Chapter 6

Scalable Inference in Models with Multiple Timescales

■ 6.1 Introduction

In many settings we may wish to learn dynamics at multiple timescales. For example, in the context of speech analysis, we may wish to model both the dynamics within individual phonemes as well as the dynamics across phonemes [68, 18]. In the context of modeling behavior, motion [51], or handwriting [67], it is natural to decompose movements into steps, while still modeling the statistics of the sequence of movements. Each of these modeling tasks involves dynamics at multiple timescales, and therefore it is natural to consider dynamical models that can capture such dynamics while maintaining tractable inference.

In this chapter, we develop a Bayesian nonparametric model and associated inference algorithms applicable to unsupervised learning of such dynamics. We combine and build on ideas developed in previous chapters. In particular, we extend the HDP-HSMM developed in Chapter 3 to include Markovian dynamics within each of its segments. The explicit duration modeling provided by the HDP-HSMM allows us to set duration priors that can disambiguate short-timescale dynamics from long-timescale dynamics and is important for identifiability in the unsupervised setting. Using ideas from Chapters 3 and 4, we develop efficient Gibbs sampling algorithms for our proposed model. Finally, extending ideas from Chapter 5, we also develop a structured Stochastic Variational Inference (SVI) algorithm, which allows inference to scale to large datasets. Developing scalable inference with efficient updates is particularly relevant when fitting rich models, since more data are often required to fit more complex models effectively.

The main contributions of this chapter are algorithmic, particularly in incorporating the algorithmic techniques developed in previous chapters. While the model we propose is new, as we discuss in Section 6.2 many similar models have been explored in the literature. The key advantage to our model is its amenability to the efficient inference algorithms we develop. Our model also benefits from explicit duration modeling and a

Bayesian nonparametric definition, which enable both explicit control over important priors and flexible learning.

In Section 6.2 we highlight some key related work. In Section 6.4 we develop several Gibbs sampling algorithms for the model, including a collapsed direct assignment sampler, a weak limit sampler, and a more efficient weak limit sampler when durations are modeled with negative binomial distributions. In Section 6.5 we develop mean field and SVI updates. Finally, in Section 6.6, we demonstrate our algorithms with an application to unsupervised phoneme discovery.

This chapter synthesizes and extends results from previous chapters and so we rely heavily on their notation and content.

■ 6.2 Related work

The model we define in this chapter is most closely related to generalizations of HMMs and HSMMs known as segment models, which can model sub-dynamics within an HMM or HSMM state. Segment models have a long history in the HMM literature; see Murphy [78] and Murphy [80, Section 17.6] and the references therein. Such models have had considerable success in modeling multiscale dynamics, particular in modeling speech dynamics at the level of words, phones, and sub-phones [80, p. 624]. Such models have typically been explored in non-Bayesian settings. Our model can be viewed as a Bayesian nonparametric segment model, where the Bayesian approach gives us explicit control over duration priors and modeling of uncertainty, and the nonparametric definition provides for flexible learning of model complexity.

A related class of models is the class of Hierarchical HMMs (HHMMs) [28] [80, Section 17.6.2], which have also been studied extensively in non-Bayesian settings. A Bayesian nonparametric HHMM, the infinite HHMM (iHHMM), has been developed and applied successfully to some small example datasets [51]. The model represents an infinite number of dynamical timescales and is extremely flexible. However, it does not provide explicit duration modeling and so it is not easy to use priors to control timescales in the learned dynamics. Furthermore, its structure is not particularly amenable to scalable inference, and in its Gibbs sampling algorithm the hidden states at each level must be sampled conditioned on the hidden states at all the other levels. The model we propose has only two timescales and so is less flexible than the iHHMM, but it allows explicit prior control over duration distributions. In addition, the algorithms we develop exploit more powerful message passing, and the SVI algorithm developed in Section 6.5 allows inference in our model to scale to large datasets.

Finally, the model that is most similar to ours is that of Lee and Glass [68], which develops a Bayesian nonparametric model for unsupervised phoneme discovery. Again, a key difference is again that our model provides explicit duration modeling and allows much more scalable algorithms. In addition, we allow for the substate dynamics themselves to be modeled nonparametrically, while the model of Lee and Glass [68] focuses on modeling each phoneme with fixed-size finite HMMs. While our model can also use fixed-size finite HMMs for short-timescale dynamics, we focus on the fully nonparametric specification.

■ 6.3 Model specification

In this section, we define our generative model, composing both the HDP-HSMM and HDP-HMM generative processes described in Chapter 3, particularly Sections 3.2.3 and 3.3.2. Recall that we write the prior measure on duration parameters as G and the corresponding duration likelihood as $p(d|\vartheta^{(i)})$, and let $\alpha, \gamma > 0$ be concentration parameters. According to the HDP-HSMM generative process, we generate the super-state sequence (z_s), the duration sequence (d_s), and the label sequence (x_t) as

$$\beta \sim \mathrm{GEM}(\gamma) \tag{6.3.1}$$

$$\pi^{(i)} \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha, \beta) \qquad \vartheta^{(i)} \overset{\text{iid}}{\sim} G \qquad i = 1, 2, \ldots \tag{6.3.2}$$

$$z_s \sim \bar{\pi}^{(z_{s-1})} \qquad d_s \sim p(d|\vartheta^{(z_s)}) \qquad s = 1, 2, \ldots \tag{6.3.3}$$

$$x_{t(s):t(s+1)-1} = z_s \qquad t(s) \triangleq \begin{cases} t(s-1) + d_{s-1} & s > 1 \\ 1 & s = 1 \end{cases} \qquad t = 1, 2, \ldots, T, \tag{6.3.4}$$

where, as in Section 3.3.2, $\bar{\pi}^{(i)} \triangleq \frac{\pi_j^{(i)}}{1-\pi_i^{(i)}}(1 - \delta_{ij})$. While the HDP-HSMM generates the observation sequence (y_t) within a segment as conditionally independent draws from an observation distribution, here we instead generate observations for each segment according to an HDP-HMM. That is, for each HDP-HSMM state $i = 1, 2, \ldots$ we have an HDP-HMM with parameters $\{\beta^{(i)}, \pi^{(i,j)}, \theta^{(i,j)}\}_{j=1}^{\infty}$ generated according to

$$\beta^{(i)} \sim \mathrm{GEM}(\gamma^{(i)}) \tag{6.3.5}$$

$$\pi^{(i,j)} \overset{\text{iid}}{\sim} \mathrm{DP}(\alpha^{(i)}, \beta^{(i)}) \qquad \theta^{(i,j)} \overset{\text{iid}}{\sim} H \qquad j = 1, 2, \ldots \tag{6.3.6}$$

where H is the prior measure over observation parameters $\theta^{(i,j)}$, $\alpha^{(i)}$ and $\gamma^{(i)}$ are concentration parameters, and each $\pi^{(i,j)}$ is a transition distribution out of the corresponding HDP-HMM's jth state. Then for a segment s in with HDP-HSMM super-state z_s and duration d_s, we generate observations from the corresponding HDP-HMM via

