APPLICATIONS OF DIRICHLET PROCESS MIXTURES TO SPEAKER ADAPTATION

Amir Hossein Harati Nejad Torbati and Joe Picone

Dept. of Elect. and Comp. Eng.
College of Engineering
Temple University, Philadelphia, USA
amir.harati@gmail.com, picone@temple.edu

Marc Sobel

Department of Statistics
Fox School of Business and Management
Temple University, Philadelphia, USA
marc.sobel@temple.edu

ABSTRACT

Balancing unique acoustic characteristics of a speaker such as identity and accent, with general acoustic behavior that describes phoneme identity, is one of the great challenges in applying nonparametric Bayesian approaches to speaker adaptation. The Dirichlet Process Mixture (DPM) is a relatively new model that provides an elegant framework in which individual characteristics can be balanced with aggregate behavior without diluting the quality of the individual models. Unlike Gaussian Mixture models (GMMs), which tend to smear multimodal behavior through averaging, the DPM model attempts to preserve unique behaviors through use of an infinite mixture model. In this paper, we present some exploratory research on applying these models to the acoustic modeling component of the speaker adaptation problem. DPM based models are shown to provide up to 10% reduction in WER over maximum likelihood linear regression (MLLR) on a speaker adaptation task based on the Resource Management database.

Index Terms— nonparametric Bayesian models, Dirichlet Process Mixture, speaker adaptation

1. INTRODUCTION

Nonparametric Bayesian methods provide a mathematically elegant framework that allows inference of model structure and complexity without diluting the purity of modes or clusters [1]. Balancing unique behaviors, such as a speaker's accent, with generalized behavior, such as an expected formant location that is tied to a phoneme's identity, is one of the most challenging aspects of speech processing. In applications such as speech recognition, the number of modalities is large and the space of potential solutions vast. For example, varying the number of states in a hidden Markov model often tends to smear information across states rather than allow states to retain an identity modeling a specific phonetic event. Similarly, clustering of formants using Gaussian mixture models often results in clusters that are averaged across unrelated individual events. Such problems can be mitigated using technologies such as phonetic decision trees, but this often results in intricate and elaborate training processes.

The Dirichlet Process Mixture (DPM) model is a popular application of nonparametric Bayesian methodologies [2]. DPM provides a framework to infer the number of clusters and their parameters jointly. Furthermore, DPM is a data-driven method. The structure evolves as more data become available. Figure 1 demonstrates this phenomenon using a simulation in which the amount of data changes from 20 to 2000 data points and the corresponding discovered clusters increase from one to three. Because of this adaptability to the amount of available data the probability of over-fitting or under-fitting is significantly reduced.

Dirichlet processes have been previously successfully applied to language modeling problems in speech recognition [3]. However, they have not been extensively applied to the acoustic modeling problem due, in part, to the computational issues involved in parameter estimation. Recent advances in fast computational methods, such as variational inference methods [4], have enabled application to computationally intensive tasks such as acoustic training.

As a proof of concept, we investigated the use of the DPM model for speaker adaptation. Specifically, we explored replacing the process of building regression trees in a maximum likelihood speaker adaption (MLLR) system [5] with a DPM-based model. This is a task well-suited to the DPM model because acoustic model adaptation involves gradually adapting Gaussian means trained on large amounts of data to speaker-specific means. All experiments have been designed using the DARPA 1000-word Resource Management (RM) task [6] and the HTK speech recognizer [7]. We have used a publicly available MATLAB library [8] for variational inference.

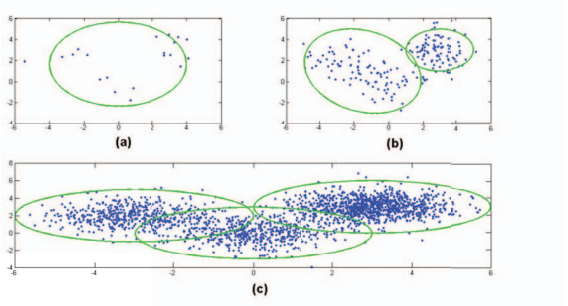

Figure 1. Model complexity as function of available data in DPM base clustering: (a) 20 (b) 200 and (c) 2000 data points.

2. DIRICHLET PROCESS MIXTURE MODELS

The traditional solution to determining the underlying distribution of some observed data is to assume a finite exponential mixture model and infer the parameters. However, the number of components (clusters) must be determined using computationally expensive model selection methods [1][2]. Results have been marginal at best and the clusters often do not capture the underlying characteristics of the data, but rather simply minimize some global distortion measure. An alternative solution that attempts to preserve underlying modalities is the Dirichlet Process Mixture (DPM) model.

