A Left-to-Right HDP-HMM with HDPM Emissions

Amir Hossein Harati Nejad Torbati, Joseph Picone
Department of Electrical and Computer Engineering
Temple University
Philadelphia, USA
amir.harati@gmail.com, picone@temple.edu

Marc Sobel
Department of Statistics
Temple University
Philadelphia, USA
marc.sobel@temple.edu

***Abstract*— In this paper we introduce a new nonparametric Bayesian HMM based on the well-known HDP-HMM model. Unlike the original ergodic model, our model has a left-to-right structure. We introduce two approaches to adding non-emitting states that are used to model the beginning and end of finite duration sequences. Finally, we extend the HDP-HMM definition by introducing an HDP-HMM with HDP mixture emissions. We demonstrate that the new model outperforms the ergodic model for problems involving temporal structure by producing a 15% increase in likelihoods. Experiments on a phoneme classification task resulted in an 15.3% relative reduction in error.**

Keywords—HDP-HMM; none-parametric Bayesian; Left-to-Right models; HMMs; Hierarchical Dirichlet Model

Regular Research Paper

I. Introduction

Hidden Markov models (HMMs) [1] are among the most powerful statistical modeling tools and have found a wide range of applications in many pattern recognition tasks such as speech recognition, machine vision, genomics and finance [2]. HMMs are parameterized both in their topology (e.g. number of states) and emission distributions (e.g. Gaussian mixtures). Model comparison methods are traditionally used to optimize the number of states and mixture components. However, these methods are computationally expensive and moreover there is no consensus on an optimum criterion for the selection [3].

An infinite HMM has been developed in the last few years [4][5][6] based on nonparametric Bayesian approaches. In this model, instead of defining a parametric prior over the transition distribution, a hierarchical Dirichlet process (HDP) prior is used. This model is known as an HDP-HMM model. HDP-HMM introduced in [5] and [6] is an ergodic model (a transition from an emitting state to all other states is allowed). However, in many pattern recognition applications involving temporal structure, such as speech processing, a left-to-right topology is preferred or sometimes required [7][8]. For example, in continuous speech recognition applications we model speech units (e.g. phonemes), which evolve in a sequential manner, using HMMs. Since we are dealing with an ordered sequence (e.g. a word is an ordered sequence of phonemes), a left-to-right model is preferred [7]. Moreover, the segmentation of speech data into these units is not known in advance, and therefore the training process must be able to connect these smaller models together into a larger HMM that models the entire utterance. Obviously, this task can easily be achieved using left-to-right (LR) HMMs.

If the data has a finite length, the beginning and end of a sequence is typically modeled as two additional discrete events – non-emitting initial and final states [1][7]. In the original HDP-HMM formulation [5][6], this problem is not addressed. Also, the original HDP-HMM, as well as parametric HMMs, models each emission distribution by data points mapped to that state. For example, if we use a Gaussian mixture model (GMM) to model the emission distribution, for every state we compute a separate GMM and components can't be shared or re-used within a model. In this paper we propose a left-to-right HDP-HMM (LR HDP-HMM) with non-emitting initial and final states. In our model, emission distributions are modeled using GMMs with an infinite number of components. Sharing components is achieved by using an HDP prior instead of Dirichlet process (DP) priors as in [6].

The paper is organized as follows. In Section 2, we introduce Dirichlet processes and the HDP-HMM model. In Section 3, our proposed model is discussed. In Section 4, we present some experimental results on two datasets. We conclude the paper in Section 5 with a discussion of the limitations of the current model and future work.

II. Background

A Dirichlet process [9] is a discrete distribution that consists of countable infinite probability masses. A DP is denoted by DP(α,H), where α is the concentration parameter and H is the base distribution. A DP can be represented by [10]:

$$G=\sum_{k=1}^{\infty}\beta_k\delta_{\theta_k}, \qquad \theta_k \sim H \ . \tag{1}$$

In this definition, δ_{θ_k} is the unit impulse function at θ_k, and is referred to as an atom [5]. The weights β_k are sampled through a stick-breaking construction [5][10]:

$$\beta_k = v_k\prod_{l=1}^{k-1}\left(1-v_l\right), \qquad v_k \mid \alpha, G_0 \sim Beta\left(1,\alpha\right). \tag{2}$$

The sequence of β_k sampled by this process satisfies the constraint $\sum_{k=1}^{\infty} \beta_k = 1$ with probability 1 and are denoted by $\beta \sim GEM(\alpha)$ [5]. One of the main applications of a DP is to define a nonparametric prior distribution on the components of a mixture model. For example, a DP can be used to define a Gaussian mixture model (GMM) with an infinite number of mixture components [11]. This is a useful model in many areas of science. For example, in speech recognition, an acoustic unit (a word or a phoneme) can be modeled using a GMM [1].

