ADAPTATION OF POLYNOMIAL TRAJECTORY SEGMENT MODELS FOR LARGE VOCABULARY SPEECH RECOGNITION

Ashvin Kannan and Mari Ostendorf

Electrical and Computer Engineering Department
Boston University, 44 Cummington Street, Boston, MA 02215
{ashvin,mo}@bu.edu

ABSTRACT

Segment models are a generalization of HMMs that can represent feature dynamics and/or correlation in time. In this work we develop the theory of Bayesian and maximum-likelihood adaptation for a segment model characterized by a polynomial mean trajectory. We show how adaptation parameters can be shared and adaptation detail can be controlled at run-time based on the amount of adaptation data available. Results on the Switchboard corpus show error reductions for unsupervised transcription mode adaptation and supervised batch mode adaptation.

1. INTRODUCTION

Conventional hidden Markov models (HMMs) do not impose any constraints on spectral trajectories, allowing them to be discontinuous in time within a state. However, we know that the smooth motion of the articulators imposes natural constraints on the trajectory. Recently, researchers have proposed parametric models for spectral trajectories in the form of polynomials [1, 2]. These models compare favorably with HMMs on constrained tasks like vowel-classification [1, 3], isolated word recognition [2] and connected digit recognition [4]; and recent results show they are also viable for large vocabulary continuous speech recognition (LVCSR) applications [6]. However, no adaptation algorithms have been reported for such models.

In this work, we develop the theory for Bayesian adaptation of a polynomial trajectory segment model (PSM), and explore transformation tying in the context of an LVCSR system. The PSM has the advantage that it allows adaptation of the *entire* trajectory with *all* observations of the segment. For HMMs, this is only possible by making assumptions such as a tied constant transformation across all states in the phone.

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

We use a PSM clustering algorithm to define parameter tying for robust context-dependent model parameter estimation, and also for variable control of adaptation detail based on amount of adaptation data. Experimental results show gains from adaptation in unsupervised transcription-mode recognition on the Switchboard corpus. We begin by defining the PSM and clustering algorithm in Section 2. Section 3 develops the theory for Bayesian adaptation of the PSM. Experimental results and conclusions are in Section 4 and Section 5.

2. BASIC PSM AND CLUSTERING

In the PSM, a sequence of d-dimensional frame-level observations is assumed to be generated by a Gaussian process with a mean modeled by a sampled polynomial trajectory. Assuming the covariance Σ is constant in a segment, the likelihood of an n-length observation segment $Y = [y_1, \ldots, y_n]$ is given by

$$p(Y|B, \Sigma) = P(n) \prod_{j=1}^{n} p(y_j|\mu_j, \Sigma), \qquad (1)$$

where $P(n)$ is the duration probability and the sequence of means used in computing the likelihood is defined as $[\mu_1 \ldots \mu_n] = BZ^n$. B is a $d \times r$ matrix of coefficients for polynomial order $r-1$, and Z^n is an $r \times n$ time sampling matrix. Maximum likelihood (ML) estimates of B and Σ are computed as in [1].

Robust context-dependent models are obtained by tying parameters via distribution clustering. ML clustering of Gaussians, e.g. [5], is a standard technique used for estimating robust context-dependent models in speech recognition. In [6], we show that this technique can be extended to cluster PSMs. For both Gaussians and PSMs, the clustering algorithm is a binary tree growing procedure that successively partitions the observations (splits a node in the tree), at each step minimizing an ML splitting criterion over a pre-determined

set of allowable binary partitions. In the PSM case, observations are whole segments and each node represents a Gaussian process with a polynomial mean and fixed covariance. Triphone models are clustered with candidate partitions found by asking linguistically motivated questions.