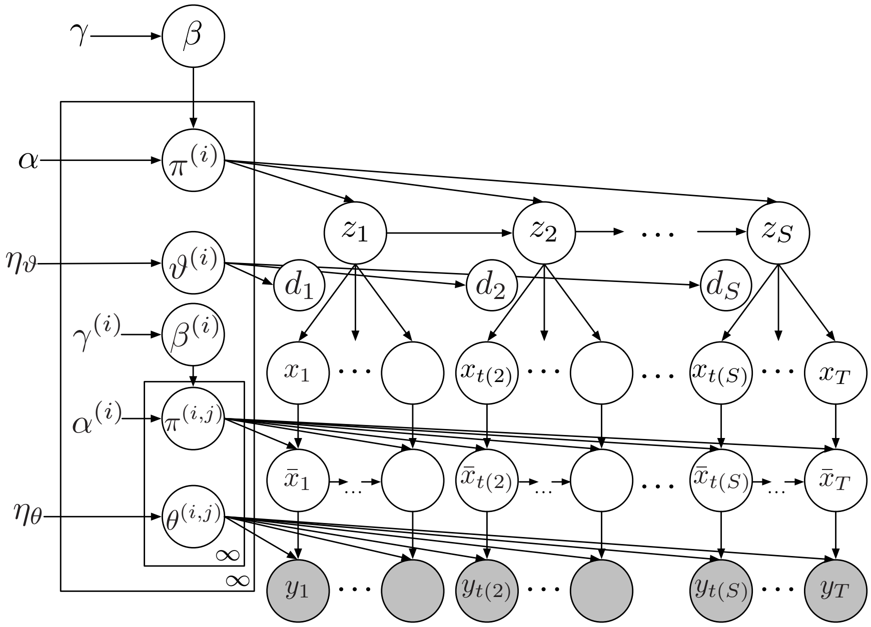


Figure 6.1: A graphical model for the HDP-HSMM with sub-HDP-HMM observations. Note that it is not formally a graphical model because the number of nodes is random due to the random durations.

$$\bar{x}_t \sim \pi^{(z_s,\bar{x}_{t-1})} \qquad y_t \sim p(y|\theta^{(z_s,\bar{x}_t)}) \qquad t = t(s), t(s)+1, \ldots, t(s+1)-1, \qquad (6.3.7)$$

where $\theta^{(z_s,\bar{x}_t)}$ is the observation parameter from the corresponding HDP-HMM's jth state, and $p(y|\theta^{(z_s,\bar{x}_t)})$ is the corresponding observation likelihood. We call $(\bar{x}_t)_{t=1}^T$ the *substate sequence*, and emphasize that it is distinct from the HDP-HSMM's super-state sequence (z_s) and label sequence (x_t). See Figure 6.1 for a graphical model.

This model definition combines the explicit duration modeling and nonparametric flexibility of the HDP-HSMM of Chapter 3 with HDP-HMM dynamics within each HSMM segment. The HDP-HSMM states can model longer-timescale dynamics, such as the dynamics between phonemes, while the HDP-HMM states can model shorter-timescale dynamics, such as the structure within an individual phoneme. As we show in the following sections, this model definition is also amenable to efficient inference.

While we have defined this model using Bayesian nonparametric priors for both layers of dynamics, it is straightforward to adapt the definition so that one or both of the layers is finite. For example, it may be desirable to use finite structured sub-HMM models to exploit some domain knowledge [68]. Alternatively, it is also possible to make the coarser-scale dynamical process a finite HSMM while allowing the substate dynamics to be generated from an HDP-HMM, thus enabling model selection via a semiparametric approach [38].

6.4 Gibbs sampling

In this section, we develop several Gibbs sampling algorithms for the model defined in Section 6.3. First, we develop a collapsed direct assignment sampler. This sampler avoids approximating the posterior with a finite distribution but, as with the direct assignment sampler developed in Chapter 3 and simulated in Figure 3.11(b), its mixing rate is far too slow to be practically useful. We include it for completeness and theoretical interest.

Second, we develop a sampler based on the weak limit approximation. Analogous to the weak limit sampler developed in Chapter 3, this sampler can use message passing to perform block sampling and therefore achieve much greater mixing.

Finally, we build on the results of Chapter 4 to develop a much faster weak limit Gibbs sampler for negative binomial duration distributions. The message passing complexity is greatly reduced, and in particular is only linear in the sequence length T.

6.4.1 Collapsed Gibbs sampler

To develop a collapsed Gibbs sampler, we extend the HDP-HSMM direct assignment Gibbs algorithm developed in Section 3.4.3. Essentially, we combine the HDP-HSMM direct assignment sampler and the HDP-HMM direct assignment sampler.

The algorithm state for our direct assignment sampler consists of a finite prefix of the HDP-HSMM β parameter and finite prefixes of each of the sub-HDP-HMM $\beta^{(i)}$ parameters. It also includes both the HDP-HSMM label sequence (x_t) and the substate sequence $(\bar{x}_t)$. That is, we write the sampler state as $(\beta_{1:N}, \beta^{(i)}_{1:N^{(i)}}, x_{1:T}, \bar{x}_{1:T})$, where we use N to represent the number of used HDP-HSMM states and $N^{(i)}$ to represent the number of used states in the ith HDP-HMM. The other parameters are integrated out analytically, including the HDP-HSMM transition parameters $\{\pi^{(i)}\}$, the sub-HDP-HMM transition parameters $\{\pi^{(i,j)}\}$, the duration parameters $\{\vartheta^{(i)}\}$, and the observation parameters $\{\theta^{(i,j)}\}$.

The algorithm proceeds by jointly resampling each pair $(x_t, \bar{x}_t)$ for $t = 1, 2, \ldots, T$ and by resampling the $\beta_{1:N}$ and $\beta^{(i)}_{1:N^{(i)}}$.

Resampling the label and substate sequence. To resample each $(x_t, \bar{x}_t)$ conditioned on all the other variables and parameters, we extend the HDP-HSMM sampling step of Section 3.4.3 That is, we resample $(x_t, \bar{x}_t)$ by considering all possible assignments (k, k') for $k = 1, 2, \ldots, N + 1$ and $k' = 1, 2, \ldots, N^{(i)} + 1$ and evaluating up to proportionality the conditional probability

$$p((x_t,\bar{x}_t)=(k,k')|(x_{\setminus t}),(\bar{x}_{\setminus t}),\beta,\{\beta^{(i)}\}), \tag{6.4.1}$$

where we suppress notation for conditioning on all the hyperparameters. Recall that as we vary the assignment of the HDP-HSMM label x_t we must consider the possible merges into adjacent label segments, and as a result there are between 1 and 3 terms in the expression for (6.4.1), each of the form

$$p((x_t,\bar{x}_t)=(k,k')|(x_{\setminus t},\bar{x}_{\setminus t}))\propto \underbrace{\frac{\alpha\beta_k+n_{x_{\text{prev}},k}}{\alpha(1-\beta_{x_{\text{prev}}})+n_{x_{\text{prev}},\cdot}}}_{\text{left-transition}}\cdot\underbrace{\frac{\alpha\beta_{x_{\text{next}}}+n_{k,x_{\text{next}}}}{\alpha(1-\beta_k)+n_{k,\cdot}}}_{\text{right-transition}}$$
$$\cdot\underbrace{f_{\text{dur}}(t_2-t_1+1)}_{\text{duration}}\cdot\underbrace{f_{\text{obs}}(y_{t_1:t_2}|k)}_{\text{observation}}, \tag{6.4.2}$$

where we have used t_1 and t_2 to denote the first and last indices of the segment, respectively, and $(x_{\setminus t})$ and $(\bar{x}_{\setminus t})$ to denote the label sequence and substate sequence assignments excluding the tth index, respectively. See Section 3.4.3 for details. In the case of the HDP-HSMM of Chapter 3, the term $f_{\text{obs}}(y_{t_1:t_2}|k)$ is computed from the independent observation model. For sub-HDP-HMM observations, we simply replace this term with the appropriate score for HDP-HMM observations, incorporating not only the data $y_{t_1:t_2}$ but also the substate assignment $\bar{x}_t = k'$ and the substate sequence $(\bar{x}_{\setminus t})$.