The general form of a K-component mixture model is:

$$P(x|\pi,\theta_1,\theta_2,\ldots,\theta_K)=\sum_{k=1}^{K}\pi_k f(x\,|\,\theta_k)\ . \tag{1}$$

In this formulation, π_k are the mixing proportions and must be positive and sum to one, f is usually a parametric distribution (i.e. Gaussian) with parameters θ_k. The finite mixture model can be expressed as a hierarchical model [1]:

$$\begin{aligned} x_i\,|\,\{\bar{\theta}_k\} &\sim f(x\,|\,\bar{\theta}_i)\\ \bar{\theta}_i &\sim G\\ G(\theta)&=\sum_{k=1}^{K}\pi_k\delta(\theta,\theta_k)\\ \pi &\sim Dir(\frac{\alpha}{k},\ldots,\frac{\alpha}{k})\\ \theta_k &\sim H\ . \end{aligned} \tag{2}$$

In this formulation α is pseudocount hyper-parameter of Dirichlet priors. H is the prior distribution over the parameters θ_k and δ is the Kronecker delta function. For a Gaussian mixture model, $\theta_k=\{\mu_k,\Lambda_k\}$ where μ_k is the mean vector and Λ_k is the covariance matrix. H is chosen to be the conjugate prior of f. For a Gaussian distribution it would be a normal-inverse-Wishart distribution.

It can be shown [1][2] that when $k\to\infty$, $G\sim DP(\alpha,H)$ would be a Dirichlet process (DP) with a base distribution of H and a concentration parameter α. One of the most important properties of G is its discrete nature that results in the clustering property of a DP.

The predictive distribution for a DP is given by [1]:

$$P(\bar{\theta}_{N+1}=\theta\,|\,\bar{\theta}_1,\ldots,\bar{\theta}_N,\alpha,H)=\frac{1}{\alpha+N}\left(\alpha H+\sum_{k=1}^{K}N_k\delta(\theta,\theta_k)\right) \tag{3}$$

In this formulation N is the total number of observations and N_k is the number of previous observations for cluster k. It states that the probability of assigning a new observation to cluster k is proportional to its size and the probability of initiating a new cluster is proportional to α.

Direct computation of the posterior probability in [3] is intractable; therefore some kind of approximation should be used. The most popular approaches are based on Monte Carlo Markov chain (MCMC) methodologies and particularly Gibbs Sampling methods [1][2]. However, MCMC based methods can be slow to converge and cannot be used in large-scale problems [1][2][4]. A different class of alternative approaches is based on variational inference, in which we recast the inference problem in terms of optimization [8] and then relax the optimization problem to obtain a tractable solution. Mean-field algorithms, which restrict the variational distribution to a factorization model, have been used in inference from DPM [4][8].

In variational inference, the posterior probability $P(Z\,|\,X)$ is approximated with an arbitrary function $q(Z)$. In other words, because the exact form of $P(Z\,|\,X)$ is not known, an approximation is assumed. In the case of mean-field algorithm this approximation is also factorized. The log marginal probability is given by [8]:

$$\begin{aligned} &\ln P(X)=L(q)+KL(q\,\|\,p)\\ &L(q)=\int q(Z)\ln\left\{\frac{P(X,Z)}{q(Z)}\right\}dZ\\ &KL(q\,\|\,p)=-\int q(Z)\ln\left\{\frac{P(Z\,|\,X)}{q(Z)}\right\}dZ\\ &q(Z)=\prod_{i=1}^{M}q_i(Z_i)\ . \end{aligned} \tag{4}$$

The goal of optimization problem is to maximize the lower bound $L(q)$ or equivalently minimize the Kullback–Leibler (KL) divergence with respect to $q(Z)$. It has been shown that the general solution to this optimization problem follows the form of [9]:

$$\ln q_i^*(Z_i)=\int\ln P(X,Z)\prod_{i\neq j}q_j\,dZ_j+const \tag{5}$$

The above expression does not give an explicit closed-form solution; instead it provides the means to obtain the solution iteratively. Because of the convexity of the bound, the convergence is guaranteed [8]. However, it might converge to a local solution. In [4], the authors used a truncated stick-breaking representation for the variational distribution. One of the downsides of this approach is that variational families are not nested over truncation level T [10] and as a result this must be optimized.

