RANDOM-FORESTS-BASED PHONETIC DECISION TREES FOR CONVERSATIONAL SPEECH RECOGNITION

Jian Xue and Yunxin Zhao

Department of Computer Science
University of Missouri, Columbia, MO 65211 USA
jxwr7@mizzou.edu, zhaoy@missouri.edu

ABSTRACT

In this paper we present a novel technique of constructing phonetic decision trees (PDTs) for acoustic modeling in conversational speech recognition. We use Random Forests (RF) to train a set of PDTs for each phone-state unit and obtain multiple acoustic models accordingly, and we extend the PDT-based state tying to RF-based state-tying. We combine acoustic scores at the model level in decoding search. Several methods are investigated to estimate the weight parameters for model combination, including maximum likelihood estimation of the weights from training data, as well as using confidence scores of *P*-value or relative entropy to obtain the weights dynamically from online data. Experimental results on a telemedicine automatic captioning task demonstrate that the proposed RF-PDT technique leads to significant improvements in word recognition accuracy.

Index Terms—Random Forests, phonetic decision trees, acoustic modeling, score combination

1. INTRODUCTION

Phonetic decision tree (PDT) based state tying is commonly used in acoustic modeling for large vocabulary continuous speech recognition. PDT can incorporate phonetic knowledge into triphone state clustering and model triphone units or contexts that do not occur in training data [1]. Several efforts have been reported to improve PDT state tying in acoustic modeling [2]-[4]. These methods aim to find one optimal PDT for each phone or phone state. On the other hand, in machine learning, using ensemble methods for classifier design has been studied and advocated [7]. Random Forests (RF) as proposed in [8] generates an ensemble of decision trees by stochastically sampling training data and variables, and it is considered unexcelled in accuracy among current classification techniques [8] [9]. A variant of RF was previously used in [5] and [6] for constructing PDTs to generate multiple 1-best decoding output hypotheses.

In this paper, we present a technique of training multiple PDTs for each phone-state unit following the RF algorithm of [8] [9]. Instead of combining output hypotheses at the word level after decoding search as in [5] and [6], we combine acoustic scores at the model level during decoding search to avoid multiple runs of decoding search and achieve better accuracy performance. We generate RF tied states by tying triphone states that belong to identical tied states across all PDTs. Several methods are investigated to estimate the model combination weights, which may be specific to RF tied states, specific to speech frame inputs, or both.

The rest of the paper is organized as follows. We introduce the background of RF in section 2 and propose the RF-based PDTs in section 3. In section 4 we discuss methods of combining acoustic scores from multiple acoustic models in decoding search. In section 5 we provide a detailed account of experimental results. We conclude our work in section 6.

2. BACKGROUND OF RANDOM FORESTS

An early example of RF is bagging [10], where to grow each tree a random sampling with replacement is made from a training data set to generate bootstrap replicated training sets. Another example is random split selection [11], where at each tree internal node a split is selected at random from among the *n*-best splits for the node. To classify an object, each tree gives a vote for the object, and the classification result is the one receiving the most votes from the collection of trees.

The common element in these procedures is that, for the *k*th tree, a random vector θ_k is generated that is independent and identically distributed with the random vectors of other trees. A tree is grown by using the training data set and its random vector. Consider a training set of *N* data samples. In bagging, the random vector θ consists of *N* random numbers generated as the counts in *N* boxes resulting from *N* darts thrown at the boxes, and a count thus corresponds to the number of times the associated data sample will be used in constructing the tree. Let *L* equal the number of non-leaf nodes in the tree. In *n*-best random split selection, θ consists of *L* independent random integers with values in the range of 1 through *n*, indicating the question in the *n*-best questions that will be used to split the node.

In the RF method of Breiman and Cutler [8], a tree of a forest is grown as follows. First, choose *N* samples randomly from the original training dataset of *N* samples as in bagging for growing the tree. Second, at each node, select *m* splitting variables randomly out of a total of *M* variables and choose the best split determined in these *m* variables at each node. Each tree is grown to the fullest extent based on some predefined thresholds without pruning.