Using the formula for the probability of an assignment sequence under an HDP model [43, 106], we can write the $f_{\text{obs}}(y_{t_1:t_2}|k,k')$ term for sub-HDP-HMM observations. Let n_{ij} be the number of transitions from substate i to substate j in the kth sub-HMM, and let $n_{i\cdot}=\sum_{j=1}^{N^{(k)}} n_{ij}$. In addition, for first and last segment indices t_1 and t_2, let

$$\bar{n}_{ij}=n_{ij}-\#\{t:\bar{x}_t=i,\bar{x}_{t+1}=j,t_1\leq t<t_2,x_t=k\} \tag{6.4.3}$$

be the number of transitions from substate i to substate j in the kth sub-HMM excluding those in the segment from t_1 to t_2, and let $\bar{n}_{i\cdot}=\sum_{j=1}^{N^{(k)}}\bar{n}_{ij}$. Then we can write

$$f_{\text{obs}}(y_{t_1:t_2}|k,k') = \left(\prod_{i=1}^{N^{(k)}} \frac{\Gamma(\alpha + \bar{n}_{i\cdot})}{\Gamma(\alpha + n_{i\cdot})} \prod_{j=1}^{N^{(k)}} \prod_{\ell=\bar{n}_{ij}}^{n_{ij}-1} (\alpha\beta_j^{(k)} + \ell) \right) \cdot \left(\int \prod_{t=t_1}^{t_2} p(y_t|\theta^{(k,k')}) p(\theta^{(k,k')}|\{y_{t'} : \bar{x}_{t'} = k', x_{t'} = k\}, \eta_\theta)\, d\theta^{(k,k')} \right) \tag{6.4.4}$$

where η_θ is the corresponding observation hyperparameter. By substituting (6.4.4) for f_{obs} in (6.4.2), we can proceed with the HDP-HSMM sampling procedure of Section 3.4.3.

Resampling $\beta_{1:N}$ and $\beta^{(i)}_{1:N^{(i)}}$. To resample $\beta_{1:N}$ conditioned on the HDP-HSMM label sequence (x_t), we use the same auxiliary variable method developed in Section 3.4.2. To resample each $\beta^{(i)}_{1:N^{(i)}}$ for each sub-HDP-HMM, $i = 1, 2, \ldots, N$, we use the standard HDP-HMM direct assignment update [106].

■ 6.4.2 Weak limit sampler

We can develop a more efficient sampling algorithm by using a weak limit approximation and exploiting dynamic programming. In particular, we develop a weak limit sampler that block resamples the label sequence and substate sequence jointly. We build on the weak limit sampler developed in Section 3.4.2. We write the weak limit truncation level of the HDP-HSMM as N and the weak limit truncation level of the ith HDP-HMM as $N^{(i)}$.

Recall that the label sequence (x_t) can be resampled by first passing HSMM messages backward and then sampling forward, as in Section 3.4.1, particularly equations (3.4.3) and (3.4.6). From Section 3.2.2, the HSMM messages (B, B^*) are defined by

$$B_{t,i} = \sum_{j=1}^{N} B^*_{t,j} p(x_{t+1} = j | x_t = i, x_{t+1} \neq x_t) = \sum_{j=1}^{N} B^*_{t,j} A_{ij}, \tag{6.4.5}$$

$$B^*_{t,i} = \sum_{d=1}^{T-t} B_{t+d,i}\, \underbrace{p(d_{s(t)} = d | z_{s(t+1)} = i)}_{\text{duration prior term}} \cdot \underbrace{p(y_{t+1:t+d} | z_{s(t+1)} = i, d_{s(t+1)} = d)}_{\text{likelihood term}} \tag{6.4.6}$$

$$= \sum_{d=1}^{T-t-1} B_{t+d,i} D_{d,i} p(y_{t+1:t+d} | z_{s(t+1)} = i, d_{s(t+1)} = d) \tag{6.4.7}$$

$$B_{T,i} \triangleq 1, \tag{6.4.8}$$

where

$$D_{d,i} = p(d|\vartheta^{(i)}) \qquad A_{ij} = p(x_{t+1} = j | x_t = i, x_{t+1} \neq x_t) \tag{6.4.9}$$

and where $s(t)$ denotes the segment index for time index t. In Chapter 3, the likelihood term $p(y_{t+1:t+d}|z_{s(t+1)} = i, d_{s(t+1)} = d)$ is computed as a product of independent likelihoods, while here we must compute it according to the HMM observation model. If we can compute these segment likelihood terms efficiently, we can use them in Eqs. (6.4.5)-(6.4.8) to compute the backward messages over the HSMM and sample the HSMM label sequence as in Section 3.4.1. Given the HSMM label sequence, we can then sample the substate sequence using the HMM state sequence sampling algorithm given in Section 2.4.1.

Therefore it remains only to compute the segment likelihood terms efficiently. We can exploit the Markov structure in the substate sequence to write these likelihoods in terms of another set of messages. For each $i = 1, 2, \ldots, N$, we define the sub-HMM backward messages for each $t = 1, 2, \ldots, T$ as

$$B^{(i,t)}_{t',j} = \sum_{k=1}^{N^{(i)}} A^{(i)}_{jk} L^{(i)}_{t'+1,k} B^{(i,t)}_{t'+1,k} \qquad t' = 1, 2, \ldots, t \qquad B^{(i,t)}_{t,j} = 1, \tag{6.4.10}$$

where $L^{(i)}_{t,k} = p(y_t|\theta^{(i,k)})$ is the observation likelihood for the jth substate of the ith HMM and $A^{(i)}_{jk} = p(\bar{x}_{t+1} = k | \bar{x}_t = j, x_t = i)$ is the probability of transitioning from the jth to the kth substate in the ith HMM. Similarly, we define the sub-HMM forward messages for each $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$ as

$$F^{(i,t)}_{t',j} = \sum_{k=1}^{N^{(i)}} A_{kj} L_{t',j} F^{(i,t)}_{t'-1,k} \qquad t = t+1, t+2, \ldots, T \qquad F^{(i,t)}_{t,k} = \pi^{(i,0)}_k \tag{6.4.11}$$

where $\pi^{(i,0)}$ is the initial state distribution for the ith sub-HMM. For any fixed time index t and superstate index i, we can compute these messages in time $\mathcal{O}(T{N^{(i)}}^2)$ time, and therefore we can compute all such sub-HMM messages in time $\mathcal{O}(NT^2{N^{(i)}}^2)$ time. Finally, we can use these messages to compute every segment likelihood term via

$$p(y_{t:t+d-1}|z_{s(t)} = i, d_{s(t)} = d) = \sum_{j=1}^{N^{(i)}} F^{(i,t)}_{t',j} B^{(i,t+d)}_{t',j}, \tag{6.4.12}$$

for any $t' = t, t+1, \ldots, t+d-1$. To compute only the backward HSMM messages, as required for the block sampling procedure, it suffices to compute only the sub-HMM forward messages.

Composing these expressions, we can write the overall HSMM messages as

$$B^*_{t,i} = \sum_{d=1}^{T-1-t} B_{t+d,i} D_{d,i} \left(\sum_{\ell=1}^{N^{(i)}} F^{(i,t)}_{t+d,\ell} \right) \qquad B_{t,i} = \sum_{j=1}^{N} A_{ij} B^*_{t,j}. \tag{6.4.13}$$

Writing $N_{\text{sub}} = \max_i N^{(i)}$, these messages require $\mathcal{O}(T^2 N N_{\text{sub}}^2 + T N^2)$ time to compute. With these messages, we can block resample the HSMM label sequence and substate sequence.

