A Comparative Analysis of Bayesian Nonparametric Inference Algorithms for Acoustic Modeling in Speech Recognition

John Steinberg, Amir Harati and Joseph Picone [1]

[1] Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA

john.steinberg@temple.edu, amir.harati@gmail.com, joseph.picone@isip.piconepress.com

Abstract

Nonparametric Bayesian models have become increasingly popular in speech recognition for their ability to discover data's underlying structure in an iterative manner. Dirichlet process mixtures (DPMs) are a widely used nonparametric method that do not require a priori assumptions about the structure of the data. DPMs, however, require an infinite number of parameters so inference algorithms are needed to make posterior calculations tractable. The focus of this work is an evaluation of three variational inference algorithms for acoustic modeling: Accelerated Variational Dirichlet Process Mixtures (AVDPM), Collapsed Variational Stick Breaking (CVSB), and Collapsed Dirichlet Priors (CDP).

A phoneme classification task is chosen to more clearly assess the viability of these algorithms for acoustic modeling. Evaluations were conducted on the CALLHOME English and Mandarin corpora, consisting of two languages that, from a human perspective, are phonologically very different. It is shown in this work that these inference algorithms yield error rates comparable to a baseline Gaussian mixture model (GMM) but with a factor of 20 fewer mixture components. AVDPM is shown to be the most attractive choice because it delivers the most compact models and is computationally efficient, enabling its application to big data problems.

Index Terms: nonparametric Bayesian methods, variational inference, CALLHOME, phoneme recognition

1. Introduction

Nonparametric Bayesian models have become increasingly popular in speech recognition due to their ability to discover data's underlying structure in an iterative manner [1]. Dirichlet process mixtures (DPMs) are a widely used nonparametric method that do not require a priori assumptions about the structure of data, such as the number of mixture components, and can learn this information directly from the data [1]. This is ideal for acoustic modeling in speech recognition where the number of mixture components is a parameter commonly found by tuning a system using a subset of the data. Typically, the number of components is assumed to be constant since it would be tedious to tune models for each phoneme. DPMs, however, are able to automatically determine an optimal number of mixtures for each individual model.

There are many depictions of Dirichlet processes but the algorithms in this work are all premised on the stick breaking approach shown in Figure 1. In this representation a stick of unit length is broken repeatedly into smaller pieces. Each break represents a new mixture component weight where the fraction of the remaining stick is given by v_i and the absolute length of each piece (i.e. the weight of the mixture component) is given by c_i.

Aside from automatic tuning of the number of mixtures, it is equally important to ensure that these models generalize well across different data. Our long-term interest in nonparametric Bayesian approaches, and advanced statistical models in general, is to develop models that are robust to significant variations in the acoustic channel. Low complexity models that have good generalization are a step in this direction. In this work, the performance of three Bayesian variational inference algorithms – Accelerated Variational Dirichlet Process Mixtures (AVDPM), Collapsed Variational Stick Breaking (CVSB), and Collapsed Dirichlet Priors (CDP) [2][3] – are assessed on both the CALLHOME English (CH-E) and the CALLHOME Mandarin (CH-M) corpora.

1.1.Variational Inference Algorithms

Nonparametric methods such as DPMs, although extremely useful for finding the underlying structure of data, often come at a cost of computational complexity. The term 'nonparametric' is something of a misnomer since DPMs require a potentially infinite number of parameters. This makes manipulating such distributions intractable, so inference algorithms are used to approximate these models. Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling, are extremely popular for their mathematical simplicity [4]-[6]. These methods approximate complex posteriors by sampling latent variables from a Markov chain that represents the distribution of interest [7]. Unfortunately, converging to optimal posterior approximations is often slow and these methods can become intractable for big data problems such as speech recognition [5][7].

Variational inference algorithms approximate a posterior, $p(y|x)$, with a simpler distribution $q(y)$ by making assumptions about the dependencies of the distribution's latent variables. The task of approximating a complex distribution is transformed into an optimization problem where an optimal q is found from a set of variational distributions $Q=\{q_1, q_2, \ldots, q_m\}$ such that an objective function, i.e. Kullback-Leibler divergence, is minimized. The introduction of these efficient inference algorithms [2][3] recently has made applications such as speech recognition computationally feasible.

