TREE-STRUCTURED MODELS OF PARAMETER DEPENDENCE FOR RAPID ADAPTATION IN LARGE VOCABULARY CONVERSATIONAL SPEECH RECOGNITION

Ashvin Kannan

Nuance Communications
1380 Willow Road
Menlo Park, CA 94025
ashvin@nuance.com

Sanjeev Khudanpur

Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
khudanpur@jhu.edu

ABSTRACT

Two models of statistical dependence between acoustic model parameters of a large vocabulary conversational speech recognition (LVCSR) system are investigated for the purpose of rapid speaker- and environment-adaptation from a very small amount of speech: (i) a Gaussian multiscale process governed by a stochastic linear dynamical system on a tree, and (ii) a simple hierarchical tree-structured prior. Both methods permit Bayesian (MAP) estimation of acoustic model parameters without parameter-tying even when no samples are available to independently estimate some parameters due to the limited amount of adaptation data. Modeling methodologies are contrasted, and comparative performance of the two on the Switchboard task is presented under identical test conditions for supervised and unsupervised adaptation with controlled amounts of adaptation speech. Both methods provide significant (1% absolute) gain in accuracy over adaptation methods that do not exploit the dependence between acoustic model parameters.

1. INTRODUCTION

Today's speaker-independent LVCSR systems contain from a few hundreds of thousands to several million acoustic model parameters. It is well known (*e.g.* [4, 8, 14]) that the performance of these systems improves dramatically if the model parameters are suitably adapted to test conditions, particularly when there is a mismatch between the acoustic environment or the speaker characteristics in the training and test speech. However, due to the large number of the parameters to be adjusted in comparison with the amount of realistically available adaptation data, one usually takes recourse to tying (perhaps hierarchically) the adjustments of individual acoustic models. The extent of tying depends on the amount of available adaptation speech – the less the data, the more one ties parameters. When only a few seconds of speech is available for adaptation, most common techniques resort to a single (global) adjustment of all acoustic model parameters.

Recently, statistical modeling of the model parameters themselves has received some attention [1, 6, 11, 12]. For the ease of discussion, consider an adaptation scheme in which one adjusts only the mean vectors (via an additive *bias*) of the Gaussian output densities of a hidden Markov model (HMM) based LVCSR system to a new speaker or environment. One may explicitly model correlation between these biases (*e.g.* [3, 7]) or have an implicit model (*e.g.* [6, 11, 12]). In either case, the resulting statistical model provides a framework to estimate *all* the biases (including those unseen in the adaptation speech) based on *all* available data. We investigate the performance of two such tree-structured schemes [6, 12] on Switchboard, a corpus of American English telephone conversations.

Section 2 begins with a brief description of a Gaussian multiscale process whose evolution (in scale) is governed by a linear dynamical system. The presentation is limited to the case of bias estimation for adaptation and the reader is referred [5] for details. This is followed by a brief review of another tree-based model, structural MAP, originally presented in [12]. A comparison of the two modeling methodologies is made. Section 3 then presents the main results[1] of this paper.

2. TREE-STRUCTURED MODELS OF DEPENDENCE

Two models of parameter dependence are considered here.

2.1. Multiscale Tree Processes

Multiscale stochastic processes are an important class of models, of which a particularly useful subclass is based on scale-recursive dynamics on trees [2, 9]. They allow efficient model estimation and likelihood calculation resulting in a variety of applications. Denoting a node in the tree by t with parent $t\bar{\gamma}$, a state-space model for the evolution in scale of the Gaussian tree-based process X and its noisy observation Y is given by

$$x(t) = A(t)x(t\bar{\gamma}) + w(t) \quad (1)$$

$$y(t) = C(t)x(t) + v(t) \quad (2)$$

where $x(t)$ is the state of the process at node t. The state $x(0)$ at the root node 0 has distribution $\mathcal{N}(0, \Sigma(0))$, where $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian density with mean μ and covariance Σ. The process noise $w(t)$ is white, independent of $x(0)$, and has distribution $\mathcal{N}(0, Q(t))$. The state $x(t)$ is observed via a noisy measurement $y(t)$, where the measurement noise $v(t)$ is white, independent of $x(0)$ and $w(t)$, and has distribution $\mathcal{N}(0, R(t))$. A tree with only one child per parent may be interpreted as the familiar linear dynamical system that evolves in time.

