MODELING DEPENDENCY IN ADAPTATION OF ACOUSTIC MODELS USING MULTISCALE TREE PROCESSES

Ashvin Kannan and Mari Ostendorf

Electrical and Computer Engineering Department
Boston University, 44 Cummington Street, Boston, MA 02215, USA
http://raven.bu.edu/~{ashvin,mo}

ABSTRACT

To adapt the large number of parameters in a speech recognition acoustic model with a small amount of data, some notion of parameter dependence is needed. We present a dependence model to relate parameters in a parsimonious framework using a Gaussian multiscale process defined by the evolution of a linear stochastic dynamical system on a tree. To adapt *all* classes from *all* adaptation data, we formulate adaptation as optimal smoothing of the tree process. This approach is used to adapt two types of models: Gaussians, and Gaussian processes (segment models) characterized by a polynomial mean trajectory. Recognition results presented on the Switchboard corpus show improvements in supervised and unsupervised modes.

1. INTRODUCTION

Adaptation of acoustic models has become increasingly important to improving performance of a speaker-independent speech recognition system. Of particular interest are approaches that allow adaptation of a large number of parameters, like that in a large vocabulary continuous speech recognition system, with a small amount of adaptation data. For example, in a system based on K Gaussian models, transform all Gaussians with observations from N Gaussians where $K \gg N$. One approach to this problem involves partitioning the K Gaussians into L classes and estimating one transformation for all Gaussians in a class, based on adaptation data for that class. To adapt classes with no observations, one can define a hierarchy of classes, e.g. a tree with L leaves, and "back-off" to a broader transformation. Alternatively (or in addition), models of dependence between means of classes have been recently attempted using extended-MAP [11, 14, 6, 2] and linear regression [4, 1]. For implementation feasibility, these methods assume that components of the feature vector are uncorrelated, along with using a small L.

An important class of dependence models relate the L classes by making Markovian assumptions on the dependence structure. They represent the joint correlation in terms of low order conditional distributions and hence use a relatively small number of parameters to characterize the dependence without heuristics or approximations. The only speech application of a Markovian model to adaptation so far is based on Markov random fields (MRF) [13]. In this paper, we present another Markovian model that relates entire vectors (the means of the classes) using a Gaussian multiscale process defined by the evolution of a linear stochastic dynamical system on a tree. To adapt *all* classes from *all* adaptation data, we formulate adaptation as optimal smoothing of the tree process. We use this approach to adapt two types of distributions for acoustic models: Gaussians, and Gaussian processes characterized by a polynomial trajectory for the mean. In Section 2, we introduce the multiscale model with the associated smoothing and training algorithms. Section 3 shows how the model can be used for adaptation, and experimental results are given in Section 4. The key results are summarized in Section 5.

2. MULTISCALE TREE PROCESSES

Multiscale stochastic processes represent an important class of models, of which a particularly useful class is based on scale-recursive dynamics on trees [3, 12]. 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 can be written as

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

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

where $x(t)$ is the state of the process at node t. The root node state $x(0)$ has distribution $\mathcal{N}(0, \Sigma(0))$, where

This work was supported by the U.S. Department of Defense, ONR Grant N00014-92-J-1778.

$\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian 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 $\mathcal{N}(0, R(t))$. We allow the branching to be non-uniform.

Given Y, the set of all available measurements at the nodes (possibly at a subset of nodes), the smoothed estimate of the state $\hat{x}_s(t) = E\{x(t)|Y\}$ and the associated error covariance $P(t|Y) = E\{[x(t) - \hat{x}_s(t)][x(t) - \hat{x}_s(t)]^T\}$ can be computed using a generalization of the Rauch-Tung-Striebel (RTS) algorithm [3, 7]. Smoothing is done in two sweeps: an upward sweep from the leaves to the root, followed by a downward one from the root to the leaves. Maximum-likelihood estimates of the parameters of the tree process $(\Sigma(0), A(t), Q(t), C(t), R(t))$ can be obtained by applying the RTS and Expectation-Maximization (EM) algorithms to multiple independent sample paths of the process [10, 7].

Multiscale models offer a number of advantages over MRFs including the ability to tie parameters of varying degrees, which is useful if we have a limited amount of training data, and efficient, non-iterative, recursive and parallelizable algorithms for state estimation.