The sampling updates to the other model parameters are identical to those described in Sections 3.4.1 and 2.4.1.

■ 6.4.3 Exploiting negative binomial durations

While the weak limit sampling procedure developed in Section 6.4.2 is general, it can be computationally expensive for long observation sequences. In this section we apply and extend the ideas developed in Chapter 4 to write an update for negative binomial duration models for which the computational complexity scales only linearly in T and is generally much more efficient. As in Chapter 4, this algorithm generalizes immediately to models in which the duration distributions are mixtures of negative binomial distributions.

Recall from Chapter 4 that with negative binomial durations we can compute the HSMM messages with more efficient recursions because the duration can be represented as an augmentation with a small number of Markov states. In particular, to represent an HSMM with negative binomial parameters $(r^{(i)}, p^{(i)})$ for $i = 1, 2, \ldots, N$, we constructed an equivalent HMM on $\sum_{i=1}^{N} r^{(i)}$ states. We can similarly embed an HSMM with HMM emissions and negative binomial durations in a stationary HMM on $\sum_{i=1}^{N} r^{(i)} N^{(i)}$ states. Using the notation of Chapter 4, we choose

$$\bar{A}^{(i)} = \hat{A}^{(i)} \otimes A^{(i)}, \qquad \bar{b}^{(i)} = \hat{b}^{(i)} \otimes \mathbb{1}^{(i)}, \qquad \bar{c}^{(i)} = \hat{c}^{(i)} \otimes \pi^{(i)}_{\text{sub}}, \tag{6.4.14}$$

where $X \otimes Y$ denotes the Kronecker product of matrices X and Y, $\mathbb{1}^{(i)}$ denotes the all-ones vector of size $r^{(i)}$, and $(\hat{A}^{(i)}, \hat{b}^{(i)}, \hat{c}^{(i)})$ denotes the HMM embedding parameters for negative binomial durations given in Eqs. (4.4.7)-(4.4.8). Note that we can index into the matrix $\bar{A}$ using the tuple (i, j, k), where $i = 1, 2, \ldots, N$ indexes the HSMM state, $j = 1, 2, \ldots, r^{(i)}$, indexes the duration pseudostate, and $k = 1, 2, \ldots, N^{(i)}$ indexes the sub-HMM substate. By comparing this construction to that of the embedding developed in Section 4.4, it is clear that it encodes the same dynamics on the HSMM label sequence: if we simply sum over the sub-HMM substates, we recover the same embedding as that of Section 4.4. That is, if we use $\hat{A}$ to denote the transition matrix of the HMM embedding of Section 4.4, then

$$\hat{A}_{(i,j),(i',j')} = \sum_{k=1}^{N^{(i)}} \sum_{k'=1}^{N^{(i')}} \bar{A}_{(i,j,k),(i',j',k')}. \tag{6.4.15}$$

Furthermore, the sub-HMM substate dynamics are faithfully represented: the last block row of $\bar{A}^{(i,j)}$ ensures that the first substate of a segment is sampled according to $\pi^{(j)}$, and the substate transition probabilities are those of $A^{(i)}$ until the superstate changes.

Note that due to the structure of the matrix $\bar{A}$, the matrix-vector multiplications required to perform message passing are especially efficient to compute. In particular, using the identity

$$(X \otimes Y)\mathrm{vec}(Z) = \mathrm{vec}(XZY^{\mathsf{T}}) \tag{6.4.16}$$

for any matrices X, Y, and Z of appropriate dimensions, we can compute the block diagonal part of a matrix-vector product in $\mathcal{O}(N(R + N_{\text{sub}}^2))$, where $R = \max_i r^{(i)}$. Furthermore, using the structure in each $\bar{A}^{(i,j)}$, we can compute the off-block-diagonal part of a matrix-vector product in $\mathcal{O}(N^2 + NN_{\text{sub}})$. Therefore, using the methods developed in Chapter 4, we can use the embedding to compute the HSMM messages in only $\mathcal{O}(TN(R + N_{\text{sub}}^2) + TN^2)$ time, avoiding the quadratic dependence on T. Finally, note that, using the methods developed in Chapter 4, this HSMM messages computation requires only $\mathcal{O}(TN + NRN_{\text{sub}})$ memory, a significant savings compared to the $\mathcal{O}(TNRN_{\text{sub}})$ memory required to compute the HMM messages in the full HMM embedding.

Given the HSMM messages, we can perform the block sampling update to the label sequence and substate sequence described in Section 6.4.2 much more efficiently.

■ 6.5 Mean field and SVI

In this section, we derive the key updates necessary to perform mean field or SVI inference in the model. This section relies heavily on the notation used in Chapter 5, and extends its results to the model developed in this chapter.

Following the notation of Chapter 5, we write our variational family as

$$q(\beta)q(\pi^{(0)}) \prod_{i=1}^{N} q(\pi^{(i)})q(\vartheta^{(i)}) \left(q(\beta^{(i)})q(\pi^{(i,0)}) \prod_{j=1}^{N^{(i)}} q(\pi^{(i,j)})q(\theta^{(i,j)}) \right) q(x_{1:T}, \bar{x}_{1:T}) \tag{6.5.1}$$

where N is the truncation parameter for the HDP-HSMM and each $N^{(i)}$ is the truncation parameter for the ith sub-HDP-HMM. The variational factors are defined analogously to those used in Chapter 5:

$$q(\theta^{(i,j)}) \propto \exp\left\{\langle \widetilde{\eta}^{(i,j)}_{\theta}, t^{(i,j)}_{\theta}(\theta^{(i,j)})\rangle\right\} \qquad q(\pi^{(i,j)}) = \mathrm{Dir}(\widetilde{\alpha}^{(i,j)}) \tag{6.5.2}$$

$$q(\pi^{(i)}) = \mathrm{Dir}(\widetilde{\alpha}^{(i)}) \qquad q(\vartheta^{(i)}) \propto \exp\left\{\langle \widetilde{\eta}^{(i)}_{\vartheta}, t^{(i)}_{\vartheta}(\vartheta^{(i)})\rangle\right\} \tag{6.5.3}$$

$$q(\beta) = \delta_{\beta^*}(\beta) \qquad q(\beta^{(i)}) = \delta_{\beta^{*(i)}}(\beta^{(i)}). \tag{6.5.4}$$

The corresponding prior densities on each term are

$$p(\theta^{(i,j)}) \propto \exp\left\{\langle \eta^{(i)}_{\theta}, t^{(i)}_{\theta}(\theta^{(i,j)})\rangle\right\} \qquad p(\pi^{(i,j)}|\beta_{1:N^{(i)}}) = \mathrm{Dir}(\alpha^{(i)}\beta_{1:N^{(i)}}) \tag{6.5.5}$$

$$p(\pi^{(i)}|\beta_{1:N}) = \mathrm{Dir}(\alpha\beta_{1:N}) \qquad p(\vartheta^{(i)}) \propto \exp\left\{\langle \eta^{(i)}_{\vartheta}, t^{(i)}_{\vartheta}(\vartheta^{(i)})\rangle\right\}. \tag{6.5.6}$$

We derive a mean field update to the variational factors over model parameters in two steps: first, we define structured mean field message-passing recursions analogous to those defined in Section 6.4.2; second, we show how to use the mean field messages to compute the expected statistics necessary for the parameter updates.