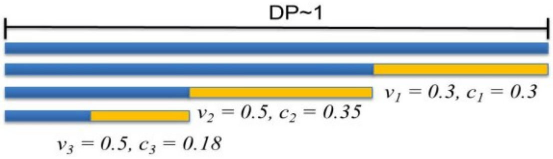

Figure 1: *A diagram of the stick breaking representation of a Dirichlet process is shown. The absolute length of each stick corresponds to a mixture component weight. These weights are constrained to sum to 1.*

1.2. English and Mandarin Speech Recognition

As of 2009 Ethnologue reported 6,909 living languages in the world and of those Mandarin and English are numbers one and three (respectively) of the most commonly spoken [8]. Moreover, these two languages come from separate families and are linguistically and phonetically very different. For these reasons English and Mandarin are selected to ensure that the performance of AVDPM, CVSB, and CDP are not heavily influenced by any language specific artifacts.

Based on NIST benchmarks Mandarin speech recognition tasks have historically yielded worse error rates than comparable English ones [9]. There are many factors that this disparity can be attributed to such as Mandarin's flexible grammatical structure, relatively high number of homophones (about 1,300 syllables compared to approximately 10,000 for English [10]) and, most conspicuously, the tonal nature of the language. Unlike English, whose phoneme labels are all unique, each vowel in Mandarin can take five different tones (4 distinct tones and 1 neutral tone). Thus, where English has approximately 40 phoneme labels, Mandarin actually has close to 90. The scope of this work is constrained to phoneme recognition so that other factors, such as language modeling, are decoupled.

2. Nonparametric Bayesian Approaches

Parameterized models have been widely applied to clustering and classification problems for their ease of use, simplicity, and reasonable performance. Unfortunately, they require making assumptions about data structure and sometimes generalize poorly. Nonparametric methods, on the other hand, do not suffer from these limitations but, due to their complex nature, require inference algorithms to make posterior calculations tractable. In this section, a brief overview of one such nonparametric method, a DPM, is provided.

2.1. Dirichlet Distributions and Dirichlet Processes

One of the main drawbacks of typical, parametric speech recognition systems is the assumption that the number of mixture components for each phoneme model is known and is held constant for every model. For complex data such as speech this is largely presumptuous and it would be more reasonable to assume that each phoneme model has its own unique structure.

Creating a model to characterize the optimal number of mixture components is best represented by a multinomial distribution. To model this in a statistically meaningful way priors are needed to ascertain information such as the number of mixture components and their respective weights. Dirichlet distributions act as the conjugate prior for the multinomial distribution, and in the case of this work, can be used to find the optimal number of mixture components. An extension of the Dirichlet distribution, the Dirichlet process (DP), is used to then generate discrete priors for modeling the respective weights of these components.

A Dirichlet distribution (DD) is often referred to as a distribution over distributions and is given by:

$$Dir(\alpha) \sim f(q;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k} q_i^{\alpha_i - 1} \qquad (1)$$

where q and α are a set of distributions and their respective concentration parameters (i.e. inverse variances) such that $q = |q_1, q_2, \ldots, q_k|$, $q_i \geq 0$, $\sum_{i=1}^{k} q_i = 1$ and, $\alpha = |\alpha_1, \alpha_2, \ldots, \alpha_k|$, $\alpha_i > 0$, and $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. Furthermore, the decimative property of DDs explains that each distribution, q_i, can be split in such a way that $(q_{11}, q_{12}, q_2, \ldots, q_k) \sim Dir(\alpha_1\beta_1, \alpha_1\beta_2, \alpha_2, \ldots, \alpha_k)$ where $q_{11}+q_{12}=q_1$ and $\beta_1+\beta_2=1$.