3. APPLICATION TO ADAPTATION

Let $\mathcal{G}_l$ denote the Gaussians in class l and μ_i the mean of one Gaussian. We model adaptation for the Gaussians in class l by a common shared shift $x(l)$:

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

where μ_i^a denotes the mean μ_i after adaptation. Such a shared shift approach has been used for Gaussians in hidden Markov models (HMMs) [14] and the stochastic segment model (SSM) [7], and for polynomial segment models (where a "shift" is also polynomial) [9].

ML estimates for the shift $\hat{x}(l)$ and associated error covariance $P(l)$ can be obtained from adaptation data for each class l independently [9]. Our goal is to obtain smoothed estimates of the shifts $\hat{x}_s(l)$ for *all* classes, using adaptation information from *all* the observed classes in the form of $\hat{x}(l), P(l)$, $l \in$ a subset of $\{1 \ldots L\}$. Define a Gaussian tree-based shift process (Equation 1) with L leaves, and associate the leaf node states with the shifts of the L classes we wish to model dependency between. Given $\hat{x}(l), P(l)$ at the leaves, we can compute $\hat{x}_s(l), P_s(l)$ using the tree RTS smoother. Due to the Bayesian nature of the smoothing, as the amount of adaptation data increases for a leaf, the smoothed shift approaches the unsmoothed shift and the estimated parameters will converge to the standard ML speaker-dependent estimate.

Observation Model. The usual dynamical system formulation includes an observation equation for the shifts (Equation 2). Here, a variable number of independent observations $y_i(l)$ are associated with leaf nodes l, and no observations are associated with internal nodes. All observations associated with a node are factored into $\hat{x}(l)$ and $P(l)$, so an explicit observation equation is not needed.

Class Definition and Topology. Context-dependent models are typically clustered in the form of a tree (e.g. ML clustering of Gaussians and PSMs [8]) for each region (or state) of a phone. Figure 1(a) indicates the tree for one region. Each node of the tree represents an equivalence class of triphones. Nodes at a certain "cut" through the tree defines terminal adaptation classes to share shifts (Equation 3), the boxes in Figure 1(a). One popular, but ad-hoc, option for adaptation is the "back-off" strategy where the shift is computed at the most detailed node which has more than T_S (shift threshold) adaptation frames, and this is copied to all child terminal adaptation classes as shown in Figure 1(b).

The topology of the clustering tree is also used for multiscale smoothing. Class-dependent shifts are computed only at these terminal adaptation nodes (Figure 1(c)), which are then smoothed using the multiscale model (Figure 1(d)) to get smoothed shift estimates at *all* nodes. For a fair comparison of MS vs. back-off approaches the same topology is used in both cases.

Parameter Estimation. The tree process parameters ($\Sigma(0)$, $A(t)$ and $Q(t)$) are estimated from training data using the EM algorithm. (There is no need to estimate $C(t)$ and $R(t)$ as we do not have an explicit observation equation, as mentioned earlier.) Here, the A and Q parameters are shared among all nodes of a phone, i.e. for N phones, there are N sets of (A, Q) parameters for the tree. Each training speaker contributes one "sample path" of the multiscale process.

To start the EM iterations, we need initial estimates of $\Sigma(0)$, the A's and the Q's. For each speaker in the training set we compute covariances of ML (unsmoothed) shifts at each terminal shift node. A frequency-weighted average of these covariances across all speakers is used *both* for $\Sigma(0)$ and Q. We initialize all the Qs in the system to be the same and all $A = I$.

4. EXPERIMENTS AND RESULTS

Experiments were conducted on the Switchboard corpus which has telephone-quality conversational speech. The feature vector consisted of the first 14 mel-warped cepstra (normalized with cepstral mean subtraction and for vocal tract length [5]) computed every 10 msec,