Given a tree topology and training observations Y, the parameters $\Sigma(0)$, $A(t)$, $Q(t)$, $C(t)$ and $R(t)$ of the tree are estimated using an Expectation-Maximization algorithm [6]. Given a tree with its parameters and observations Y at (possibly only a subset of) the nodes, the smoothed (MAP or MMSE) estimates of the states $\hat{x}(t|Y) = E\{x(t)|Y\}$ are computed via a generalization of the Rauch-Tung-Striebel algorithm [2].

Supported by National Science Foundation Grant No. IIS-9732388.

[1] This research was conducted at the 1998 Johns Hopkins University Summer Research Workshop on Language Engineering. See the "Rapid Recognizer Adaptation" project web page [15] for various details.

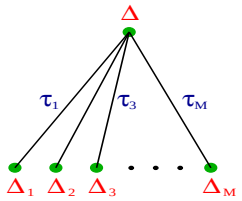
Let μ_i^{SI}, $i = 1, \ldots, M$, denote means of the Gaussian densities constituting the acoustic models in a speaker-independent LVCSR system, and $\mu_i^{SA} = \mu_i^{SI} + \Delta_i$ denote the corresponding speaker-adapted means. For the purpose of robust estimation, the M densities are partitioned into L classes $\mathcal{G}_l$, $l = 1, \ldots, L$, and one estimates only a *common*- or class-bias $x(l)$ shared by all the densities in the class, so that

$$\mu_i^{SA} = \mu_i^{SI} + x(l), \ \forall\, i \in \mathcal{G}_l. \tag{3}$$

The usual practice is to adjust the number of classes L in accordance with the amount of available adaptation data, and is usually of the order of 10 or smaller for rapid adaptation based on a few seconds of speech. We propose to use large values of L (150-250) even when a very small amount of speech is available for adaptation. To obtain robust estimates of the class-biases $x(l)$, we define a tree with L leaves, associate the leaf nodes with the $x(l)$'s, and model the biases as a Gaussian multiscale process given by (1). From a block of adaptation data, ML estimates $y(l)$ of biases at a subset of the leaves are used to obtain *a posteriori* estimates of all the biases, $\hat{x}(l)$. These smoothed bias estimates are used in (3).

2.2. Structural MAP Adaptation

The scheme described above bears strong resemblance to SMAP, an adaptation scheme based on maximum *a posteriori* probability under a structured prior presented in [12]. The SMAP scheme is briefly described next, so as to enable drawing parallels with multiscale models.



The speaker-dependent biases Δ_i are assumed to be jointly Gaussian with a hierarchical dependence structure. One may hypothesize that the class-bias Δ of the M densities is a random variable with a prior $\mathcal{N}(0, \tau_0^{-1})$, and that each individual bias Δ_i has a conditional prior density $\mathcal{N}(\Delta, \tau_i^{-1}\sigma_i^2)$. If $\gamma_i(n)$, $n = 1, \ldots, N$, denotes the posterior probability, computed by the forward-backward procedure, that the n-th acoustic frame o_n of the adaptation speech comes from the i-th output density, $\tilde{\Delta}$ denotes the ML estimate of Δ, and $\tilde{\Delta}_i$ that of Δ_i, then the MAP estimate of Δ has the form

$$\begin{aligned} \Delta &= \frac{\Gamma \cdot \tilde{\Delta} + \tau_0 \cdot 0}{\Gamma + \tau_0}, \quad \text{where} \quad \Gamma = \sum_i \frac{1}{\sigma_i^2} \sum_n \gamma_i(n), \\ \tilde{\Delta} &= \frac{1}{\Gamma} \sum_i \left[\frac{1}{\sigma_i^2} \sum_n \gamma_i(n)(o_n - \mu_i^{SI}) \right], \end{aligned} \tag{4}$$