A DP is a DD split infinitely many times, ultimately generating discrete values that serve as priors. This can be seen in Figure 2 where a DD is initially set to a uniform distribution. After an infinite number of splits, the resulting distributions are infinitely narrow and essentially discrete values are obtained which serve as priors for the models in this work. Although there are many representations of DPs, all three algorithms used in this work focus on the stick breaking approach shown in Figure 1.

2.2. Variational Inference Algorithms

As mentioned earlier, variational inference converts the sampling problem of MCMC methods into an optimization problem. A variational distribution, $q(y)$, which has made independence assumptions about model parameters, is used to approximate the posterior, $p(y|x)$. More specifically, these algorithms assume that the distributions that represent stick lengths (and by extension, mixture component weights), component structure (i.e. means and covariances of a Gaussian for this work), and mixture assignments are all independent. This relationship can be seen in (2), (3), and (4) below. By using optimization techniques such as the EM algorithm and the Kullback-Leibler (KL) divergence as a cost function, an optimal $q(y)$ can be found from a set of distributions $Q = \{q_1, q_2, \ldots, q_k\}$. Thus, new stick breaks, i.e. mixture components, are released as the KL divergence is minimized.

Even variational inference algorithms can be computationally inefficient and often require additional constraints to make their use viable. AVDPM incorporates KD-trees which can be used during preprocessing to organize the data by partitioning them across hyperplanes in the feature vector space. Lower initial depths essentially result in shorter training times at the expense of accuracy. Moreover, AVDPM limits the number of mixture components to a truncation level, T, such that additional components, $L>T$, can exist but are tied to their priors. For AVDPM the factorized variational

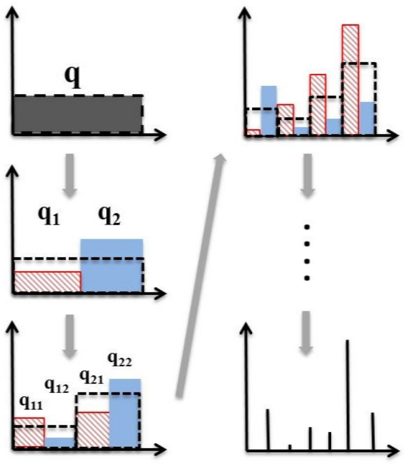

Figure 2: *A diagram showing how splitting a Dirichlet distribution infinitely many times yields discrete values.*

distribution is given by [3]

$$q(z,\eta,v) = \prod_{i=1}^{L}\left[q_{\phi_v}(v_i) q_{\phi_\eta}(\eta_i) \right] \prod_{n=1}^{N} q_{z_n}(z_n) \qquad (2)$$

where $q_\phi(v_i)$, $q_\phi(\eta_i)$, and $q_z(z_n)$ represent parametric models for stick lengths, the components' structures (e.g. μ and σ for Gaussians), and mixture component assignments respectively. Each of the parametric models' respective parameters are given by ϕ.

CVSB and CDP, on the other hand, do not incorporate KD-trees but instead use a "hard" truncation level. This essentially limits the DPM to a finite, but large number of mixture components, *T*. The variational distribution for CVSB is almost identical to that used for AVDPM [2]

$$q(z,\eta,v) = \left[\prod_{n}^{N} q(z_n) \right] \left[\prod_{t=1}^{T} q(\eta_t) q(v_t) \right] . \qquad (3)$$

While CVSB can have variable stick lengths, CDP imposes a symmetric prior on the variational distributions, i.e. the lengths of *k* stick breaks are all equal and thus weights of mixture components are all equal. This essentially reduces the DP to a DD and allows for the exchangeability of labels. The factorized variational distribution for CDP is [2]:

$$q(z,\eta,c) = \left[\prod_{n}^{N} q(z_n) \right] \left[\prod_{k=1}^{T} q(\eta_k) \right] q(c) . \qquad (4)$$

The primary difference between (3) and (4) is the replacement of *q(v)* by *q(c)*. The i^{th} stick break, v_i, represents the fraction of the remaining stick length and is modeled with a beta distribution [7] while c_i is the actual mixture weight (i.e. the fraction of the original, whole stick). Since the length of each stick break is held constant, the effect from the stick lengths can be removed from the product in (3) and replaced by *q(c)*.