This issue is addressed using an accelerated variational Dirichlet process mixture (AVDP) algorithm [10] that can handle extremely large data sets. In [11] two other extensions of the variational inference named collapsed variational stick-breaking (CSB) and collapsed Dirichlet priors (CDP) have been introduced. CSB is based on stick-breaking representation, but the difference here is to

integrate out mixture weights. For CSB, the truncation level T should be specified. CDP, on the other hand, is based on a finite symmetric Dirichlet distribution approximation of a Dirichlet process. For this algorithm, the size of Dirichlet distribution K needs to be specified. Both of these algorithms give comparable results and are considerably faster than Gibbs sampling. In this research we have used AVDP, CSB and CDP inference algorithms.

3. APPLICATION TO SPEAKER ADAPTATION

Maximum Likelihood Linear Regression (MLLR) is a well-known speaker adaptation technique in which mean transformation matrices are estimated using a tree-based clustering process [5]. Clustering is usually accomplished using a regression class tree which is constructed using a centroid splitting algorithm. This algorithm begins with a single node and recursively grows a tree using an ML-based distance measure. However, this is an ad hoc algorithm and its performance is sensitive to the specific training recipe and the amount of data. Further, it is difficult to determine when the algorithm should be stopped.

In this paper, we explore the use of DPM as an alternate clustering algorithm to investigate the potential advantages of this approach. The procedure we employ is as follows:

1. Train speaker independent (SI) models, collecting all mixture components and their frequency of occurrence.
2. Generate samples for each component and cluster them using one of the DPM inference algorithms.
3. Construct a tree structure of the final result.
4. Assign clusters to each component.

Training SI model is done using HTK [7]. MFCC features plus energy and their first and second derivatives have been used. After training, Gaussian components are extracted. Data points are generated for each component proportional to the number of occurrences in the training data and then clustered using one of the DPM inference methods.

Clusters were reorganized in form of a tree for two reasons. First, we need a method to merge clusters to deal with insufficient data. Second, we need this mechanism to be consistent with HTK. The difference with a regular regression tree is in the construction process. While centroid splitting algorithm is a top-down approach, the proposed algorithm starts from the terminal nodes that are obtained using DPM and merges them using a bottom-up Euclidean distance-based approach. Finally components are labeled using a majority-voting scheme. The result of this section is used to compute transformation matrices using a maximum likelihood approach.

4. RESULTS AND DISCUSSION

In this section we summarize the results of several experiments conducted on a speaker adaptation task based on the Resource Management Corpus [6].

4.1. Monophone Models

In this experiment, monophone models with a single Gaussian mixture have been trained. Before training on the speaker-dependent data, speaker-independent models were trained. MLLR models were trained for 12 different speakers. There are 600 training utterance for each speaker in the dataset. Figure **2** shows the word error rate (WER) as number of utterances in training dataset for each speaker changes for both the regression tree and ADVP approaches.

It is evident that ADVP works significantly better than regression tree. The reason is related to the number of clusters discovered by each method. To some extent, result of ADVP resembles broad phonetic classes. For example, distributions related to phoneme "w" and "r" (which are both liquids) are in the same cluster. This is also true for phonemes belonging to other classes however there are some exceptions too since the clustering is done automatically in the feature space and based on the similarities of the distributions.

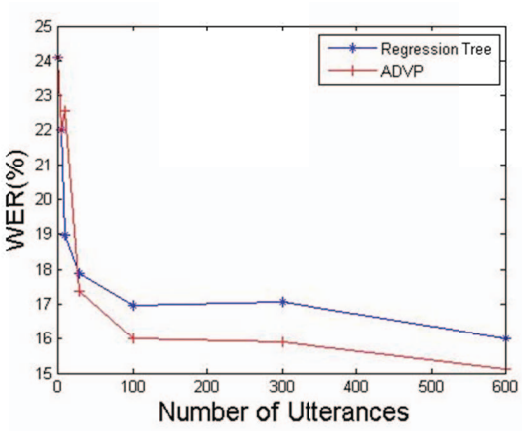

Figure 2. A comparison of regression tree and ADVP approaches for monophone models.

4.2. Cross-word Models

In Figure **3**, we compare results for the speaker adaptation task, but this time cross-word models are used. Several DPM approaches are evaluated: ADVP, CSB and CDP. ADVP performs slightly better than the regression tree approach for medium amounts of training data but works slightly worse when the amount of training data increased. For example, WER for 100 utterances of training data drops from 5.17% to 4.77% and for 600 utterances increases from 3.79% to 4.3%.

It can be seen that CDP and CSB work slightly better than the regression tree approach when the amount of training data is increased. For example, WER decreases from 3.79% for regression tree to 3.53% for both of CDP and CSB algorithms.

