multivariate normal distribution* with mean m_i and precision matrix R_i, the exact method uses the *normal-Wishart* conjugate prior as follows [2],

$$\begin{aligned}&p\left(m_i,R_i\mid\tau_i,\mu_i,\alpha_i,\Psi_i\right)\propto\left|R_i\right|^{(\alpha_i-p)/2}\\&\times\exp\left\{-\frac{\tau_i}{2}\left(m_i-\mu_i\right)^T R_i\left(m_i-\mu_i\right)\right\}\exp\left\{-\frac{1}{2}tr\left(\Psi_i R_i\right)\right\}\end{aligned} \tag{8}$$

where $\tau_i,\alpha_i,\mu_i,\Psi_i$ are hyper-parameters. Analytical results show that the evidence $p\left(Y\mid X,T\right)$ is in the form of *d-dimensional multivariate student t distribution*

$$\begin{aligned}p\left(Y_i\mid X,T\right)\propto&\left(\frac{\tau_i}{\tau_i+n_i}\right)^{1/2}\left|\Psi_i\right|^{-(\alpha_i+n_i)/2}\\&\times\left(1+s_i+t_i\right)^{-(\alpha_i+n_i)/2}\end{aligned} \tag{9}$$

where $s_i = \sum_{t=1}^{n_i} (y_{it} - \bar{y}_i)^T \Psi_i^{-1} (y_{it} - \bar{y}_i)$, and

$$t_i = \frac{\tau_i n_i}{\tau_i + n_i} (\bar{y}_i - \mu_i)^T \Psi_i^{-1} (\bar{y}_i - \mu_i) \cdot$$

The *Laplace* approximation method for exponential family as described in [6] has been extensively used in the literature to evaluate the integral in (6). 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

$$\begin{aligned} &\log \int p(Y_i \mid \theta_i) p(\theta_i \mid T) d\theta_i \approx \log p(Y_i \mid \hat{\theta}_i) \\ &+ \log p(\hat{\theta}_i \mid T) + \frac{D}{2} \log(2\pi) - \frac{D}{2} \log n_i - \frac{1}{2} \log \left| I_y(\theta_i) \right| \\ &\overset{n_i >> 0}{\approx} \log p(Y_i \mid \hat{\theta}_i) - \frac{D}{2} \log n_i = BIC \end{aligned} \tag{10}$$

where D is the number of free parameters in the model and $I_y(\theta_i)$ is the Fisher information matrix for a single observation, which is bounded and hence becomes insignificant when sample size grows to infinity. The resulting value is equivalent to the well known Bayesian information criterion (BIC), also known as Schwarz information criterion (SIC) [6].

In experiments presented in this paper, we adopt the approximate method to compute BTIC because of its computation convenience. The exact form will be investigated in another work. After standard analytical simplification, the Bayesian tree information criterion as defined in (5) is derived to be

$$BTIC(T) = BIC(C) + \gamma \sum_{i=1}^{k-1} \log p(s_i^{\text{var}}) \tag{11}$$

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

$$BIC(C) = \sum_{c_i \in C} BIC(Y_i, c_i) = \sum_{i=1}^{k} \left(\log p(Y_i \mid \hat{\theta}_i) - \frac{D}{2} \log n_i \right) \tag{12}$$

3. KNOWLEDGE-BASED ADAPTIVE PHONETIC DECISION TREE

Recently, much attention has been drawn to employ knowledge-based features for speech recognition [7, 8]. The rational behind these methods is that incorporating more knowledge of acoustic-phonetics into acoustic modeling will provide more accurate and robust models of conversational speech. Our knowledge-based adaptive decision tree (KBA-PDT) approach fits in this scenario in that it extracts the knowledge from a relatively large corpus and provides beliefs of phonetic questions for construction of target PDTs. The key idea is to provide a reasonable way to model knowledge variation, which represents the intra-speaker/inter-speaker variations in their understandings on how to make a correct pronunciation for a given word or subword unit in a certain context, and is achieved by appropriate estimates of the prior probabilities of phonetic questions 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 [5]. 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

$$p(s_i^{\text{var}}) \propto \begin{cases} \Delta BTIC, & \text{if } s_i^{\text{var}} \in \text{ top } h \text{ variables} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where