and the conditional MAP estimates of the individual biases are

$$\begin{aligned} \Delta_i &= \frac{\Gamma_i \cdot \tilde{\Delta}_i + \tau_i \cdot \Delta}{\Gamma_i + \tau_i}, \quad \text{with} \quad \Gamma_i = \sum_n \gamma_i(n), \\ \tilde{\Delta}_i &= \frac{1}{\Gamma_i} \sum_n \gamma_i(n)(o_t - \mu_i^{SI}). \end{aligned} \tag{5}$$

The analysis and the formulae extend to a tree structured organization of the M biases Δ_i as shown in [12].

2.3. Comparison of SMAP and Multiscale Models

The SMAP methodology *imposes* a dependence structure between the Δ_i's through the choice of the τ_i's, while the multiscale approach estimates this dependence structure via training data. Equation (4) for obtaining Δ in the SMAP scheme is identical to the upward sweep in the state estimation formulae for the internal nodes of a multiscale tree with $A(t) = I$, $Q(t) = 0$, $\Sigma(0) = 0$. Once Δ has been estimated, the multiscale tree uses the *estimated* error covariances of the parent and the i-th child to obtain the MMSE (also MAP) estimate of Δ_i. In the SMAP procedure, the choice of the hyperparameter τ_i, together with the variance σ_i^2 of the i-th density imposes a covariance $\tau_i^{-1}\sigma_i^2$ on the parent's *a priori* estimate of Δ_i. Thus multiscale models enjoy the advantage that the designer need only be concerned with constructing a reasonable topology (and parameter tying) and the model parameters are estimated from training data. Multiscale trees are also more general in terms of the kinds of processes they can model (*e.g.* [9]), and while we have not formally demonstrated it here, we view SMAP as a special case of a multiscale tree model.

Our implementations of the SMAP and multiscale tree models are not perfectly comparable[2]. However, their adaptation performance may still be compared, as the task definition (unadapted system, adaptation speech, test set, *etc.*) is identical for both cases.

3. EXPERIMENTAL RESULTS ON SWITCHBOARD

The two adaptation schemes described above are tested on the Switchboard corpus of conversational telephone speech. We begin with a description of the LVCSR system and present the performance of the two schemes under various conditions.

3.1. Setup for Rapid Adaptation

The speaker-independent LVCSR system is HTK-based [13], with state clustered triphone acoustic models trained on about 60 hours of speech[3]. The test set comprises 19 conversations ($\sim$ 2 hours) with over 2400 utterances containing a total of about 18000 words. The speaker-independent system has a word error rate (WER) of **45.2%** on this test set. Bigram word lattices generated using the speaker-independent system are rescored by the adapted acoustic models described in the following.

In this paper, each adaptation scheme is tested under four conditions: • 30 seconds of supervised adaptation, • 30 seconds of unsupervised adaptation on transcriptions produced by the speaker-independent system, • 60 seconds of supervised adaptation and • 60 seconds of unsupervised adaptation.

[2] The topology of the trees used for SMAP and the multiscale model are different due to our current implementation. Furthermore, SMAP is implemented component-wise in the bias vector, corresponding to a diagonal covariance assumption, while the multiscale trees are implemented with full error covariance (albeit based on empirical variance of the class-biases across the training speakers).

[3] The front-end performs a PLP analysis at a 10ms frame rate, and subtracts the cepstral mean of each individual utterance from all the samples of that utterance. The system comprises about 7000 tied states each with an output modeled by a mixture of up to 6 Gaussian densities. The vocabulary contains 22K words. A standard back-off bigram language model trained on about 2.1 million words of transcribed text is used during decoding. No speaker normalization (*e.g.* Vocal Tract Length Normalization) is performed during training or testing.