3. Experimental Setup

In this work, the performance of AVDPM, CVSB, and CDP was compared to a standard Gaussian mixture model. This section outlines some of the key details used in this work.

Labels for the CH-E Corpus consisted of the 39 phonemes found in the CMU7 dictionary [14] as well as three additional labels – sp, sil, and a garbage phoneme – which were added to account for any partial words or sounds in the data. The CH-M Corpus contains 92 phoneme labels consisting of the labels found in the CH-M lexicon and the 3 additional labels used in CH-E Corpus. Furthermore, English words that exist in CH-M are added to the CH-M lexicon where any English vowel sounds are assigned to the neutral tone. The relatively high number of labels is due to the tonal nature of Mandarin which requires all vowel sounds to have 5 labels (e.g. vowel "a" is actually "a1", "a2", "a3", "a4", and "a5").

Phoneme alignments were generated by training a hidden Markov model (HMM) based acoustic model using a flat start and training up to 16 monophone mixtures. Finally, a Viterbi alignment was performed to identify phoneme segments. Any utterances from the corpora that contained simultaneous speech from multiple speakers were discarded.

Using the generated segmentations, 13 MFCC features and their first and second derivatives were extracted using a frame duration and window duration of 10 ms and 25 ms respectively. The frame-based features from each phoneme segment were averaged in a 3-4-3 manner so that the number of features per segment was constant despite duration (although duration was added as a single additional feature). Models were trained for each phoneme label and predictions were generated using maximum likelihood. Diagonal covariances were used to train the GMM models and the number of mixture components was held constant for all phoneme labels. Conversely, AVDPM, CVSB, and CDP found this number, and the corresponding means and covariances, automatically.

The best of 10 iterations of the GMM baseline was compared to the average performance of AVDPM, CVSB, and CDP over 10 iterations. Performance was evaluated using both error rates and the average number of mixture components per phoneme label.

These algorithms were initially evaluated on the well-calibrated TIMIT Corpus to confirm that this setup produced comparable performance to other published results. Following the methods in [11]-[13], the corpus was partitioned into training, validation, and evaluation sets. The 61 original phonemes that exist in TIMIT were collapsed to 39 labels. GMMs were first fit using the phoneme alignments provided with TIMIT. The number of mixture components was varied for the GMMs and an optimal performance of 31.56% misclassification error was achieved for 4 mixture components per phoneme label. This was comparable to the results found in [13] although for a much lower number of mixture components (i.e. 4 mixtures vs. 64 mixtures). This discrepancy was due to the use of features only from the central portion of each phoneme segment instead of the 3-4-3 approach used in this work [13]. With this confirmation, phoneme alignments were then generated for the collapsed 39 labels in the same manner used for CH-E and CH-M. These results are discussed in the following section and allowed for a better comparison to the performance on CH-E and CH-M.

4. Results and Discussion

The truncation level for CVSB and CDP was varied to determine an optimal operating point for each corpus. Similarly, the initial depth of the KD tree was adjusted for AVDPM to determine the effect on performance. Each algorithm was iterated ten times and an average misclassification error rate was calculated. A table of the best error rates on their respective evaluation sets are shown in Table 1, along with the associated parameter values.

It can be seen that the average misclassification error of all three variational inference algorithms yield comparable error rates and require significantly fewer mixture components than the baseline GMM model where the number of components is assumed to be known a priori. This is due to the ability of DPMs to discover the underlying structure of the data and consequently less prone to overfitting.

It is interesting to note that relative performance of CVSB and CDP was worse for TIMIT than both CH-E and CH-M. This is most likely an artifact of the studio recorded, read speech of TIMIT which allows for the fixed number of mixture components of the GMM to reasonably approximate the underlying structure of the data. Conversely, CVSB and CDP are better suited to conversational telephone speech where the underlying structure is less apparent. Finally, the relatively small disparity between Mandarin and English can easily be attributed to Mandarin having more than double the number of phoneme labels as English, i.e. each phoneme's model is trained on less than half the number of segments as those for English.