3. ADAPTATION THEORY

If the time warping matrices ($\{Z^n\}$) are shared by all PSMs, the parameters B_k and Σ_k characterize a model k. We investigate adapting the trajectory mean, given by B_k, taking either an ML or Bayesian approach. In conventional Bayesian or Maximum *a posteriori* probability (MAP) estimation, the new mean B_k^a is estimated from a prior for B_k and sample observations of model k. Here, we instead choose to estimate and have a prior for the shift, F, where $B_k^a = B_k + F$, as in [7] for HMMs. Estimating B_k^a indirectly in this way has the advantage that the shift can be shared across a class of models. We first consider the case where each model has a different shift followed by the case where a class of models share a shift. Then we describe a method to estimate the shift priors and define classes of models that share a shift.

MAP estimate of model-dependent shift

If $B = [b_1 \dots b_r]$, define $B_V \stackrel{\triangle}{=} [b_1' \dots b_r']'$, i.e. a $(dr \times 1)$ column vector formed by stacking the columns of B. We can write the adaptation equation for the new mean B^a as a shift $B_V^a = B_V + F_V$. The MAP estimate of the shift for a model given a set of independent segment observations $\mathcal{Y} = \{Y_i\}$ where $Y_i = [y_{i1}, \dots, y_{in_i}]$ is

$$\begin{aligned}\hat{F}_V &= \underset{F_V}{\operatorname{argmax}}[p(F_V|\mathcal{Y}, B^a, \Sigma)] \\ &= \underset{F_V}{\operatorname{argmax}} \log[p(F_V) \prod_i p(Y_i|B^a, \Sigma)] \stackrel{\triangle}{=} \underset{F_V}{\operatorname{argmax}} Q.\end{aligned}$$

The log-likelihood of the observations is

$$\begin{aligned}\log \prod_i p(Y_i|B^a, \Sigma) &= -\frac{1}{2} \sum_{i,j} g_{ij}' \Sigma^{-1} g_{ij} + K \\ \text{where } g_{ij} &= y_{ij} - (B+F) z_{ij} \\ &= e_{ij} - F z_{ij} \\ &= e_{ij} (z_{ij}' \otimes I_d) F_V,\end{aligned}$$

$e_{ij} = y_{ij} - B z_{ij}$, z_{ij} is the jth column of Z for Y_i, I_d is a $d \times d$ identity matrix, $\otimes$ denotes the Kronecker product, and K includes all terms without F. Assuming $F_V \sim \mathcal{N}(0, \Sigma_{F_V})$, we set $\frac{\partial Q}{\partial F_V} = \mathbf{0}$ to get

$$\begin{aligned}\Sigma_{F_V}^{-1} \hat{F}_V &+ \sum_{i,j} [-(z_{ij} \otimes I_d) \Sigma^{-1} e_{ij} \\ &+ (z_{ij} \otimes I_d) \Sigma^{-1} (z_{ij}' \otimes I_d) \hat{F}_V] = \mathbf{0},\end{aligned}$$

or

$$[\Sigma_{F_V}^{-1} + C \otimes \Sigma^{-1}] \hat{F}_V = \sum_{i,j} z_{ij} \otimes \Sigma^{-1} e_{ij}, \qquad (2)$$

where C is a $r \times r$ matrix with entries

$$c_{mn} = \sum_{i,j} z_{ij(m)} z_{ij(n)}.$$

This represents a linear system of dr equations which can be solved for $\hat{F}_V$. Note that the Gaussian prior is similar to that used in [4], but the solution in our case is slightly different because a shift F is modeled rather than the mean B for parameter tying in adaptation.

MAP and ML estimate of class-dependent shift

If a class of models is constrained to share the same shift F, the adaptation equation in this case is $B_{V,k}^a = B_{V,k} + F_V$, where k is the model index. The resulting equation for the MAP estimate has sums over k for all models that share the same shift.

$$[\Sigma_{F_V}^{-1} + \sum_k C_k \otimes \Sigma_k^{-1}] \hat{F}_{V,MAP} = \sum_{i,j,k} z_{ij} \otimes \Sigma_k^{-1} e_{ij} \quad (3)$$

where C_k is the C matrix computed for class k. If the ML estimate of the shift is desired, it can be found by ignoring the prior term $\Sigma_{F_V}^{-1}$ in Equation 3.