A hierarchical Dirichlet process extends a DP to grouped data [5]. In this case there are several related groups and the goal is to model each group using a mixture model. These models can be linked using traditional parameter sharing approaches. For example, consider the problem of modeling acoustic units, such as phonemes, in continuous speech recognition using a mixture model in which parameters of different acoustic units can be shared. One approach is to use a DP to define a mixture model for each group and to use a global Dirichlet process, $DP(\gamma,H)$, as the common base distribution for all DPs [5]. An HDP is defined as:

$$\begin{aligned} &G_0 \mid \gamma, H \sim DP(\gamma, H) \\ &G_j \mid \alpha, G_0 \sim DP(\alpha, G_0)\ , \end{aligned} \tag{3}$$

where H provides the prior for the parameters and G_0 represents the average of the distribution of the parameters (e.g. means and covariances).

An alternative analogy, which is useful for gaining insight into the inference algorithms, is based on the concept of a Chinese restaurant franchise (CRF) [5]. In a CRF, a franchise consists of several restaurants with a common franchise-wide menu. Customers represent observed data, tables represent clusters and restaurants represent groups. The first customer entering restaurant j sits at one of the tables and orders an item from the menu. The next customer either sits at one of the occupied tables and eats the food served at that table or sits at a new table and orders new food from the menu. The probability of sitting at a table is proportional to the number of customers already seated at that table. However, if a customer starts a new table (with probability proportional to α), he or she orders food from the menu with a probability proportional to the number of tables serving that food in the franchise, or alternately orders a new food item with a probability proportional to γ.

An HDP-HMM [4][5][6] is an HMM with an unbounded number of states. In a typical ergodic HMM, the number of states is fixed so a matrix of dimension N states by N transitions per state is used to represent the transition probabilities. In an HDP-HMM, the transition matrix is replaced by an infinite, but discrete transition distribution, modeled by an HDP for each state. This lets each state have a different distribution for its transitions while the set of reachable states would be shared among all states. Fox et al. [6] extended the definition of HDP-HMM to HMMs with state persistence by introducing a sticky parameter κ. The definition for HDP-HMM is given by:

$$\begin{aligned} &\beta \mid \gamma \sim GEM(\gamma) \\ &\pi_j \mid \alpha, \beta \sim DP(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}) \\ &\psi_j \mid \sigma \sim GEM(\sigma) \\ &\theta_{kj}^{**} \mid H, \lambda \sim H(\lambda) \\ &z_t \mid z_{t-1}, \left\{\pi_j\right\}_{j=1}^{\infty} \sim \pi_{z_{t-1}} \\ &s_t \mid \left\{\psi_j\right\}_{j=1}^{\infty}, z_t \sim \psi_{z_t} \\ &x_t \mid \left\{\theta_{kj}^{**}\right\}_{k,j=1}^{\infty}, z_t \sim F\left(\theta_{z_t s_t}\right)\ . \end{aligned} \tag{4}$$

The state, mixture component and observation are represented by z_t, s_t and x_t respectively. The indices j and k are indices of the state and mixture components respectively. The base distribution that links all DPs together is represented by β and can be interpreted as the expected value of state transition distributions. The transition distribution for state j is a DP denoted by π_j with a concentration parameter α. Another DP, ψ_j, with a concentration parameter σ, is used to model an infinite mixture model for each state (z_j). The distribution H is the prior for the parameters θ_{kj}. If we want the posterior distribution over the parameters to remain in the same family as the prior, then H should be chosen to be a conjugate prior to the observation likelihood. Since the likelihood has a multivariate normal distribution, H should have normal inverse Wishart (NIW) distribution.