As in Section 6.4.2, it is useful to define sub-HMM messages for each $i = 1, 2, \ldots, N$ and each time index $t = 1, 2, \ldots, T$:

$$\widetilde{B}^{(i,t)}_{t',j} = \sum_{k=1}^{N^{(i)}} \widetilde{A}^{(i)}_{jk} \widetilde{L}^{(i)}_{t'+1,k} \widetilde{B}^{(i,t)}_{t'+1,k} \qquad t' = 1, 2, \ldots, t \qquad \widetilde{B}^{(i,t)}_{t,j} = 1 \tag{6.5.7}$$

$$\widetilde{F}^{(i,t)}_{t',j} = \sum_{k=1}^{N^{(i)}} \widetilde{A}^{(i)}_{kj} \widetilde{L}^{(i)}_{t',j} \widetilde{F}^{(i,t)}_{t'-1,k} \qquad t = t+1, t+2, \ldots, T \qquad \widetilde{F}^{(i,t)}_{t,k} = \widetilde{\pi}^{(i,0)}_{k} \tag{6.5.8}$$

where

$$\widetilde{L}^{(i)}_{t,j} = \mathbb{E}_{q(\theta^{(i,j)})}\left[\ln p(y_t|\theta^{(i,j)}\right] \qquad \widetilde{\pi}^{(i,j)} = \mathbb{E}_{q(\pi)}\left[\ln \pi^{(i,j)}\right] \tag{6.5.9}$$

and where $\widetilde{A}^{(i)}$ is a matrix with its kth row as $\widetilde{\pi}^{(i,k)}$. Then we can write the overall message recursions as

$$\widetilde{B}^*_{t,i} = \sum_{d=1}^{T-1-t} \widetilde{B}_{t+d,i}\widetilde{D}_{d,i}\left(\sum_{\ell=1}^{N^{(i)}} \widetilde{F}^{(i,t)}_{t+d,\ell}\right) \qquad \widetilde{B}_{t,i} = \sum_{j=1}^{N} \widetilde{A}_{ij}\widetilde{B}^*_{t,j} \tag{6.5.10}$$

$$\widetilde{F}_{t,i} = \sum_{d=1}^{T-t-1} \widetilde{F}^*_{t-d,i}\widetilde{D}_{d,i}\left(\sum_{\ell=1}^{N^{(i)}} \widetilde{B}^{(i,t)}_{t-d,\ell}\widetilde{\pi}^{(i,0)}_{\ell}\right) \qquad \widetilde{F}^*_{t,i} = \sum_{j=1}^{N} \widetilde{A}_{ji}\widetilde{F}_{t,j} \tag{6.5.11}$$

where

$$\widetilde{D}_{d,i} = \mathbb{E}_{q(\vartheta^{(i)})}\left[\ln p(d|\vartheta^{(i)})\right] \qquad Z = \sum_{i=1}^{N} F_{T,i} \tag{6.5.12}$$

and where $\widetilde{A}$ is a matrix with its ith row as $\widetilde{\pi}^{(i)}$. As in Section 6.4.2, these messages can be computed in time $\mathcal{O}(T^2 N N_{\text{sub}}^2 + TN^2)$.

Next, we calculate expected statistics in terms of these messages. To simplify notation, we write the event $\{x_{t:t+d-1} = i, x_{t-1} \neq x_t \neq x_{t+d}\}$ simply as $\{x_{t:t+d-1} = i\}$. First, note that we can write

$$E_{q(x_{1:T},\bar{x}_{1:T})}\left[\mathbb{I}[x_{t:t+d-1} = i]\right] = \left(\widetilde{F}^*_{t,i}\widetilde{B}_{t+d-1,i}\widetilde{D}_{d,i}\sum_{\ell=1}^{N^{(i)}} \widetilde{F}^{(i,t)}_{t,\ell}\widetilde{B}^{(i,t+d)}_{t,\ell}\right)/Z. \tag{6.5.13}$$

This decomposition follows from the definitions of the HSMM messages. Using the definition of the sub-HMM messages, we can similarly write

$$\mathbb{E}_{q(x_{1:T},\bar{x}_{1:T})}\left[\mathbb{I}[\bar{x}_{t'} = j, \bar{x}_{t'+1} = k \big| x_{t:t+d-1} = i\right] = \frac{\left(\widetilde{F}^{(i,t)}_{t',j}\widetilde{B}^{(i,t+d)}_{t'+1,k}\widetilde{L}^{(i)}_{t'+1,k}\widetilde{A}^{(i)}_{j,k}\right)}{\sum_{\ell=1}^{N^{(i)}}\widetilde{F}^{(i,t)}_{t,\ell}\widetilde{B}^{(i,t+d)}_{t,\ell}} \tag{6.5.14}$$

for any $t' = t, t+1, \ldots, t+d-2$. To compute the expected statistics, we compose these two expressions and use the basic identity that for any random variable X we have

$$\begin{aligned}
\mathbb{E}\big[\mathbb{I}[X \in A]\,\mathbb{I}[X \in B]\big] &= \mathbb{P}\big[X \in A, X, \in B\big] && (6.5.15)\\
&= \mathbb{P}\big[X \in A\big]\,\mathbb{P}\big[X \in B | X \in A\big] && (6.5.16)\\
&= \mathbb{E}\big[\mathbb{I}[X \in A]\big]\;\mathbb{E}\big[\mathbb{I}[X \in B]\big| X \in A\big]. && (6.5.17)
\end{aligned}$$

Therefore we can compute the expected statistics for each sub-HDP-HMM factor as

$$\begin{aligned}
\hat{t}_y^{(i,j)} &\triangleq \mathbb{E}_{q(x_{1:T},\bar{x}_{1:T})}\left[\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\mathbb{I}\left[x_{t:t+d-1}=i\right]\sum_{t'=t}^{t+d-1}\mathbb{I}\left[\bar{x}_{t'}=j\right](t_y^{(i,j)},1)\right] \\
&=\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\left(\widetilde{F}^{*}_{t,i}\widetilde{B}_{t+d-1,i}\widetilde{D}_{d,i}\right)\left(\sum_{t'=t}^{t+d-1}\widetilde{F}^{(i,t)}_{t',j}\widetilde{B}^{(i,t+d)}_{t',j}\widetilde{L}^{(i)}_{t',j}\right)/Z \qquad (6.5.18) \\
(\hat{t}^{(i,j)}_{\text{subtr}})_k &\triangleq \mathbb{E}_{q(x_{1:T},\bar{x}_{1:T})}\left[\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\mathbb{I}\left[x_{t:t+d-1}=i\right]\sum_{t'=t}^{t+d-2}\mathbb{I}\left[\bar{x}_{t'}=j,\bar{x}_{t'+1}=k\right]\right] \\
&=\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\left(\widetilde{F}^{*}_{t,i}\widetilde{B}_{t+d-1,i}\widetilde{D}_{d,i}\right)\left(\sum_{t'=t}^{t+d-2}\widetilde{F}^{(i,t)}_{t',j}\widetilde{B}^{(i,t+d)}_{t'+1,k}\widetilde{L}^{(i)}_{t'+1,k}\widetilde{A}^{(i)}_{j,k}\right)/Z \qquad (6.5.19) \\
(\hat{t}^{(i)}_{\text{subinit}})_j &\triangleq \mathbb{E}_{q(x_{1:T},\bar{x}_{1:T})}\left[\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\mathbb{I}\left[x_{t:t+d-1}=i\right]\mathbb{I}\left[\bar{x}_{t}=j\right]\right] \\
&=\sum_{t=1}^{T-1}\sum_{d=1}^{T-1-t}\left(\widetilde{F}^{*}_{t,i}\widetilde{B}_{t+d-1,i}\widetilde{D}_{d,i}\right)\left(\widetilde{F}^{(i,t)}_{t,j}\widetilde{B}^{(i,t+d)}_{t,j}\right)/Z \qquad (6.5.20)
\end{aligned}$$