Prior Estimation

To estimate the prior of the shared shift for models within a class, we first compute the best shared shift in the ML sense using all the observations for the class in the utterance. Collecting these shared shifts, we compute the covariance Σ_{F_V} for the prior $F_V \sim \mathcal{N}(0, \Sigma_{F_V})$. The shifts can be based on a "leave-some-out" paradigm, i.e. train the B, Σ parameters on a part of the data and estimate the shifts on the remaining part, cycling over all parts of the data or by using all the data without cycling for B, Σ and the shifts. In practice, we found the latter approach much cheaper to implement with no performance loss over the former.

PSM Class Definition and Shift Usage

The nodes of the clustering tree created in Section 2 can also be used to define the classes of models that share the same shift. During the prior estimation phase, ML shifts are computed at pre-determined intervals at *all* nodes of the tree where there are any observations. During recognition, shifts from the most detailed node with more than T_S (shift threshold) adaptation frames

are used, i.e. the class definition becomes more specific (dynamically at run-time) while maintaining robustness of the estimates as adaptation data increases. All leaves in the subtree below the lowest node that passes the T_S threshold share the same shift. The variable threshold is primarily useful for ML estimation, but may also be used in MAP estimation to discount the effect of the prior.

4. EXPERIMENTS AND RESULTS

Experimental Conditions

A phone is modeled here with two regions each characterized by a linear PSM (i.e. $B = [b_1\ b_2]$) with a single full covariance. Z is a linear time-sampling matrix. Using two linear trajectories per phone can be thought of as approximating a quadratic trajectory. It also allows us to tie (cluster) parameters at a sub-phone level, which has been shown to be more effective than tying at a phone level. Feature vectors include the first 14 mel-warped cepstra with first order differences plus the first difference of log energy, computed every 10 ms. Cepstral mean subtraction is done on a per utterance basis, and feature vectors are normalized for vocal tract length [8] (which gives less of an opportunity to show adaptation gains). Experiments were conducted on the Switchboard corpus, which has telephone-quality conversational speech [9]. Sixty hours of speech (a subset of the training corpus) are used for training gender-dependent models.

To reduce the computational cost of segment-based models, we use the N-best rescoring formalism. Specifically, the top N (=100) word-sequence hypotheses provided by BBN's Byblos system are rescored by the PSM and reranked by linearly combining the PSM log acoustic score with insertion penalties (the number of words and phones in the sentence), a trigram language model score and a duration score (based on relative frequency) to minimize average word error in the top ranking hypotheses [10].

Adaptation Choices

There are many implementation choices for adaptation and we explored the following –

- ML vs. MAP
- adapt full B ($F = [f_1\ f_2]$, offset and slope) vs. part of B ($F = [f_1\ 0]$, offset only)
- transcription vs. batch mode
- in batch mode, supervised vs. unsupervised.

Statistics for shift estimation ($\{e_{ij}\}$ in Equation 3) are collected from the adaptation data (entire conversation for transcription-mode or a fraction of the conversation for batch-mode) based on the hypothesized segment labels and times from recognition with a speaker-independent model. Speaker adapted models, obtained from these statistics, are then used to score the entire conversation in a second pass for transcription-mode or the rest of the conversation for batch-mode. For unsupervised adaptation, the adaptation data are segment labels and times from the top ranking hypothesis, while for supervised adaptation, they are from the "forced-alignment" of the true word sequence.

The implementation cost for adaptation is a small part of the recognizer needs. Memory is needed to accumulate statistics for the shift and priors (for the MAP case) and scaling from half B to full B results in an increase in memory needs of 3-4 times. However the entire memory usage of the PSM recognizer with adaptation are still about the same as that of an unadapted non-parametric mean trajectory segment model system (BU-SSM).

Adaptation Experiments

We present recognition results for a development test set (dev96) comprising 7 conversations (14 speakers and 6381 words, with an average of 2.3 min speech/speaker). Score combination to determine the best hypotheses used weights optimized over the same test set. The baseline recognition system has an average word error rate of 43.1%. In comparison, the BU-SSM system results in 43.0% with three times as many parameters as the baseline PSM system, indicating that the PSM system is parsimonious and viable for LVCSR [6].