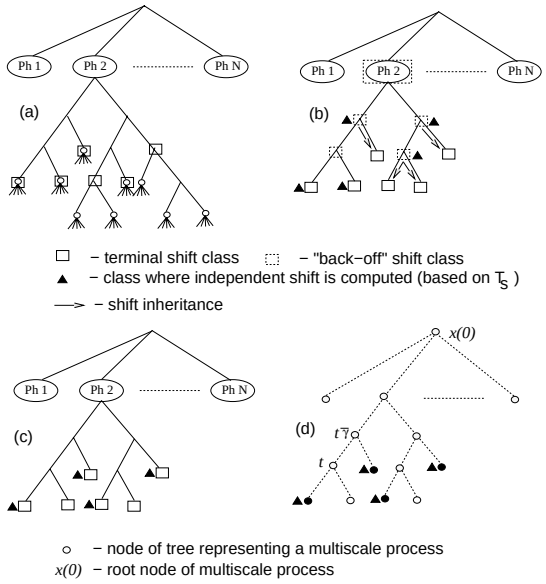

Figure 1: *Trees used for adaptation: (a) shows the clustering tree with terminal adaptation classes, (b) is for the "back-off" method of adaptation, while (c) and (d) are for the multiscale smoothing approach.*

their first differences and the first difference of log energy. Results for adaptation of both SSMs (Gaussians) and PSMs are reported. The PSM systems used a 2-region model, with each region modeled by a linear trajectory Gaussian process with a single full covariance. The SSM systems used a 5-region model, with each region represented by a full covariance Gaussian. Both cases used gender-dependent models and ML clustered triphones. The PSM and SSM adaptation systems had 300 and 150 terminal adaptation classes/region respectively. Sixty hours of speech are used for training the acoustic models and multiscale model for most experiments; 136 hours are used in the guided adaptation experiments.

Recognition is done using N-best rescoring: the top N (=100) word-sequence hypotheses provided by BBN's Byblos system are rescored by the acoustic model (SSM or PSM) and reranked by linearly combining the log acoustic score with the number of words and phones in the sentence (insertion penalties), the trigram language model score and the duration score (based on relative frequency) to minimize average word error in the top ranking hypotheses. Recognition is performed on a development test set comprising 7 conversations (14 speakers and 6381 words, with an average of 2.3 min speech/speaker).

Batch Mode Adaptation. In batch adaptation, the first half of each conversation is used as adaptation data and the second half for testing. Results in Table 1 for supervised adaptation indicate the MS-smoothing is better than the back-off MAP approach. In other experiments [7], we found gains from MS-smoothing relative to the back-off approach to increase as the amount of adaptation data is reduced, demonstrating the potential of the multiscale model for rapid adaptation. In the unsupervised mode, both the back-off and MS systems were usually worse than the speaker-independent baseline.

Table 1: *Supervised batch recognition with 2-region PSMs. Error rates on the second half.*

SI baseline	MAP back-off	MS
44.5%	44.1%	43.9%

Transcription Mode Adaptation. In transcription mode adaptation two passes are made over the speech: the first to collect statistics for adaptation after Viterbi alignment, and the second to perform recognition with the adapted models. The baseline unadapted error rate was 43.2% while the best back-off (ML) and MS case with the PSM resulted in 42.7% and 42.8% respectively. These sets of experiments do not indicate any advantage of using MS-smoothing over the back-off system.

Guided Adaptation. In unsupervised batch and transcription modes, we do not see a gain in using the multiscale model over the back-off method. We conjecture that this is explained by two main limitations in the use of any model of dependence: 1) the model is trained to "learn" dependence between correct observations of different sound classes but we use it at a high error rate, and 2) in conventional transcription-mode adaptation the *same* data used to estimate the adaptation transformation in the first pass is scored with the adapted models in the second pass, i.e. there are potentially no new classes for the algorithm to generalize to. In this case, there is little advantage to using any model of dependence, and it is likely that ML techniques for adaptation will be better.

Both limitations are addressed with *guided* adaptation, i.e. adapting only with data from a subset of words recognized with a high degree of confidence. This serves to lower the error-rate for the speech used in adaptation, as well as tests the ability of the adaptation approach to handle unseen classes (in the "incorrect" parts of the speech).

We use a simplistic measure of word confidence based on the relative frequency of that word appearing in that position (according to a dynamic programming alignment) in the hypotheses. Experiments show that, in

unsupervised transcription mode, guided ML adaptation is better than non-guided ML adaptation [7]. In Table 2 we show that guided-MS adaptation is better than guided-ML adaptation for a 5-region SSM. The conditions are different from earlier experiments in that the test set consists of 7 conversations each of Switchboard and CallHome, improved language models and signal processing, and acoustic models trained on 123 hours of speech.