III. A LEFT-TO-RIGHT HDP-HMM WITH HDPM EMISSIONS

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain [1]. The state of a Markov chain at time t is denoted by z_t and an observation is denoted by $x_t \sim F(\theta_{z_t,s_t})$ where F is the emission distribution (e.g., a Gaussian mixture) and s_t is a mixture component index. In an HMM, there is a probability distribution to transit into state z_t. In an infinite HMM, this transition distribution should have infinite support and is modeled using HDP. For state j this transition distribution is denoted by π_j:

$$\pi_j \mid \alpha, \beta \sim DP(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}). \tag{5}$$

From (5) we can see that the transition distribution has no topological restriction and therefore (4) defines an ergodic HMM. In this section we introduce a left-to-right HDP-HMM with initial and final non-emitting states. Moreover, we replace DP with HDP to model multimodal emission distributions that allow states to share mixture components.

A. *Left-to-Right Transition Distributions*

In order to obtain a left-to-right (LR) topology we need to force the base distribution of the Dirichlet distribution in (5) to only contain atoms to the right of the current state. This means β should be modified so that the probability of transiting to states left of the current state (i.e. states previously visited) becomes zero. For state j we define $V_j=\{V_{ji}\}$ as:

$$V_{ji} = \begin{cases} 0, & i < j \\ 1, & i \geq j \end{cases} \tag{6}$$

where i is index for all states. Then we can modify β by multiplying it with V_j:

$$\beta' = \frac{\beta \bullet V_j}{\sum_i \beta_i V_{ji}}. \tag{7}$$

Therefore to obtain a left-to-right HDP-HMM, which we refer to as LR HDP-HMM, we simply replace β' with β in (5). The rest of the definition remains the same. Also notice that different topologies can be achieved by defining an appropriate V_j .

B. *Initial and Final Non-Emitting States*

In many applications, such as continuous speech recognition, a LR HMM begins from and ends with non-emitting states. These states are required to model the beginning and end of finite duration sequences. Adding a non-emitting initial state is trivial: the probability of transition into the initial state is 1 and the probability distribution of a transition from this state is equal to π_{init} which is the initial probability distribution for an HDP-HMM without non-emitting states. However, adding a final non-emitting state is more complicated. In the following we will discuss two approaches to solving this problem.

1) *Maximum Likelihood Estimation*

Consider state z_i depicted in Figure 1. The outgoing probabilities for any state can be classified into three categories: (1) a self-transition (P1), (2) a transition to all other states (P2), and (3) a transition to a final non-emitting state (P3). These probabilities must sum to 1: P1+P2+P3=1. Suppose that we obtained P2 from the inference algorithm. We will need to reestimate P1 and P3 from the data. This problem is, in fact, equivalent to the problem of tossing a coin until we obtain the first tails. Each head is equal to a self-transition and the first tails triggers a transition to the final state. This can be modeled using a geometric distribution [12]:

$$P(x=k) = (1-\rho)^{k-1}\rho. \tag{8}$$

Equation (8) shows the probability of *K-1* heads before the first tail. In this equation $1-\rho$ is the probability of heads (success). We also have:

$$\frac{P_1}{1-P_2} = 1-\rho, \qquad \frac{P_3}{1-P_2} = \rho. \tag{9}$$

Suppose we have a total of N examples but for just M examples the state z_i is the last state of the model (S_M). It can be shown [12] that the maximum likelihood estimation is obtained by:

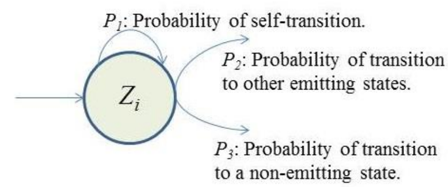

Figure 1- Outgoing probabilities for state z_i

$$\hat{\rho} = \frac{M}{\sum_{i \in S_M} k_i} \tag{10}$$

where k_i are the number of self-transitions for state i. Notice that if z_i never happens to be the last state (*M=0*), *P3=0*.

2) *Bayesian Estimation*

Another approach to estimate ρ is to use a Bayesian framework. Since a beta distribution is the conjugate distribution for geometric distribution [13], we can use a beta distribution with hyperparameters *(a,b)* as the prior and obtain a posterior as [13][14]:

$$\rho \sim Beta\left(a+M, b+\sum_{i \in S_M}(k_i - 1)\right) \tag{11}$$

where M and S_M are same as in the previous section. Hyperparameters (a,b) can also be estimated using a Gibbs sampler if required [15].