Table 1: *A comparison of best misclassification error and number of mixture components for the evaluation sets of the TIMIT, CH-E, and CH-M corpora. The parameters corresponding to these operating point are also given.*

Model	TIMIT		CH-E		CH-M	
	Error %	Notes	Error %	Notes	Error %	Notes
GMM	38.02%	# Mixt. = 8	58.41%	# Mixt. = 128	62.65%	# Mixt. = 64
AVDPM	37.14%	Init. Depth = 4	57.82%	Init. Depth = 6 Avg. # Mixt. = 5.14	63.53%	Init. Depth = 8 Avg. # Mixt. = 5.01
CVSB	40.30%	Truncation Level = 4	58.68%	Truncation Level = 6 Avg. # Mixt. = 5.89	61.18%	Truncation Level = 6 Avg. # Mixt. = 5.75
CDP	40.24%	Truncation Level = 4	57.69%	Truncation Level = 10 Avg. # Mixt. = 9.67	60.93%	Truncation Level = 6 Avg. # Mixt. = 5.75

It can be seen in Table 1 that both CH-E and CH-M have the same optimal truncation levels for CVSB and CDP with the exception of CDP on CH-E. This is not unexpected since the symmetric prior CDP imposes on the lengths of the stick breaks indicates that there should be an equal or greater number of mixture components compared to those found by CVSB to compensate for that assumption.

AVDPM's performance and average number of mixture components are comparable to both CVSB and CDP. However, the incorporation of KD trees make it more attractive for acoustic modeling since larger data sets can be managed by trading off the depth of the KD tree. The computational complexity of this algorithm grows rapidly as depth increases [3], but it can be seen in Table 1 that speech from significantly different recording environments have optimal operating points at similar initial depths of the KD tree. Although the optimal depth for CH-E and CH-M are 6 and 8, reducing the depths to 4 was found to only marginally worsen the error rates (by 1.32% and 1.14% respectively).

This is particularly interesting as Figure 3 shows the actual measured CPU times for training as a function of the amount of training data for AVDPM, CVSB, and CDP. CPU times were obtained using optimal operating points on TIMIT when the initial depth of the KD tree is set to 4 for AVDPM and the truncation level to 4 for both CDP and CVSB. These plots were extrapolated to show the theoretical training time for a much larger corpus such as Fisher [15]. We have generated these extrapolated results using simulated data since we do not have access to corpora of this size. These run-time differences held for simulated data and should not be data dependent.

It can be seen here that the required training times of CVSB and CDP grow rapidly as the number of training samples increases. Furthermore, CH-E and CH-M require higher truncation levels. As can be seen in Table 1, these algorithms generally choose the maximum number of mixture components (this is at least true for relatively low truncation levels). This indicates that the training time should increase linearly as the truncation level increases. The error rates generated by AVDPM are optimal (or very near optimal) at a low initial depth of the KD tree. The complexity of initially building the KD tree has a significant cost which accounts for the relatively large gap in training times for small amounts of training data between AVDPM and CVSB or CDP. However, it is shown in Figure 3 that the training time required by AVDPM is significantly less affected as the amount of data increases and would be almost two orders of magnitude faster when training on a large corpus such as Fisher.

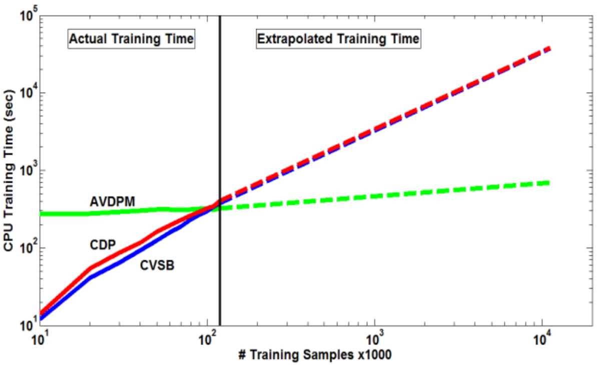

Figure 3: *A diagram showing how the CPU training time changes as the amount of training data increases.*