3. RANDOM FORESTS-BASED PDTS

In the current work, we explore the potential power of RFs of [8] to generate multiple PDTs for each HMM state of a phone unit. We propose a modified question selection method to train PDTs in order to simplify the construction of random forests of acoustic models which are much more complex than the problems

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC04340-01A2-06.

considered in [8]. We randomly choose a subset of *m* phonetic questions out of a total of *M* questions to train one set of PDTs for all phone-state units, with one PDT for one phone-state unit, and refer to the acoustic models thus produced as one set of acoustic models. The procedure of constructing a PDT is the same as the commonly adopted deterministic greedy method. We then randomly choose another subset of *m* questions to generate another set of PDTs and so on. Different from the method of [8], we use the full set of training data of each phone state instead of randomly sampling the data to avoid completely dropping out some triphone state data in constructing a tree since our training data set is small.

Multiple sets of PDTs are thus generated for all speech units, and from which we obtain multiple sets of acoustic models $MS_1, \cdots, MS_K$. We define RF-tied states as the following. Each PDT set k has N_k models $M_{1k}, \cdots, M_{N_k k}$ corresponding to the N_k tied states. Each triphone state S_r therefore has K models $M_{r_1 1}, \cdots M_{r_K K}$, where r_k is a mapping from the triphone state S_r to a tied state in the model set k. If for two triphone states S_i and S_j, $i_k = j_k$ for $k = 1, \cdots K$, that is, S_i and S_j belong to the same tied state in every PDTs, then we say that S_i and S_j belong to the same RF tied state ES_l. Fig. 1 illustrates the construction of RF tied states from two PDTs. The triphone states of a RF tied state share the same set of models $M_{l_1 1}, \cdots, M_{l_K K}$. Since in the single PDT-based method data distribution within a tied state is usually modeled by a Gaussian mixture density (GMD), a RF tied state is modeled by multiple GMDs.

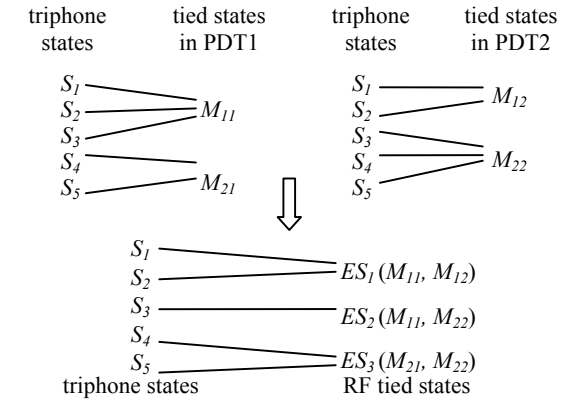

Fig. 1 An illustration of RF tied states.

4. SCORE COMBINATION

We investigate methods of combining the multiple acoustic models to improve accuracy of decoding search. For each speech feature vector $\underline{x}_t$, we need to combine the multiple GMDs in each RF tied state l to compute an acoustic score, that is

$$P(\underline{x}_t \mid ES_l) = F(\underline{x}_t \mid M_{l_1,}, \cdots, M_{l_K K}). \quad (1)$$

Considering a linear combination of acoustic scores gives

$$P(\underline{x}_t \mid ES_l) = \sum_{k=1}^{K} w_{tlk}\, p(\underline{x}_t \mid M_{l_k k}). \quad (2)$$

We need to estimate the weights w_{tlk}, which may vary with feature $\underline{x}_t$ or model $M_{l_k k}$, with $\sum_{k=1}^{K} w_{tlk} = 1$ and $w_{tlk} \geq 0$.

4.1 Maximum likelihood based weight estimation

First, we consider weights that are specific to each RF tied state, that is

$$P(\underline{x}_t \mid ES_l) = \sum_{k=1}^{K} w_{lk}\, p(\underline{x}_t \mid M_{l_k k}). \quad (4)$$

Let $X = (\underline{x}_1, \cdots, \underline{x}_T)$ be i.i.d observations drawn from a RF tied state ES_l in training data set, and denote $\underline{w}_l = (w_{l1}, \cdots w_{lK})$. The likelihood function is

$$L(X \mid \underline{w}_l) = \prod_{t=1}^{T} (\sum_{k=1}^{K} w_{lk}\, p(\underline{x}_t \mid M_{l_k k})). \quad (5)$$

The maximum likelihood estimation (MLE) of $\underline{w}_l$ gives the following:

$$\begin{aligned} \hat{\underline{w}}_l &= \arg\max_{\underline{w}_l} \{\prod_{t=1}^{T} (\sum_{k=1}^{K} w_{lk}\, p(\underline{x}_t \mid M_{l_k k}))\} \\ &= \arg\max_{\underline{w}_l} \{\sum_{t=1}^{T} \log(\sum_{k=1}^{K} w_{lk}\, p(\underline{x}_t \mid M_{l_k k}))\}, \end{aligned} \quad (6)$$

with $\sum_{k=1}^{K} w_{lk} = 1$ and $w_{lk} \geq 0$.