Transcription Mode

Unsupervised transcription-mode rescoring with the adapted models results in consistent improvement for all cases with the best case being 0.8% absolute better than the baseline, as shown in Table 1. ML and MAP result in similar performance; adapting full B is better than adapting half B. We find that a non-zero threshold T_S is useful for MAP as well as ML adaptation, probably because a low T_S makes the shifts sensitive to recognition errors.

Batch Mode

The batch mode experiments use the first half of the conversations as adaptation data and the second half for reporting results. The baseline word error rate for

Table 1: *Unsupervised transcription mode results. Baseline (unadapted) WER = 43.1%.*

T_S	avg # of shifts	word error rate %			
		$F = [f_1\ 0]$		$F = [f_1\ f_2]$	
		ML	MAP	ML	MAP
10	2112	42.7	42.7		
25	828	42.4	42.5	42.4	42.4
50	374	42.5	42.6	42.3	42.3
75	223		42.7	42.6	42.6

dev96 is 44.6% and adaptation results are given in Table 2. Again, we see that ML and MAP schemes give similar performance. All supervised cases show an improvement over the baseline, but unsupervised cases do not and require a higher T_S, probably because of the high baseline error rate. A smaller than optimal T_S in supervised batch mode hurts performance because of the inability to generalize to unseen contexts.

Table 2: *Batch mode results. Baseline (unadapted) WER = 44.6%.*

T_S	avg # of shifts	word error rate %			
		$F = [f_1\ 0]$		$F = [f_1\ f_2]$	
		ML	MAP	ML	MAP
Unsupervised batch					
75	84		45.0	44.6	44.8
100	50	44.8	44.8	44.7	44.7
250	4	44.6	44.6		
Supervised batch					
10	1167	44.0	43.8	43.9	43.7
25	423	43.8	43.9	43.5	43.5
50	168	44.1	44.1	44.0	44.1
75	84	44.2	44.3		

5. CONCLUSION

In this work, we have developed a theory for Bayesian and ML adaptation of polynomial trajectory segment models. To be useful in LVCSR systems which have a large number of parameters, adaptation is modeled as a (shared) shift of the mean matrix and the shift detail is determined dynamically at run-time to yield better performance with more adaptation data. Adaptation gains of about 1% absolute on the Switchboard corpus are demonstrated for unsupervised transcription-mode adaptation.

We find adaptation of full B is slightly better than partial B, and adaptation using ML or MAP give roughly the same performance. We conjecture that the high operating word error rate of the Switchboard corpus impacts the choice of the shift threshold T_S (a larger T_S is needed for more robust transformations), as well as limits the usefulness of unsupervised batch adaptation.

6. REFERENCES

[1] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 447–450, 1993.

[2] L. Deng, M. Aksmanovic, X. Sun, and C. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 507–520, Oct. 1994.

[3] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proc. of the Inter. Conf. on Spoken Language Processing*, vol. 1, pp. 466–469, Oct. 1996.

[4] W. Holmes and M. Russell, "Speech recognition using a linear dynamic segmental HMM," in *Proc. European Conference on Speech Comm. and Tech.*, pp. 1611–1614, Sept. 1995.

[5] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 453–455, July 1994.

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

[7] G. Zavaliagkos, R. Schwartz, J. McDonough, and J. Makhoul, "Adaptation algorithms for large scale HMM recognizers," in *Proc. European Conference on Speech Comm. and Tech.*, pp. 1131–1134, 1995.

[8] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, pp. 346–348, 1996.

[9] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. of the Inter. Conf. on Acoust., Speech and Signal Proc.*, vol. 1, pp. 517–520, 1992.

[10] A. Kannan, M. Ostendorf, and J. R. Rohlicek, "Weight estimation for N-best rescoring," in *Proc. DARPA Speech and Natural Language Workshop*, pp. 455–456, Feb. 1992.