5. Conclusions

Dirichlet distributions, and by extension DPMs, can be used to find underlying structure of data, e.g. the number of mixture components in a GMM. For further improvements these nonparametric models can be extended to HMMs to not only find the structure of each state's distribution but to also find the structure of the HMM itself, i.e. the number of states and the transitions between them. However, due to these methods' infinite parameters variational inference algorithms are needed to make posterior calculations tractable. In this work, it is shown that three variational methods – AVDPM, CVSB, and CDP – are not subject to language specific artifacts and yield comparable performance to baseline GMMs but with significantly fewer parameters.

CVSB and CDP have optimal truncation levels between 4 and 10 for speech data and can perform well on small corpora such as TIMIT. However, AVDPM is best suited to acoustic modeling since controlling KD tree depth allows for the tradeoff between accuracy with available computational resources, thereby making training on large corpora possible. An initial depth of 4 for AVDPM yielded optimal, or very near optimal, results for data ranging from cleanly recorded read speech to noisy conversational telephone speech. Furthermore, this algorithm is significantly less affected by the amount of training data and is theoretically able to train large corpora orders of magnitude faster than CVSB or CDP.

6. Acknowledgements

This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854. Our research group would also like to thank the Linguistic Data Consortium (LDC) for awarding a data scholarship to this project and providing the lexicon and transcripts for CALLHOME Mandarin.

7. References

[1] C. Antoniak, “Mixtures of Dirichlet Process with Applications to Bayesian Nonparametric Problems,” *The Annals of Statistics*, vol. 2, no. 7, pp. 1152–1174, 1974.

[2] K. Kurihara, “Collapsed variational Dirichlet process mixture models,” in *International Joint Conference on Artificial Intelligence*, 2007, pp. 1–6.

[3] K. Kurihara, M. Welling, and N. Vlassis, “Accelerated Variational Dirichlet Process Mixtures,” in *Advances in Neural Information Processing Systems*, 1st ed., B. Scholkopf, J. Platt, and T. Hofmann, Eds. Boston, Massachusetts, USA: The MIT Press, 2007, pp. 1–8.

[4] R. Neal, “Bayesian Mixture Modeling by Monte Carlo Simulation,” 1991.

[5] J. Paisley, “Machine learning with Dirichlet and beta process priors: Theory and Applications,” Duke University, 2012.

[6] C. E. Rasmussen, “The Infinite Gaussian Mixture Model,” in *In Advances in Neural Information Processing Systems*, MIT Press, 2000, pp. 554–560.

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

[8] L. Paul, G. Simons, and C. Fennig, “Ethnologue: Languages of the World,” 2009. [Online]. Available: http://www.ethnologue.com. [Accessed: 03-Feb-2013].

[9] “The History of Automatic Speech Recognition Evaluations at NIST,” *NIST*, 2009. [Online]. Available: http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html. [Accessed: 03-Feb-2013].

[10] W. Gu, K. Hirose, and H. Fujisaki, “Comparison of Perceived Prosodic Boundaries and Global Characteristics of Voice Fundamental Frequency Contours in Mandarin Speech,” in *Chinese Spoken Language Processing*, 2006, pp. 31–42.

[11] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden Markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[12] A. Halberstadt and J. Glass, “Heterogeneous acoustic measurements for phonetic classification,” in *Proceedings of Eurospeech*, 1997, pp. 401–404.

[13] M. Ager, Z. Cvetkovic, and P. Sollich, “Robust phoneme classification: Exploiting the adaptability of acoustic waveform models,” in *EUSIPCO*, 2009.

[14] “The CMU Pronouncing Dictionary,” 2008. [Online]. Available: https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict. [Accessed: 03-Feb-2013].

[15] C. Cieri, D. Miller, and K. Walker, “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text,” in *Proceedings of the International Conference on Language Resources and Evaluation*, 2004, pp. 69–71.