We use the EM algorithm [12] to iteratively compute $\underline{w}_l$. The reestimation formula is straightforwardly derived as

$$w_{lk}^{r+1} = \frac{1}{T} \sum_{t=1}^{T} \frac{w_{lk}^{r}\, p(\underline{x}_t \mid M_{l_k k})}{\sum_{j=1}^{K} w_{lj}^{r}\, p(\underline{x}_t \mid M_{l_j j})}, \quad (7)$$

where $w_{lk}^{0} = 1/K$ for $k = 1, \cdots, K$, and $r = 0, 1, \cdots$.

4.2 Confidence score based weight estimation

We consider using confidence scores to determine the time-dependent weights. The idea is that if for a RF tied state the confidence score of a certain model $M_{l_k k}$ on the feature $\underline{x}_t$ is larger than other models, i.e., we have higher confidence that $\underline{x}_t$ is generated by $M_{l_k k}$, then $M_{l_k k}$ should contribute more to $\underline{x}_t$'s acoustic score, and vice versa. Here we use the confidence score *P*-value. Unlike likelihood scores, *P*-value measures some area under a pdf curve with the area size depending on the distance between a data sample and the distribution mean as well as the distribution shape. *P*-value was first proposed in [13] as a confidence feature for speech recognition.

The *P*-value of the feature $\underline{x}_t$ evaluated on the Gaussian mixture density function or the tied state $M_{l_k k}$, $P_v(\underline{x}_t \mid M_{l_k k})$, can be used to measure the fit between $\underline{x}_t$ and $M_{l_k k}$. The larger the $P_v(\underline{x}_t \mid M_{l_k k})$, the closer $\underline{x}_t$ is to $M_{l_k k}$. We set the weight of $M_{l_k k}$ for $\underline{x}_t$ in the RF tied state l, i.e., w_{tlk}, to be directly proportional to the *P*-value $P_v(\underline{x}_t \mid M_{l_k k})$, that is

$$w_{tlk} = \frac{P_v(\underline{x}_t \mid M_{l_k k})}{\sum_{j=1}^{K} P_v(\underline{x}_t \mid M_{l_j j})}, k = 1, 2, \cdots, K. \quad (8)$$

4.3 Relative entropy based weight estimation

We next consider using relative entropy (R-entropy) in a set of acoustic models as the weight of the models. As discussed above, through using RF-based PDTs we obtain multiple sets of acoustic

models $MS_1,\cdots,MS_K$, with the set k having N_k physical models (GMDs of tied states) $M_{1k},\cdots,M_{N_k k}$. For a feature vector $\underline{x}_t$, the acoustic score $p(\underline{x}_t \mid M_{l_k k})$ measures the likelihood of $\underline{x}_t$ being emitted by a RF tied state ES_l using the model $M_{l_k k}$. If the values of $p(\underline{x}_t \mid M_{1k}),\cdots,p(\underline{x}_t \mid M_{N_k k})$ are spread out, then the kth set of acoustic models is discriminative. Relative entropy can be used to measure the distribution spread of acoustic scores within each set of acoustic models, which is defined as the Kullback–Leibler divergence (KLD) from the distribution of acoustic scores among the models in the kth set of acoustic models to the uniform distribution $1/N_k$, that is,

$$D_{tk} = \sum_{j=1}^{N_k} p_{tjk} \log \frac{p_{tjk}}{1/N_k} = \sum_{j=1}^{N_k} p_{tjk} \log p_{tjk} + \log N_k, \quad (9)$$

where the distribution of acoustic scores is defined by $p_{tjk} = p(\underline{x}_t \mid M_{jk}) / \sum_{n=1}^{N_k} p(\underline{x}_t \mid M_{nk})$. A large value of D_{tk} would indicate that the scores of the kth set of acoustic models deviate significantly from the uniform distribution of $1/N_k$. Therefore the weight of each model set can be made to be proportional to its relative entropy, defined as

$$w_{tk} = D_{tk} / \sum_{n=1}^{K} D_{tn}, \quad k = 1,2,\cdots K. \quad (10)$$

In addition to the three methods presented above, the acoustic scores may also be combined in a time-dependent way by using the maximum score (MAX), or an average of n-best scores out of K models (n-K). The weights may also be set uniformly (Uniform). In our experiments, these methods were evaluated along with the above proposed three methods and the results were compared.