Table 2: *Guided unsupervised transcription mode adaptation with a 5-region SSM system.*

baseline	guided-ML	guided-MS
40.9%	40.4%	40.0%

Cost of Multiscale Smoothing. The computational complexity of the tree RTS smoother is $O(d^3 n)$ where d is the dimensionality of the state (shift process) and n is the number of nodes (internal+leaf) in the tree. The memory requirement is $O(d^2 n)$. The algorithm is inherently parallelizable, though our implementation was on serial machines. Since the E-step of the EM algorithm for parameter estimation uses the RTS smoother, the complexity of training is $O(d^3 nrp)$, where r is the number of runs of the multiscale process (i.e. the number of training speakers) and p is the number of EM iterations. The M-step computational needs are insignificant in comparison to the E-step.

The computational and storage costs of using multiscale smoothing is a small part of the recognition needs. For two trees of $L = 300$ each and $d = 29$, the memory image of the recognizer for adaptation increases by about 9% relative to the unadapted baseline. An iteration of the EM algorithm runs in 0.2 times real time on a Sun Ultra-1, and 3 training iterations were run. Adaptation costs for recognition are similar.

5. SUMMARY

We proposed a new dependence model based on a multiscale tree process, which allows one to optimally estimate shifts to adapt *all* models taking into consideration *all* adaptation data. Our approach provides a unified framework to handle classes with and without observations, and the adaptation converges asymptotically to standard ML speaker-dependent estimates as data from a particular speaker increases. The dependence model permits parameter tying of varying degrees, which is useful if we have a limited amount of training data. Efficient algorithms exist for smoothing and parameter estimation of such a process.

Experimental results on the Switchboard corpus indicate improvements with small amounts of data in supervised adaptation using multiscale smoothing, relative to ML and standard MAP adaptation. For unsupervised transcription mode adaptation, we show that guided multiscale smoothing gives maximum adaptation gains.

6. REFERENCES

[1] S. Ahadi and P. Woodland, "Rapid speaker adaptation using model prediction," *ICASSP-95*, pp. 684–687.

[2] S. Chen and P. DeSouza, "Speaker adaptation by correlation (ABC)," In *CSR Hub-4 DARPA Speech Workshop*, February 1997.

[3] K. Chou, A. Willsky and A. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. on Auto. Control*, 39(3):464–478, 1994.

[4] S. Cox, "A speaker adaptation technique using linear regression," *ICASSP-95*, pp. 700–703.

[5] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *ICASSP-96*, pp. 346–348.

[6] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *ICSLP-96*, pp. 985–988.

[7] A. Kannan, *Adaptation of Spectral Trajectory Models for Large Vocabulary Continuous Speech Recognition.* PhD thesis, Boston University, 1997. Available from `ftp://raven.bu.edu/pub/reports`.

[8] A. Kannan and M. Ostendorf, "A comparison of constrained trajectory models for large vocabulary speech recognition," Technical Report ECE-96-007, Boston University, September 1996. Available from `ftp://raven.bu.edu/pub/reports`.

[9] A. Kannan and M. Ostendorf, "Adaptation of polynomial trajectory segment models for large vocabulary speech recognition," *ICASSP-97*, pp. 1411–1414.

[10] A. Kannan, M. Ostendorf, D. Castañon, and W. Karl, "ML parameter estimation of a multiscale tree process using the EM algorithm," Technical Report ECE-96-009, Boston University, November 1996. Available from `ftp://raven.bu.edu/pub/reports`.

[11] M. J. Lasry and R.M. Stern. "*A Posteriori* estimation of correlated jointly Gaussian mean vectors." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(4):530–535, 1984.

[12] M. Luettgen, W. Karl, A. Willsky, and R. Tenney, "Multiscale representations of Markov random fields," *IEEE Trans. on Signal Proc.*, 41(12):3377–3396, 1993.

[13] B. Shahshahani, "A Markov random field approach to Bayesian speaker adaptation," *IEEE Trans. on Speech and Audio Proc.*, 5(2):183–191, 1997.

[14] G. Zavaliagkos, R. Schwartz, J. McDonough, and J. Makhoul, "Adaptation algorithms for large scale HMM recognizers," *Eurospeech-95*, pp. 1131–1134.