Furthermore, we can compute the expected statistics for each HDP-HSMM factor as

$$\begin{aligned}
(\hat{t}^{(i)}_{\text{dur}})_d &\triangleq \mathbb{E}_{q(x_{1:T})}\left[\sum_{t=1}^{T-1}\mathbb{I}\left[x_{t:t+d}=i\right]\right] \\
&=\sum_{t=1}^{T-1}\widetilde{F}^{*}_{t,i}\widetilde{B}_{t+d,i}\widetilde{D}_{d,i}\left(\sum_{\ell=1}^{N^{(i)}}\widetilde{F}^{(i,t)}_{t,\ell}\widetilde{B}^{(i,t+d)}_{t,\ell}\right)/Z \qquad (6.5.21) \\
(\hat{t}^{(i)}_{\text{trans}})_j &\triangleq \mathbb{E}_{q(x_{1:T})}\left[\sum_{t=1}^{T-1}\mathbb{I}\left[x_t=i,x_{t+1}=j\right]\right]=\sum_{t=1}^{T-1}\widetilde{F}_{t,i}\widetilde{B}^{*}_{t,j}\widetilde{A}_{i,j}/Z \qquad (6.5.22) \\
(\hat{t}_{\text{init}})_i &\triangleq \mathbb{E}_{q(x_{1:T})}\mathbb{I}[x_1=i]=\widetilde{F}^{*}_{1,i}\widetilde{B}_{1,i}/Z \qquad (6.5.23)
\end{aligned}$$

While these expected statistics expressions appear complex when fully expanded, the expressions are in fact quite modular: each involves the expectation of an HSMM segment indicator, which is computed using the HSMM messages, and possibly an expectation in terms of a sub-HMM statistic, which is computed using the sub-HMM messages.

Using the notation of Chapter 5, the corresponding SVI updates to the variational factors on the model parameters are then

Algorithm 6.1 Sub-HMM SVI

```
Initialize global variational parameters $\widetilde{\eta}_\theta^{(i,j)}$, $\widetilde{\alpha}^{(i,j)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\eta}_\vartheta^{(i)}$
for $t = 1, 2, \ldots$ do
    Sample minibatch index $\hat{k}$ uniformly from $\{1, 2, \ldots, K\}$
    Using minibatch $\bar{y}^{(\hat{k})}$, compute sub-HMM messages using (6.5.7)-(6.5.8)
        and HSMM messages using (6.5.10)-(6.5.11)
    Using the messages, compute $\hat{t}_y^{(i,j)}$, $\hat{t}_{\text{subr}}^{(i,j)}$, $\hat{t}_{\text{subinit}}^{(i)}$, $\hat{t}_{\text{dur}}^{(i)}$, $\hat{t}_{\text{trans}}^{(i)}$, and $\hat{t}_{\text{init}}$
        using (6.5.18)-(6.5.23).
    Update each $\widetilde{\eta}_\theta^{(i,j)}$, $\widetilde{\alpha}^{(i,j)}$, $\widetilde{\alpha}^{(i)}$, and $\widetilde{\eta}_\vartheta^{(i)}$ using (6.5.24)-(6.5.29)
```

$$\widetilde{\eta}_\theta^{(i,j)} \leftarrow (1-\rho)\widetilde{\eta}_\theta^{(i,j)} + \rho(\eta_\theta^{(i,j)} + s \cdot \hat{t}_y^{(i,j)}) \tag{6.5.24}$$

$$\widetilde{\alpha}^{(i,j)} \leftarrow (1-\rho)\widetilde{\alpha}^{(i,j)} + \rho(\alpha^{(i)} + s \cdot \hat{t}_{\text{subtr}}^{(i,j)}) \tag{6.5.25}$$

$$\widetilde{\alpha}^{(i,0)} \leftarrow (1-\rho)\widetilde{\alpha}^{(i,0)} + \rho(\alpha^{(i)} + s \cdot \hat{t}_{\text{subinit}}^{(i)}) \tag{6.5.26}$$

$$\widetilde{\alpha}^{(i)} \leftarrow (1-\rho)\widetilde{\alpha}^{(i)} + \rho(\alpha + s \cdot \hat{t}_{\text{trans}}^{(i)}) \tag{6.5.27}$$

$$\widetilde{\alpha}^{(0)} \leftarrow (1-\rho)\widetilde{\alpha}^{(0)} + \rho(\alpha + s \cdot \hat{t}_{\text{init}}^{(i)}) \tag{6.5.28}$$

$$\widetilde{\eta}_\vartheta^{(i)} \leftarrow (1-\rho)\widetilde{\eta}_\vartheta^{(i)} + \rho(\eta_\vartheta^{(i)} + s(\textstyle\sum_{d=1}^{T}(\hat{t}_{\text{dur}}^{(i)})_d \cdot (t_d(d), 1))). \tag{6.5.29}$$

for some stepsize ρ and minibatch scaling s as in Section 5.1.2. We summarize the overall algorithm in Algorithm 6.1.

6.6 Experiments

As discussed in Section 6.1, one natural motivation for models with multiple timescales is speech analysis. Individual phonetic units, such as phonemes, have internal dynamical structure that can be modeled by an HMM with Gaussian emissions [58, 68]. At the same time, it is desirable to model the dynamical patterns among the phonemes themselves. The model and inference algorithms developed in this chapter can be applied to capture these two timescales of dynamics. Furthermore, by utilizing both explicit duration modeling and a Bayesian approach we can easily incorporate informative prior knowledge and encourage the model to learn interpretable representations.

Similar models have been applied successfully to tasks in speech analysis. In particular, Lee and Glass [68] develop a Bayesian nonparametric approach in which phonetic units are each modeled as fixed-size HMMs and the number of such units is discovered using a Dirichlet process mixture. The authors show that the discovered phonetic units are highly correlated with English phones and that the model can be used for some speech tasks to achieve state-of-the-art performance relative to other unsupervised methods. Our model can be viewed as a refinement of this approach in two ways: first, our model admits explicit phonetic unit duration and transition modeling, and second, the inference algorithms we develop allow our model and similar models to be fit to large datasets much more efficiently. Indeed, our algorithms allow such models to be fit in minutes or hours of computation time instead of days or weeks.

In this section we describe an application of our models and algorithms to semi-supervised phonetic unit modeling based on the approach of Lee and Glass [68]. In particular, we demonstrate the advantages of using explicit duration priors, of modeling dynamics within phonetic units, and of using scalable inference algorithms.

The remainder of this section is organized as follows. In Section 6.6.1 we describe the dataset and features we use to train and evaluate the model. In Section 6.6.2 we describe a general approach to set the hyperparameters for informative duration priors. Finally, in Section 6.6.3 we describe our training procedure and experimental results.

■ 6.6.1 Dataset and features

Our setup follows Lee and Glass [68] closely. We use the TIMIT dataset, which consists of recordings of 630 speakers each reading 10 sentences, for a total of 6300 example sequences. We process these recordings into 13-dimensional MFCC features [21] using sliding windows of width 25ms spaced every 10ms. We concatenate the MFCCs with their first- and second-order numerical time derivatives to form a 39-dimensional feature vector. We also center and whiten these features to have zero mean and identity covariance.