5. EXPERIMENTAL RESULTS

5.1 Experimental setup

The proposed methods were evaluated on the Telemedicine automatic captioning system developed at the University of Missouri-Columbia [14]. The training and test speech datasets were extracted from healthcare providers' 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 analyses were made at a 10 ms frame rate with a 20 ms window size. Gaussian mixture density based hidden Markov models (GMD-HMM) were used for within-word triphone modeling, where each HMM had 3 emitting states, each state was modeled by a GMD with 16 Gaussian components, and the size of mixtures was optimized for the baseline. Speaker dependent acoustic models were trained for five speakers Dr1 - Dr5. The task vocabulary size is 46,480. Baseline acoustic models used the single PDT-based state tying, with the number of tied states or physical models averaged over the five speakers being 1429. The decoding engine is TigerEngine 1.0 [15]. Please refer to [4] for a detailed description of the experimental setup.

5.2 Experimental results

a. Performance vs. model combinations and forest sizes

We trained multiple sets of PDTs by using the proposed RF technique and obtained multiple sets of acoustic models. The number of phonetic questions m was set to be 200 out of the M = 216 questions (see below for a discussion on m) as defined in HTK [16]. The number of RF tied-states in general increases with m and the forest size K. The resulting number of RF-tied states averaged over the five speakers was 7478 when K equals 50. The acoustic scores were combined by using the proposed methods of MLE, P-value, and R-entropy, as well as the methods of maximum score (MAX), n-best out of K models (n-K), and simple average (Uniform). In addition, weights from MLE and relative entropy were averaged (MLE+R-entropy). Table I gives the performance in word recognition accuracy averaged over the 5 speakers' test sets.

Table I Word accuracies (%) averaged over five speakers.

(a) Baseline and n-best methods with different n-K.

Baseline	78.96
n-best (5-20)	80.49
n-best (5-50)	80.62
n-best (10-50)	80.79
n-best (10-100)	80.31
n-best (20-100)	80.56

(b) Other methods of model combination.

	K			
	10	20	50	100
MAX	80.35	80.41	79.95	79.70
Uniform	80.39	80.57	80.71	80.80
MLE	80.47	80.81	80.90	80.92
P-value	80.43	80.69	80.85	80.90
R-entropy	80.39	80.64	80.88	80.91
MLE+R-entropy	80.39	80.72	80.95	80.96

From Table I we observe that the proposed RF-based PDTs improved word recognition accuracy significantly, and the effect was dependent on the score combining method and the forest size. When the forest size was large (K=50, 100), using the averaged weights from MLE and relative entropy yielded best results. Uniform weight was a good choice as well, since it was simple to implement and its performance was competitive to the best results. The MAX method's performance deteriorated when K became large, since maximum score is susceptible to unreliable models, and as K increases, the possibility that the score of some unreliable model turns into a maximum score becomes larger.

b. Performance vs. size of question subset

The subset size m of phonetic questions for individual PDTs needs to be chosen to balance the need for maintaining qualities of individual PDTs and reducing correlations among the PDTs. We investigated the effect of different values of m on word recognition accuracy for a fixed forest size K=50. Table II summarizes the results, where we used the MLE method to estimate the weights of multiple acoustic models. We observe that when m=15, the word recognition accuracy is lower than the baseline, since now all the PDTs are very weak. As m increases, word recognition accuracy improves, but the accuracy differences among m=100, 150, 200 are small. As m increases further, word accuracy gets noticeably lowered again, since now the correlations among the trees become too high. It appears that for the current task m in the range of 100 to 200 can all be considered as good choices.

Table II The effect of question subset size m on word accuracy.

Subset size m	15	20	100	150	200	210
Word accuracy (%)	77.68	80.38	80.92	80.96	80.90	80.65

c. Performance vs. model complexity

We evaluated the performance of RFs with respect to the complexity of Gaussian mixture densities, i.e., the mixture size or the number of Gaussian components per GMD, with the state tying thresholds in the PDTs kept unchanged. Table III summarizes the word recognition accuracies with the number of Gaussian components per GMD varied to be 8, 16, 20 and 24. For the RF method, we set the question subset size m = 150 and used the MLE method to estimate weights for model combination. From Table III we observe that the RF method improved accuracy performance over the baseline in every evaluated mixture size. In the baseline, using a mixture size of 16 yielded best performance, but further increasing the size led to overfitting and thus decreased accuracy. In RF, accuracy performance improved with the mixture size, and at the mixture size of 24 overfitting still did not occur, indicating the robustness of the RF-based state tying to overfitting.