3.2. Multiscale Trees

To use a multiscale model in an adaptation experiment, we need to first define (i) the L bias-classes, (ii) the multiscale tree topology with bias classes as the leaves, (iii) the model parameter tying (if any), and then estimate the model parameters from the class-biases of the training speakers (EM training).

Bias class definition: The HMM states from which the acoustic models of the triphones are constructed have an inherent hierarchical partition due to the triphone state-clustering used in our system. For instance, the initial states of all the triphones of the phoneme "aa" correspond to the leaves of a decision tree, all the middle states to the leaves of a second decision tree, and the final states to that of yet another decision tree. There are about 150 such trees (three corresponding to each phoneme) in the system. In the experiments described here, we take advantage of this fact to define: (i) 150 bias-classes, each containing all states of one state-clustering tree, and (ii) 250 bias-classes obtained by starting with the 150 classes and dividing them further exactly as we do for constructing the state-clustering trees in HTK [13], and stopping when 250 leaves are obtained. Thus the 250 class partition is a refinement of the 150 classes but coarser than the inherent system partition. Gaussians in each class are tied for adaptation according to (3).

Tree topology definition: In this paper, we do not address the issue of optimal dependence tree topology [10] for adaptation. We instead use a reasonable topology which groups biases of Gaussian means of one phoneme closer to each other and models the dependence across phonemes by estimating the parameters of the multiscale tree model. We model the overall dependence of the biases by defining a superstructure on top of the 150 roots of the state-clustering trees. The resulting trees, viewed top-down, divide all

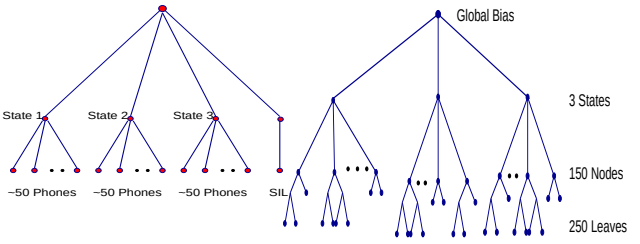


Figure 1: Modeling dependence above the 150 triphone state clusters using multiscale trees with (a) 150 and (b) 250 bias-classes.

the Gaussian densities in the system according to their state number within the model of the triphone and then between phonemes. Figure 1 illustrates dependence structures with 150 and 250 leaves that we have investigated.

Another axis that we explore is that of structure in the dependence between phonetic classes: (i) a flat structure in which all the 50 phonemic classes are descendants of a common ancestor as shown in Figure 1(a), and (ii) a structure with 11 intermediate nodes each governing a subset of the 50 phonemic classes. The partitioning of the 50 phonemes into the 11 subsets is based on articulatory-phonetic features. Figure 2 illustrates a modification of Figure 1(a) to provide such additional structure.

The multiscale model also allows the use of additional independent observations $y(t)$ in the internal nodes of the tree. For our bias adaptation experiments, this amounts to providing additional estimates of the bias of a *cluster* of some of the 150 or 250 biases. We investigate the use of one such additional measurement: the

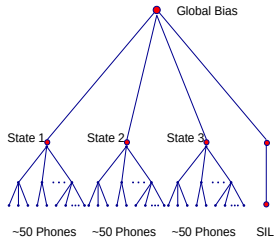


Figure 2: A 150-leaf multiscale tree with richer structure

Mode and Dur (sec.)	ML Est.	150 Tree	Richer Struct	Global Bias
Sup (30)	44.3%	43.2%	43.2%	43.1%
Unsup (30)	45.0%	44.1%	43.8%	44.0%
Sup (60)	42.9%	42.2%	42.1%	42.3%
Unsup (60)	44.6%	43.7%	43.6%	43.6%

Table 1: Recognition WER for 150 bias classes with • ML estimates, • multiscale tree of Figure 1(a), • Figure 2, and • Figure 1(a) with additional global information

global bias of all the Gaussian components in the system. By providing this estimate at the *root* of the tree, one may expect that a robust "anchor" is provided at the root when the data at the leaves are really sparse, and the internal bottom-up estimate at the root is susceptible to "system noise." The use of such additional information is investigated for both dependence structures of Figure 1.