C. *HDP Mixture Emission Distributions*

In previous works [5][6], emission distributions for each state of an HDP-HMM were modeled using a Dirichlet process mixture (DPM) as shown in (4). While this model is reasonably flexible, each data point is strictly associated with a single state and hence statistical estimation of each parameter would be less reliable. This is a more serious problem for HDP-HMMs with a left-to-right topology since these models will discover more states. As a result the available data for estimating the emission distribution for each state would be more limited. The solution proposed here is to replace the DPM with an HDP mixture (HDPM) defined for the entire HMM. The final model without non-emitting states, which we refer to as LR HDP-HMM/HDPM, is defined by (12) and is displayed in Figure 2-(b). For comparison purposes, we display the original HDP-HMM in Figure 2-(a) [6].

$$
\begin{aligned}
&\beta \mid \gamma \sim GEM(\gamma) \\
&\beta' = \frac{V_j \cdot \beta}{\sum_i V_{ji}\beta_i}, V_{ji} = \begin{cases} 0, & i < j \\ 1, & i \ge j \end{cases} \qquad 1 \le i < \infty \\
&\pi_j \mid \alpha, \beta' \sim DP(\alpha + \kappa, \frac{\alpha\beta' + \kappa\delta_j}{\alpha + \kappa}) \\
&\xi \mid \sigma \sim GEM(\sigma) \\
&\psi_j \mid \tau, \xi \sim DP(\tau, \xi) \\
&\theta_{kj}^{**} \mid H, \lambda \sim H(\lambda) \\
&z_t \mid z_{t-1}, \{\pi_j\}_{j=1}^{\infty} \sim \pi_{z_{t-1}} \\
&s_t \mid \{\psi_j\}_{j=1}^{\infty}, z_t \sim \psi_{z_t} \\
&x_t \mid \{\theta_{kj}^{**}\}_{k,j=1}^{\infty}, z_t \sim F\left(\theta_{z_t s_t}\right)
\end{aligned}
\tag{12}
$$

D. *Modified Block Sampler*

A block sampler for HDP-HMM with a multimodal emission distribution has been introduced by Fox et al. [6]. In this section we review the modifications of this algorithm needed for our new model. The interested reader should refer to [6][16] for additional details. The central idea is to jointly sample the state sequence $z_{1:T}$ given the observations, model parameters and transition distribution π_j. A variant of forward-backward procedure [1] is utilized that allows us to exploit the Markovian structure of the HMM. However it requires approximation of the theoretically infinite distributions with a "degree L weak limit" approximation that truncates a DP into a Dirichlet distribution with L dimensions [17]:

$$GEM_L(\alpha) \triangleq Dir\left(\frac{\alpha}{L}, ..., \frac{\alpha}{L}\right). \tag{13}$$

The sampling of the transition distribution is similar to [6]. The only difference is to replace β with β' given in (7). Using a similar approximation we can write the following prior distributions for the global weights ξ and state-specific weights ψ_j used in the HDPM emission distributions.

$$\xi \mid \sigma \sim Dir\left(\frac{\sigma}{L'}, ..., \frac{\sigma}{L'}\right) \tag{14}$$

$$\psi_j \mid \xi, \tau \sim Dir\left(\tau\xi_1, ..., \tau\xi_{L'}\right) \tag{15}$$

where L' is the order of approximation in this case. For the posterior distribution we can write:

$$\xi \mid M, \sigma \sim Dir\left(\frac{\sigma}{L'} + M_{\bullet 1}, ..., \frac{\sigma}{L'} + M_{\bullet L'}\right) \tag{16}$$

$$\psi_j \mid \tau, \xi, Z_{1:T}, S_{1:T} \sim Dir\left(\tau\xi_1 + n'_{j1}, ..., \tau\xi_{L'} + n'_{jL'}\right) \tag{17}$$

where M_{jk} is the number of tables (clusters) in restaurant (state) j that serves dish (mixture component) k; $M_{\bullet k}$ is total number of tables in the franchise that serves dish k. The number of observations in state j that are assigned to component k is denoted by n'_{jk} . Estimating transition probabilities for the final non-emitting state can be done as a last step and after estimating the other parameters.