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

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 target tree is built top-down in a recursive fashion. A node split is made by sequentially evaluating each splitting question that has a nonzero probability at this node, and by taking the split that results in the largest increase in BTIC. The priors $p(s_i^{\text{var}})$ that are used by the target tree to evaluate BTIC are estimated by the oracle tree, which copies the current structure of the target tree but keeps its own observation data. Starting from the root node, the oracle tree tries each splitting variable on the current node and gets the estimates of $p(s_i^{\text{var}})$ based on equation (13); the oracle tree forwards these probability estimates to the target tree to assist the split of its current node in the way described above; the best splitting variable found by the target tree is then used to split the current node of the oracle tree. This interleaved tree growing process is repeated for every node in the two trees until some stopping criterion is met. These stopping criteria include thresholds on occupancy count at leaf nodes, and on information gain obtained from a split. To evaluate BTIC, recall that we use the approximated BTIC given by

$$BTIC(T) \approx \log L(T) - \frac{D}{2} \sum_{i=1}^{k} \log n_i + \gamma \sum_{i=1}^{k-1} \log p(s_i^{\text{var}}) \tag{14}$$

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 of state i which are estimated from the Baum-Welch algorithm.

4. EXPERIMENTS

4.1. 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 [9]. Speaker dependent acoustic models were trained for 5 speakers, including two females (D1 and D5) and three males (D2, D3, D4). A description of the data sets is given in [9]. The training and test datasets were extracted speech data from the health care provider speakers' conversations with clients in mock Telemedicine interviews. 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.

4.2. 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 [10]. 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 1, where the tuning factors are fixed with $h=10$ and $\gamma=10$. 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 0.5% absolute) and reduced model complexity (by 27% relative).

Table 1. 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	***73.73***	*73.15*
D3	Number of states	799	1156
	Word accuracy	***74.98***	*73.95*
D4	Number of states	1098	1521
	Word accuracy	***78.29***	*77.96*
D5	Number of states	1397	1838
	Word accuracy	***83.02***	*82.80*
W. Avg.	Number of states	1232	1700
	Word accuracy	***78.90***	*78.39*

The value h for the prior probability of splitting variables as defined in (13) is a tuning constant; a small value of h implies a strong belief on the knowledge extracted from the large corpus, which leads to resistance to noise and uncertainty in the domain-specific training data, but at the risk of low robustness when the large data set is not representative of the target task. The performance of KBA-PDT versus different values of h is given in Table 2.

Table 2. Effects of Number of Active Questions h ($\gamma = 10$)

	h	1	5	10	200
D1	Number of states	1250	1463	1611	1804
	Word accuracy	*80.83*	*80.83*	***81.75***	*81.28*
D2	Number of states	871	1021	1119	1278
	Word accuracy	*73.13*	*73.01*	***73.73***	*73.20*
D3	Number of states	581	727	799	944
	Word accuracy	*74.67*	*74.67*	***74.98***	*74.73*
D4	Number of states	852	1027	1098	1204
	Word accuracy	*77.57*	***78.35***	*78.29*	*78.35*
D5	Number of states	1002	1216	1397	1552
	Word accuracy	*82.95*	*83.40*	*83.02*	***83.55***
W. Avg	Number of states	935	1119	1232	1380
	Word accuracy	*78.33*	*78.63*	***78.90***	*78.80*

5. 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.

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 0.5% (absolute) on average with 27% reduced model complexity.

6. 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] R-S. Hu and Y. Zhao, "Bayesian decision tree state tying for conversational speech recognition," Proc. INTERSPEECH06, pp. 1738-1741, Pittsburgh, PA, 2006.

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

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

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

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

[8] B. Launay, O. Siohan, A. Surendran and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," Proc. ICASSP02, vol. 1, pp. I-817-I820, 2002.

[9] Y. Zhao, X. Zhang, R-S. Hu, J. Xue, X. Li, L. Che, R. Hu and L. Schopp, "An automatic captioning system for telemedicine," Proc. ICASSP06, pp. I-957-960, Toulouse, France, 2006.

[10] 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.