The resulting dataset contains 6300 sequences of 39-dimensional features, where the sequence lengths vary from 90 to 777 with an average length of 305. See Figure 6.2 for a histogram of sequence lengths. The total number of features is 1,925,362 frames.

In addition, we follow Lee and Glass [68] and use the changepoint detector of Glass [41] to accelerate our training algorithm. We include these detected possible changepoints while training our model to reduce the complexity of the message passing computation using the methods developed for the energy disaggregation application of Chapter 3. See Figure 6.3 for a typical set of examples showing the detected changepoints and the true changepoints.

The TIMIT dataset is also fully expert-labeled with phonetic units, and we make use of a small subset of these labels to set our priors and initialize our fitting procedure, as we describe in the subsequent sections.

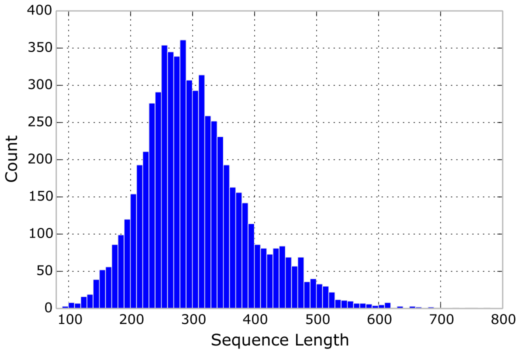


Figure 6.2: Histogram of TIMIT sequence lengths.

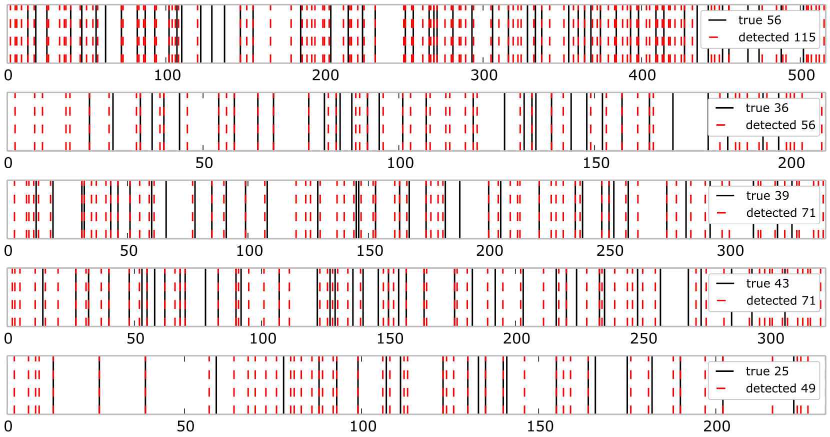
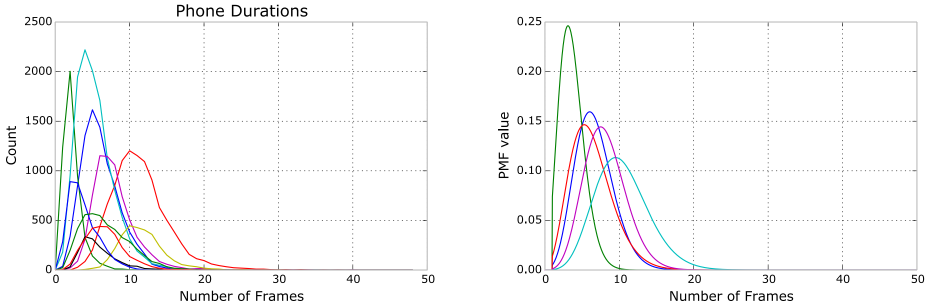


Figure 6.3: Detected possible changepoints

■ 6.6.2 Setting informative duration priors

We wish to use informative duration priors to encourage the model to learn interpretable phonetic units. In this subsection we describe a general method for setting duration hyperparameters.

Phonetic units have durations that are well-modeled by negative binomial distributions; see Figure 6.4(a) for typical phonetic unit duration distributions from the labels in the TIMIT dataset. Recall from Chapter 4 that a negative binomial distribution has parameters (r, p), where $r \in \{1, 2, \ldots, r_{\max}\}$ and $0 < p < 1$. We use priors of the form



(a) Ground-truth durations

(b) Prior samples

Figure 6.4: Phonetic unit durations in the labeled dataset and informative priors set using the method of Section 6.6.2

$p(r, p) = p(r)p(p|r)$ where

$$p(r = j|\nu) = \nu_j \qquad p(p|r = j) = \text{Beta}(a_j, b_j) \tag{6.6.1}$$

where (ν_j, a_j, b_j) for $j = 1, 2, \ldots, r_{\max}$ are the hyperparameters we wish to determine from labeled examples.

A natural way to set hyperparameters is via empirical Bayes [38], in which one chooses hyperparameters to maximize the likelihood of an observed training set. While we have no training set of (r, p) parameter pairs available for such a procedure, we can simulate an appropriate set of parameters by using the Gibbs sampling procedure developed in Section 4.4.2 and some labeled durations. Using a set of durations $\{d_i\}_{i=1}^{S}$ drawn from the expert labels in the TIMIT dataset, we collect samples of (r, p) pairs from $p(r, p|\{d_i\}, \nu^0, a^0, b^0)$, where $\nu_j^0 = \frac{1}{r_{\max}}$ and $a_j^0 = b_j^0 = 1$ for $j = 1, 2, \ldots, r_{\max}$ are chosen to be non-informative. Using these simulated samples $\{(\hat{r}_k, \hat{p}_k)\}_{k=1}^{K}$, we then choose hyperparameters via maximum likelihood. We choose S, the number of duration examples used to set the hyperparameters, to correspond to 2.5% of the labeled examples, and we set K, the number of simulated samples, to be equal to S.

See Figure 6.4(b) for samples drawn from this prior. By comparing these duration distribution samples to the histograms in Figure 6.4(a), it is clear that the prior summarizes the empirical distribution over typical phoneme duration means and variances well. In addition, the negative binomial duration distribution class is able to represent the empirical phoneme duration distributions, which look substantially different from the geometric durations to which we would be restricted with a purely HMM-based approach.

Table 6.1: Fit times and per-frame predictive likelihoods

	Sub-HMM	Sub-GMM	HDP-HMM
Pred. Like. (nats)	-44.046	-47.599	-47.940
Fit Time (min.)	110	67	10

■ 6.6.3 Experimental procedure and results

In this subsection we fit three alternative models, two of which are developed in this thesis, and compare their performance both at prediction and on a phonetic unit segmentation task. First, we fit the nonparametric model developed in this chapter, which we refer to as the HDP-HSMM Sub-HMM model. Second, we fit an HDP-HSMM with Gaussian mixture model (GMM) emissions, which we refer to as the HDP-HSMM Sub-GMM model. This second model is different from the first only in that by using GMM emissions instead of sub-HMM emissions, the internal structure of the phonetic units is not modeled. Finally, we fit an HDP-HMM for comparison. The HDP-HMM does not include the duration prior that the other two models can incorporate. Each model has 39-dimensional Gaussian emission distributions, with Normal Inverse-Wishart priors with hyperparameters set as $\mu_0 = 0$, $\Sigma_0 = I$, $\kappa_0 = 0.5$, and $\nu_0 = 45$. For each of the three models, we are able to scale efficient inference to this large dataset using the algorithms developed in this chapter and in Chapter 5, allowing the models to be fit orders of magnitude faster than previous methods.