IV. EXPERIMENTS

Synthetic data. In the first experiment, we generate data from a left-to-right HMM without non-emitting states that consists of four states. The emission distribution for each state is a GMM with up to three components, each consisting of a two-dimensional normal distribution. Three synthetic data sequences totaling 1900 observations were generated for training. Three configurations have been studied: (1) an ergodic HDP-HMM, (2) a LR HDP-HMM with DPM emissions and (3) a LR HDP-HMM with HDPM emissions. An NIW prior is used for the mean and covariance. The truncation levels are set to 10 for both the number of states and the number of mixture components. Parameters of the NIW are set as follows: pseudocounts, the number of pseudo observations for the sample mean, is set to 0.1; the sample mean and covariance are set to the empirical mean and covariance; and degree of freedom, which is the precision on sample covariance, is set to 5.

Figure 3-(a) shows the average likelihoods for different models for held-out data by averaging five independent chains. Figure 3-(b) shows the structure of the models. The LR HDP-HMM/HDPM discovers the correct structure while the ergodic HDP-HMM finds a more simplified HMM. Moreover, we can see using HDP emissions improves the likelihood. While LR HDP-HMM/DPM can find the structure close to the correct one (not shown here), its likelihood is slightly less than that for the ergodic HDP-HMM. However, LR HDP-HMM/HDPM

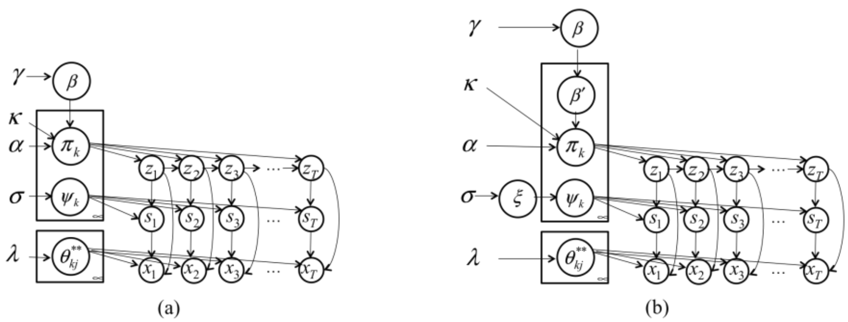

Figure 2- A comparison of models: (a) ergodic HDP-HMM [6] (b) proposed LR HDP-HMM/HDPM.

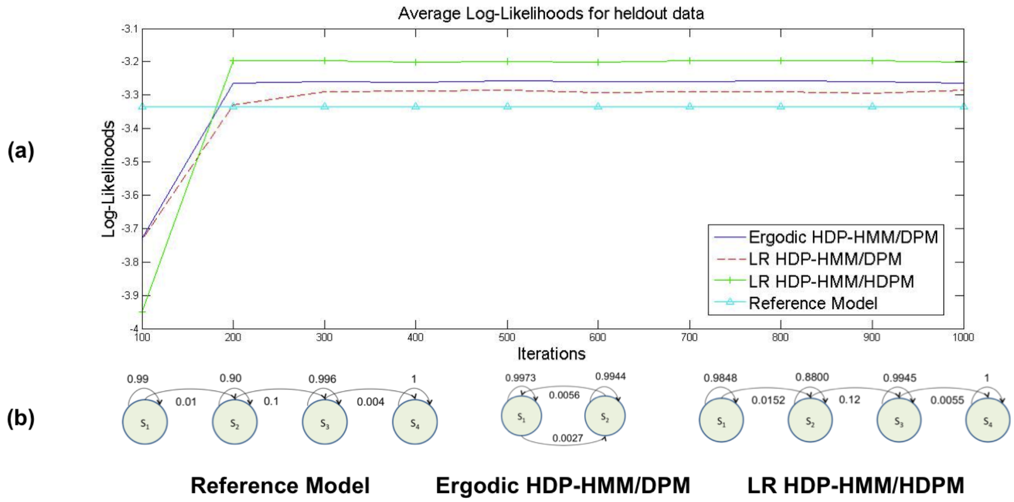

Figure 3- A comparison of (a) log-likelihoods of the proposed models to an ergodic model, and (b) the corresponding model structures.

produces a 15% improvement in likelihoods compared to the ergodic model. It is also interesting to note that the likelihoods of models discovered by all HDP-HMM algorithms are superior to the likelihood of the reference model itself.