Table III Word accuracies vs. mixture size, with m=150 for RFs.

	Number of Gaussian components per GMD			
	8	16	20	24
Baseline	77.65	78.96	78.68	78.15
RF method, K=10	78.08	80.47	81.57	81.70
RF method, K=20	78.06	80.81	81.86	81.92

d. Comparison with output hypothesis integration

We next compared performance of our proposed RF method with the method of integrating output hypotheses proposed in [5]. In doing so, we trained multiple sets of PDTs by using n-best random split selection, where n equals 10, and each PDT tied state has 16 GDFs as in the baseline. We then carried out speech decoding evaluation K times by using K sets of acoustic models. For each speech utterance, multiple 1-best recognition hypotheses were generated with each obtained from one set of the acoustic models. We used confusion network (CN) [4] to combine the multiple hypotheses output, which is made equivalent to the hypothesis integration method of ROVER as used in [5]. For details of using CN to integrate recognition hypotheses, please refer to [4]. Table IV summarizes recognition word accuracies by using the PDTs generated by n-best random split selection followed by CN based hypothesis integration.

Table IV Word accuracy by n-best random split selection and CN.

K	10	20	50
Word accuracy (%)	79.88	79.97	80.24

From Table IV we observe that n-best random split selection with hypothesis integration at the output word level also improved accuracy performance, but the gain is less than what are obtained by our proposed method of RFs for acoustic model combination. Furthermore, combining output hypothesis at the word level requires running decoding search K times, which is much slower than our methods of combining acoustic scores when single processor computers are used.

We also conducted a significance test on performance differences between the proposed methods and the baseline method, as well as the method of n-best random split selection with hypothesis integration. The results show that our proposed methods improved the word recognition accuracy significantly over the two comparison cases. For details, please refer to [17].

6. SUMMARY

In this paper we have presented a novel Random Forests based technique of constructing phonetic decision trees for acoustic modeling in conversational speech recognition. We have introduced a mechanism of tying triphone states over PDTs of a RF. The method allows using more specific models than the conventional single PDT method for acoustic modeling without running into the problem of overfitting. We have proposed several methods to combine the acoustic scores of multiple models in decoding search. We have demonstrated through experimental results on a large vocabulary conversational speech recognition task that the proposed techniques significantly improved performance of word recognition accuracy, and on this task it achieved higher accuracy performance and faster decoding speed than using word hypothesis integration at recognition output.

7. REFERENCE

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

[2] C. Chesta, P. Laface and F. Ravera, "Bottom-up and top-down state clustering for robust acoustic modeling," *Proc. Eurospeech*, pp. 11-14, 1997.

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

[4] J. Xue and Y. Zhao, "Novel lookahead decision tree state tying for acoustic modeling," *Proc. ICASSP*, pp. IV-1133 — IV-1136, 2007.

[5] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," *Proc. ICASSP*, pp. I-197 – I-200, 2005.

[6] C. Breslin and M.J.F. Gales, "Complementary system generation using directed decision trees," *Proc. ICASSP*, pp. IV-337–IV-340, 2007.

[7] T. G. Dietterich, "Ensemble methods in machine learning," *Proc. of the First International Workshop on Multiple Classifier Systems*, pp. 1-15, 2000.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp.5-32, 2001.

[9] Random Forests classifier description, website of Leo Breiman, http://www.stat.berkeley.edu/~breiman/.

[10] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[11] T. Dietterich, "An experimental comparison of three methods for constructing ensembles of Decision Trees: Bagging, Boosting and Randomization," *Machine Learning*, vol. 1, no. 22, 1998.

[12] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. B*, vol. 39, no. 1, 1-38, 1977.

[13] J. Xue and Y. Zhao, "Random forest-based confidence annotation using novel features from confusion network," *Proc. ICASSP*, pp. I-1149 –I-1152, 2006.

[14] 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. ICASSP*, pp. I-957 – I-960, 2006.

[15] X. Li and Y. Zhao, "A fast and memory-efficient N-gram language model lookup method for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 21, pp. 1-25, 2007.

[16] HTK Toolkit, http://htk.eng.cam.ac. UK.

[17] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," to appear in *IEEE Trans. Audio, Speech, and Language Processing,* vol. 16, no. 2, 2008.