Often in speech analysis there is an abundance of unlabeled data but only a very limited amount of labeled data. In such settings, labeled data is used to set priors and initializations, while an unsupervised inference procedure is used with the large amount of unlabeled data. Accordingly, we use the labels from 2.5% of the full dataset to set our prior hyperparameters using Gibbs sampling. We also use this small subset of labeled data to initialize our inference procedure. We perform inference over the unlabeled data in a single pass over the dataset using a minibatch size of 50 sequences and a stepsize sequence $\rho^{(t)} = (t + \tau)^{-\kappa}$ where we chose $\tau = 0$ and $\kappa = 0.6$.

Table 6.1 summarizes both the fitting runtimes and the predictive performance of each model. We measure predictive performance by computing the per-frame predictive likelihood on 20 held-out sequences, where a larger value indicates a higher average likelihood assigned to each held-out frame and hence better predictions. The per-frame predictive likelihoods are very similar, indicating that the alternative models perform comparably well on predictive measures. However, their predictive performance does not give any insight into the interpretability of the latent structure learned, which we discuss next.

To evaluate the quality and interpretability of the learned latent parameters, we

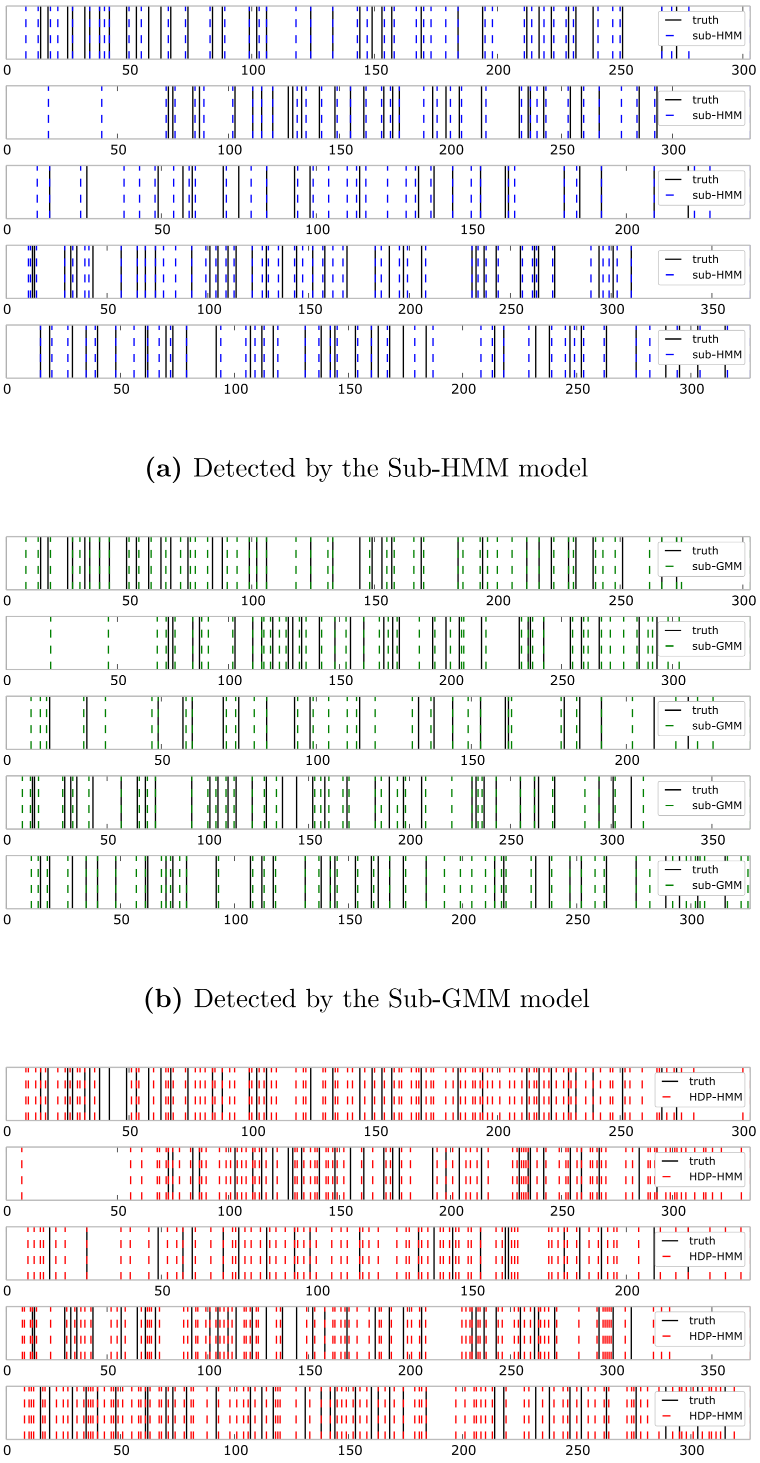
Table 6.2: Error rates for the segmentation task

	Sub-HMM	Sub-GMM	HDP-HMM
Missed Detections	22.0	21.9	24
False Positives	31.9	35.0	59.8

consider a segmentation task similar to the one considered by Lee and Glass [68]. On the 20 held-out sequences and using no changepoint information from the changepoint detector, we compute the optimal variational factor over the label sequence (or state sequence in the case of the HDP-HMM) and then perform a Viterbi decoding to find the most probable joint assignment according to that variational factor. Finding this most probable label sequence (or state sequence) assignment evaluates each model's ability to discover modes that correspond to phonemes, where the HDP-HMM is unable to distinguish the dynamics at multiple timescales present in the data. We then compare the changepoints in the Viterbi sequence to the true changepoints and measure both the missed detection and false positive error rates. Following Lee and Glass [68] and Scharenborg et al. [100], we allow a 20ms tolerance window to compute detections.

We summarize the segmentation performance of the three models in Table 6.2. We find that both of the models which include explicit duration modeling perform significantly better than the HDP-HMM at both missed detection and false positive error rates. In addition, we find that modeling the dynamics within each phonetic unit with the Sub-HMM model further reduces the false positive rate. The HDP-HMM, which cannot separate timescales because it lacks explicit duration modeling, tends to over-segment relative to the intended phonetic unit segmentation, leading to a very high false positive error rate. The Sub-HMM changepoints also perform well qualitatively; in Figure 6.5 we show 5 typical examples of the changepoints detected by each model.

These experiments demonstrate advantages of both our algorithms and our models. With our SVI algorithms we are able to perform inference in a single pass over the dataset in the time it would require to compute a single Gibbs sampling or batch mean field update. Thus our algorithms allow inference in each of these models to scale to large datasets efficiently, reducing the computation time by orders of magnitude and enabling even larger datasets to be explored. By comparing three related models, we also show that explicit duration modeling provides a significant boost to segmentation performance, with the Sub-HMM model refinement providing a further increase in performance. These models and algorithms may provide new tools for speech researchers to analyze detailed structure while imposing model regularities with interpretable prior information.



(a) Detected by the Sub-HMM model

(b) Detected by the Sub-GMM model

(c) Detected by the HDP-HMM model

Figure 6.5: Phonetic unit boundaries detected by the three models.

■ 6.7 Conclusion

This chapter composes the ideas of Chapters 3, 4, and 5 to develop both new models and new efficient and scalable algorithms. In particular, it shows that the ideas developed in this thesis can be readily extended. The flexible Bayesian nonparametric approach to modeling dynamics at multiple timescales may provide new insights into complex phenomena, and the algorithms we develop enable such rich models to be fit to large enough datasets. Finally, our speech application shows the promise and potential utility of explicit duration modeling in a Bayesian framework, both for performance and for interpretability.