Tree model parameter tying definition: The tree model parameters can be tied, i.e., a group of nodes may share the same A parameter. Of course, tying the tree parameters does not imply tying the smoothed estimates $\hat{x}(l)$ at the nodes. For the 150-leaf models, each node in the tree has its own set of parameters (A, Q). For the 250-leaf models, the nodes below the corresponding leaves of the 150-leaf tree share the same (A, Q). Therefore the number of parameters is equal in the 150- and 250-leaf models.

Parameter training: The L biases $\{y(l),\ l = 1, \ldots, L\}$ of each training speaker contribute one sample of the multiscale model. The tree model parameters $(\Sigma(0), A(t), Q(t))$ are estimated from the samples over all the training speakers using an EM algorithm. Details of the training procedure are in [5, 6].

Results: Recognition results for three dependence models with 150 leaves are presented in Table 1. The columns labeled ML correspond to using the ML estimate $y(l)$ of the bias of class l where available and backing off to the global ML bias where it is not. The recognition results for two dependence models with 250 leaves are presented in Table 2. There is thus a 0.6-1.1% (absolute) reduction in WER over the ML scheme depending on adaptation conditions. Also, multiscale models provide a 1% improvement even when only 30 seconds of unsupervised adaptation is performed, where the ML scheme provides almost no gains or even degrades. In the-

Mode and Dur (sec.)	ML Est.	250 Tree	Global Bias
Sup (30)	44.9%	43.0%	43.2%
Unsup (30)	45.5%	44.1%	44.1%
Sup (60)	43.0%	42.1%	42.0%
Unsup (60)	44.5%	43.2%	43.2%

Table 2: Recognition WER for 250 bias classes with • ML estimates, • multiscale tree of figure 1(b) without and • with additional global information

ory, a larger tree (250) should be no worse than a smaller one (150) even if there is little data (30 sec.) since the estimates are smoothed in both cases, and we do find this to be true. Larger trees perform better, as expected, when more data is available.

3.3. Structural MAP

A hierarchical prior is used in the SMAP scheme which assumes that the biases of the Gaussian densities in the 50 phonetic classes are mutually independent. Within each phonetic class, the topology of Figure 3 is used to first estimate a bias for all the components of a phone, then a bias for each of the three states of the phone, and finally that of each individual Gaussian density. Note

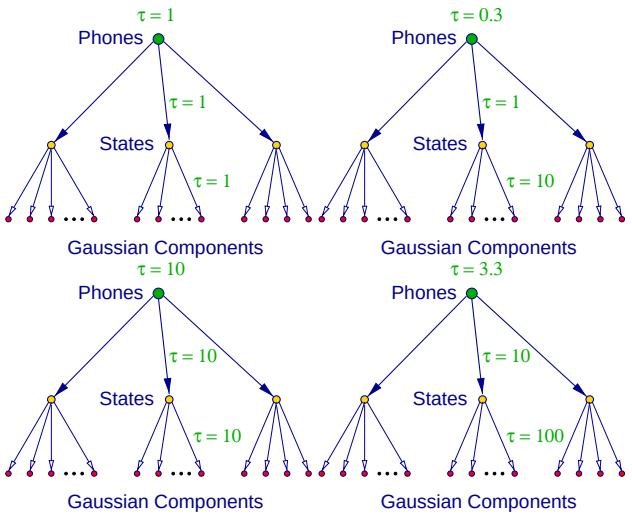


Figure 3: Structured priors used for SMAP adaptation

Mode and Dur (sec.)	Hyperparameters τ_0, τ_1, τ_2			
	1,1,1	10,10,10	0.3,1,10	3.3,10,100
Sup (30)	44.0%	43.8%	43.2%	44.1%
Unsup (30)	45.4%	44.8%	44.5%	44.6%
Sup (60)	43.3%	43.2%	42.2%	43.6%
Unsup (60)	45.5%	44.7%	44.5%	44.7%

Table 3: Recognition WER (%) for SMAP adaptation for different values of the hyperparameters (c.f. Figure 3).

from the figure that different values of the hyperparameters (τ) are investigated and the recognition performance is presented in Table 3. Also observe that for supervised adaptation, SMAP provides almost the same gains over a comparable ML adaptation scheme as the corresponding multiscale methods in Table 2 (1% absolute reduction in WER). SMAP seems to be less effective than multiscale trees in the unsupervised case.