TIMIT Classification. The TIMIT Corpus [18] is one of the most cited evaluation data sets used to compare new speech recognition algorithms. The data is segmented manually into phonemes and therefore is a natural choice to evaluate phoneme classification algorithms. TIMIT contains of 630 speakers from eight main dialects of American English [18]. The total numbers of utterances are 6300 where 3990 utterances are the standard training set and 150 utterances are core test set. We followed the standard practice of building models for 48 phonemes and then map them into 39 phonemes [19]. The first 12 Mel-Frequency Cepstral Coefficients (MFCCs) plus energy and their first and second derivatives features have been used to convert speech data into 39-dimensional feature streams. In this experiment, LR HDP-HMMs with Gaussian and DPM emissions have been used. We have used non-conjugate priors and placed a Gaussian prior on the mean and inverse-Wishart prior on the covariance matrix. Truncation levels are set to 10.

Table 1 compares the classification error of the left-to-right models and the parametric models. Since the maximum number of mixture components is set to 10, we have compared our systems to parametric HMMs with 10 components per state. As this table shows, even left-to-right HDP-HMM with Gaussian emissions outperforms the parametric model.

Table 1- A comparison of classification error rates

Model	Classification Error Rate
Parametric HMM [19] (10 mixtures)	27.8%
LR HDP-HMM with Gaussian emissions	26.7%
LR HDP-HMM with DPM emissions	24.1%

Figure 4 shows the discovered structure for phonemes /aa/ and /sh/ using the proposed model. As the amount of data increases the system can learn a more complex model for the same phone. It is also important to note that the structure learned for each phone is different and reflects underlying differences between phonemes. Also note that the learned structure models multiple modalities by learning several parallel left-to-right paths. This is shown in Figure 4-(c), where S1-S2, S1-S3 and S1-S4 depict three parallel models.

V. Conclusion

In this paper we introduced a left-to-right HDP-HMM with HDPM emissions. We have shown that the new model can successfully learn the underlying structure when the data is generated using a generative left-to-right model. Moreover, it has been shown that the likelihood of the learned model is higher than the ergodic model. In this paper we have also introduced two approaches to adding non-emitting initial and final states to the left-to-right HDP-HMM model. Finally we presented the modifications needed in the block sampler to implement the inference algorithm for the new model. Through experimentation on TIMIT, we have shown that the proposed model outperforms parametric HMMs and can learn multimodal structure from the data.

One of the current problems of the HDP-HMM model (including left-to-right model) is that the inference algorithm is still computationally expensive. It is a serious problem when we are dealing with large datasets such as in speech or video processing applications. Therefore, our next task is to improve the inference algorithm specifically for left-to-right HDP-HMMs with HDPM emissions using its specific properties and structure. For example, due to the left-to-right constraints, the number of possible transitions in state 1 is L, in state 2 is L-1 and in state L is 1. We can exploit this fact to reduce the computational complexity.