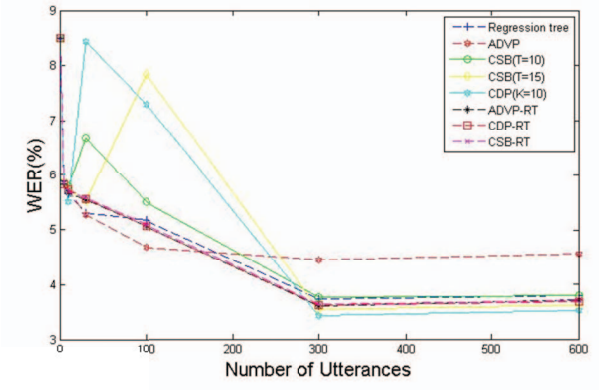

Figure 3. A comparison of WERs between regression tree-based MLLR and several DPM inference algorithms for cross-word acoustic models. ADVP works better for medium data sizes, while other DPM methods work better for large amounts of data.

The clusters generated using DPM have acoustically and phonetically meaningful interpretations. For instance, CSB generates 20 clusters in which distributions corresponding to states two and four of the "silence" model consist a single cluster. The distribution related to the center state of "silence" model (which is also tied to the distribution of short pause model) is not a member of this particular cluster. This is not unexpected since many words in the corpus are articulated with no pause between them and therefore, the model corresponding to that would be more similar to speech models.

Figure **4** shows the number of discovered clusters for each method. DPM-based clustering generally works better when we have a large amount of training data. Therefore, we can combine these systems in order to get better results. The result of cascading ADVP, CDP and CSB with regression tree is presented in Figure **2**. It is evident that cascading a DPM clustering with a regression tree-based clustering slightly improves the results, demonstrating that the two approaches can complement one another.

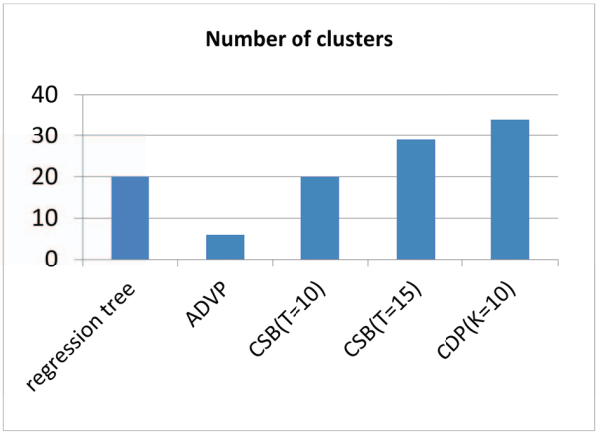

Figure 4. The number of discovered clusters is shown. ADVP generates the fewest clusters.

5. CONCLUSION

In this paper we examined an application of Dirichlet Process Mixtures to an acoustic modeling problem in speaker adaptation. Experimental results were promising, showing that the DPM approach can outperform the classic MLLR approach and decrease the error rate up to 10%. DPM appears to do better when there are larger amounts of data, suggesting application to LVCSR tasks could produce promising results.

In order to fully utilize the potential of nonparametric Bayesian methods we need to redesign the architecture of the speech recognizers and investigate a better reestimation process. Also, in this work we assigned each "distribution" to just one cluster. An obvious extension is to use some form of soft tying. Series and parallel combinations of these systems might also lead to improved results.

6. REFERENCES

[1] E. Sudderth, "Graphical models for visual object recognition and tracking," Ph.D. dissertation, Massachusetts Institute of Technology, May 2006.

[2] J. Paisley, "Machine learning with Dirichlet and beta process priors: Theory and Applications", Ph.D. Dissertation, Duke University, May 2010.

[3] P. Liang, M. I. Jordan, and D. Klein. In T. O'Hagan and M. West "Probabilistic grammars and hierarchical Dirichlet processes", *The Handbook of Applied Bayesian Analysis*, Oxford University Press, 2010.

[4] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[5] C. J. Leggetter, "Improved acoustic modeling for HMMs using linear transformations," Ph.D. Dissertation, University of Cambridge, February 1995.

[6] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-Word Resource Management Database for continuous speech recognition," *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 651-654, New York, New York, USA, April 1988.

[7] *http://htk.eng.cam.ac.uk/docs/docs.shtml*

[8] *https://sites.google.com/site/kenichikurihara/academic-software/variational-dirichlet-process-gaussian-mixture-model*

[9] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, New York, USA, 2007.

[10] K. Kurihara, M. Welling, and N. Vlassis, "Accelerated variational Dirichlet process mixtures," *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, Massachusetts, USA, 2007 (editors: B. Schölkopf and J.C. Hofmann).

[11] K. Kurihara, M. Welling, and Y. W. Teh, "Collapsed variational Dirichlet process mixture models," *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, Jan. 2007.