4. CONCLUSIONS

In this paper we compared two tree-based models: SMAP and Multiscale models for rapid speaker adaptation. Experimental results on Switchboard show that both models yield greater reduction in WER ($\sim$1% absolute) across a range of conditions than ML techniques which do not exploit parameter dependence.

We find the two schemes to be competitive, with the multiscale tree performing a little better, particularly for unsupervised adaptation. We also find richer dependence structure to be slightly beneficial. There is no gain from the use of the global ML bias in the multiscale model, probably due to the fact that is not independent of the class-biases (considered jointly).

5. ACKNOWLEDGMENTS

William Byrne implemented the baseline HTK system and generated the lattices for further experimentation. Vassilis Digalakis generated the ML estimates of the biases for the training and test speakers which form the inputs to the multiscale tree models. Mari Ostendorf contributed significantly to the development of the framework of multiscale tree processes for adaptation [5].

6. REFERENCES

[1] S. Chen and P. DeSouza, "Speaker Adaptation by Correlation (ABC)," in *CSR Hub-4 DARPA Speech Workshop*, 1997.

[2] K. Chou, A. Willsky, and A. Benveniste, "Multiscale recursive, Data Fusion and Regularization," *IEEE Transactions on Automatic Control*, pp. 464–478, 1994.

[3] S. Cox, "A Speaker Adaptation Technique Using Linear Regression," in *Proc. ICASSP*, pp. 700–703, 1995.

[4] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Trans. Speech Audio*, Sept. 1995.

[5] A. Kannan, *Adaptation of Spectral Trajectory Models for LVCSR*, PhD thesis, Boston University, 1997.

[6] A. Kannan and M. Ostendorf, "Modeling Dependence in Adaptation of Acoustic Models Using Multiscale Tree Processes," in *Proc. EUROSPEECH*, pp. 1863–1866, 1997.

[7] M. Lasry and R. Stern, "*A Posteriori* Estimation of Correlated Jointly Gaussian Mean Vectors," *IEEE Trans. PAMI*, vol. PAMI-6, pp. 530–535, July 1984.

[8] C. Leggetter and P. Woodland, "Speaker Adaptation Using Maximum Likelihood Linear Regression," *Comp Speech Lang*, vol. 9, no. 2, pp. 171–185, 1995.

[9] M. Luettgen, W. Karl, A. Willsky and R. Tenney, "Multiscale Representations of Markov Random Fields," *IEEE Trans. Sig Proc.*, vol. 41, pp. 3377–3396, Dec. 1993.

[10] O. Ronen and M. Ostendorf, "A Dependence Tree Model of Phone Correlation," in *Proc. ICASSP*, vol. 2, pp. 873–876, May 1996.

[11] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *IEEE Trans. Speech Audio*, vol. 5, pp. 183–191, Mar. 1997.

[12] K. Shinoda and C-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE ASRU*, pp. 381–387, 1997.

[13] S. Young, J. Jansen, J. Odell, D. Ollasen, P. Woodland, *The HTK Book (Version 2.0)*, Entropic Cambridge Research Laboratory, 1995.

[14] G. Zavaliagkos, R. Schwartz, J. McDonough, and J. Makhoul, "Adaptation algorithms for large scale HMM recognizers," in *Proc. EUROSPEECH*, pp. 1131–1134, 1995.

[15] http://www.clsp.jhu.edu/ws98/projects/adapt, Rapid Recognizer Apadtaion team web page.