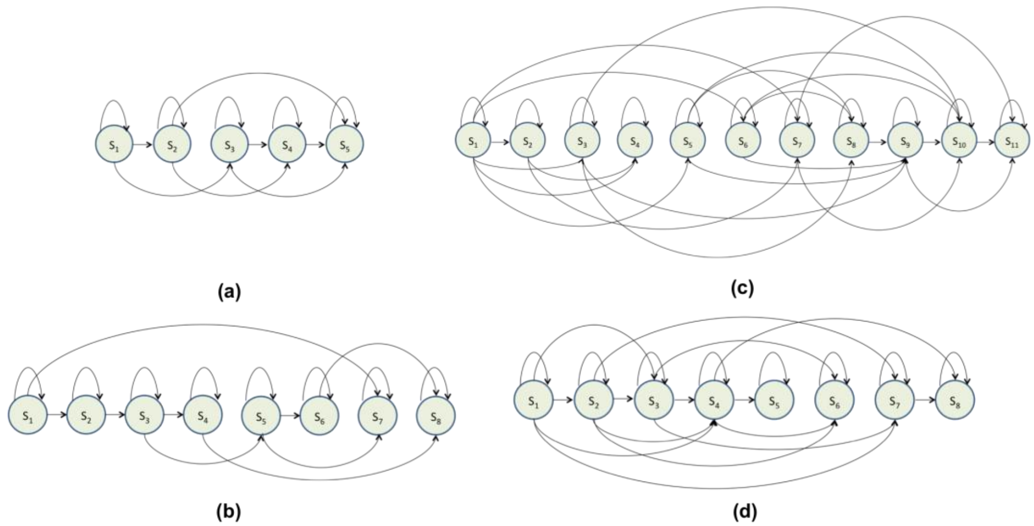

Figure 4- An automatically derived model structure for a left-to-right HDP-HMM model (without the first and last non-emitting states) for (a) */aa/* with 175 examples (b) */sh/* with 100 examples (c) */aa/* with 2,256 examples and (d) */sh/* with 1,317 examples. The data used in this illustration was extracted from the training portion of the TIMIT Corpus.

Another possible direction is to replace HDP emissions with more general hierarchical structures such as a Dependent Dirichlet Process [20] or an Analysis of Density (AnDe) model [21]. It has been shown that the AnDe model is the appropriate model for problems involves sharing statistical strength among multiple set of density estimators [5][21].

ACKNOWLEDGMENT

This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854.

REFERENCES

[1] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989

[2] P. Dymarski, Hidden Markov Models, Theory and Applications. InTech Open Access Publishers, 2011

[3] J. B. Kadane and N. A. Lazar, "Methods and Criteria for Model Selection," Journal of the American Statistical Association, vol. 99, no. 465, pp. 279–290, 2004.

[4] M. Beal, Z. Ghahramani, and C. E. Rasmussen, "The Infinite Hidden Markov Model," in Proceedings of Neural Information Processing Systems, 2002, pp. 577–584.

[5] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet Processes," Journal of the American Statistical Association, vol. 101, no. 47, pp. 1566–1581, 2006.

[6] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "A Sticky HDP-HMM with Application to Speaker Diarization.," The Annalas of Applied Statistics, vol. 5, no. 2A, pp. 1020–1056, 2011.

[7] B.-H. Juang and L. Rabiner, "Hidden Markov Models for Speech Recognition," Technometrics, vol. 33, no. 3, pp. 251–272, 1991.

[8] G. A. Fink, "Configuration of Hidden Markov Models From Theory to Applications," in Markov Models for Pattern Recognition, Springer Berlin Heidelberg, 2008, pp. 127–136.

[9] Y.-W. Teh, "Dirichlet process," in Encyclopedia of Machine Learning, Springer, 2010, pp. 280–287.

[10] J. Sethuraman, "A constructive definition of Dirichlet priors," Statistica Sinica, vol. 4, no. 2, pp. 639–650, 1994.

[11] C. E. Rasmussen, "The Infinite Gaussian Mixture Model," in Proceedings of Advances in Neural Information Processing Systems, 2000, pp. 554–560.

[12] J. Pitman, Probability. Springer-Verlag, 1993.

[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Bayesian Data Analysis, 2nd ed. Chapman & Hall, 2004.

[14] P. Diaconis, K. Khare, and L. Saloff-Coste, "Gibbs Sampling, Conjugate Priors and Coupling," Sankhya A, vol. 72, no. 1, pp. 136–69, 2010.

[15] F. A. Quintana and W. Tam, "Bayesian Estimation of Beta-binomial Models by Simulating Posterior Densities," Journal of the Chilean Statistical Society, vol. 13, no. 1–2, pp. 43–56, 1996.

[16] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Supplement to ' A Sticky HDP-HMM with Application to Speaker Diarization'," The Annals of Applied Statistics, vol. S, no. 2A, pp. S1–S32, 2010.

[17] H. Ishwaran and M. Zarepour, "Exact and approximate sum representations for the Dirichlet process.," Canadian Journal of Statistics, vol. 30, no. 2, pp. 269–283, 2002.

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," The Linguistic Data Consortium Catalog, 1993.

[19] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, "Hidden Conditional Random Fields for Phone Classification," in Proceedings of INTERSPEECH, 2005, pp. 1117–1120.

[20] S. N. MacEachern, "Dependent Nonparametric Processes," in ASA Proceedings of the Section on Bayesian Statistical Science, 1999, pp. 50–55.

[21] G. Tomlinson and M. Escobar, "Analysis of Densities," Technical Report, University of Toronto, Toronto, Canada, 1999.