IDIAP RESEARCH REPORT

Infinite Models for Speaker Clustering

Fabio Valente [a]

IDIAP–RR 06-19

June 2006

published in
ICLSP 2006

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-19

Infinite Models for Speaker Clustering

Fabio Valente

June 2006

published in
ICLSP 2006

Abstract. In this paper we propose the use of infinite models for the clustering of speakers. Speaker segmentation is obtained trough a Dirichlet Process Mixture (DPM) model which can be interpreted as a flexible model with an infinite *a priori* number of components. Learning is based on a Variational Bayesian approximation of the infinite sequence. DPM model is compared with fixed prior systems learned by ML/BIC, MAP/BIC and a Variational Bayesian method. Experiments are run on a speaker clustering task on the NIST-96 Broadcast News database.

1 Introduction

Speaker clustering is a main task in many audio processing systems. Most common approaches are based on statistical models in which data are represented by an ergodic HMM (each state represents a speaker) with emission probability modeled by GMM ([1]). The actual number of speaker is generally not known and must be estimated from data using a model complexity criterion (e.g. Bayesian Information Criterion [6] or Variational free energy [14]).

We propose here the use of a flexible model based on the Dirichlet Process Mixture (DPM). DPM can be interpreted as a bayesian model with an *a priori* infinite number of components. The learning algorithm infers the actual number of components out of the initial infinite number. In other words our prior model is an ergodic HMM with an unbounded number of states (speakers) emitting according to a GMM with an unbounded number of gaussian components.

In general speaker clustering situations, number of speakers is not known and can considerably change from file to file. Furthermore in the same file, amount of data available per speaker can be very heterogeneous (e.g. many speakers provide just few minutes of speech). In those cases we would like to represent speakers with a variable number of gaussian components proportional to the amount of available data. This issue was addressed for example in [13] where a complex model is used if enough data from a given speaker is provided and a simpler model is used if only poor amount of data is available. DPM provides an elegant framework for handling this problem because the initial number of speakers (i.e. HMM states) and gaussian components per speaker is infinite and final complexity is inferred during the training according to statistical properties of data.

Dirichlet Process Mixture Models have been introduced in the framework of non-parametric Bayesian statistics in [2] and [3] but only in recent years efficient training techniques have been proposed and they have been applied to many machine learning, language modeling and image processing problems [7]. We describe here a first investigation in the field of audio processing.

The paper is organized as follows: in sections 2 and 3 we introduce basic concepts of Dirichlet Process and Dirichlet Process Mixture, in section 4 we discuss learning algorithms, in section 5 we describe a speaker clustering model based on DPM and finally we discuss results on Broadcast News data.

2 Dirichlet process

A Dirichlet Process designated as $DP(G_0, \alpha)$ is a measure on measures, i.e. a stochastic process, and is parametrized using a probability measure G_0 known as base measure and scalar value α. The original definition proposed in [2] says that a measure G is distributed according to a Dirichlet process $DP(G_0, \alpha)$ if for all natural numbers k and measurable partition of an ensemble $\{B_1, ..., B_k\}$,

$$(G(B_1), ..., G(B_k)) \sim Dir(\alpha G_0(B_1), ..., \alpha G_0(B_k) \qquad (1)$$

where Dir designates a Dirichlet distribution. The general definition provided in equation (1) is not very self explicative and DP clustering properties for infinite series are not easily deductible from definition. We refer to the review done in [7] which gives three different representations of the DP. A DP has inherently some clustering properties. Let us consider N random variables $\{y_1, ..., y_N\}$ drawn according to $G \sim DP(G_0, \alpha)$ and let us compute the probability of sample y_{N+1} conditioned on the previous samples. It can be shown [8] that:

$$p(y_{N+1}|y_1, ..., y_N) \propto \alpha\, G_0(y_{N+1}) + \sum_{i=i}^{N} \delta_{y_i}(y_{N+1}) \qquad (2)$$

where G_0 is the base measure and $\delta_a(b)$ is a function equal to one if $a = b$ and zero elsewhere. This is the so called Polya Urn scheme [8]. Let us designate the values of variables y_n with $\{c_1, ..., c_N\}$, expression (2) can be rewritten as:

$$y_{N+1} = c_i \text{ with probability } \frac{\#(j : y_j = c_i)}{N+\alpha} \tag{3}$$

$$y_{N+1} = c_{new}, c_{new} \sim G_0 \text{ with probability } \frac{\alpha}{N+\alpha} \tag{4}$$

where $\#(j : y_j = c_i)$ is the number of time y_j is equal to c_i. From equation (3) we notice that the probability of the observation $y_{N+1} = c_i$ is proportional to the number of times the value c_i was seen before i.e., the probability of seeing a value we have seen before is higher if this value appeared already many times. Furthermore there is always a probability $\alpha/(N+\alpha)$ to explore new values that have not been seen before (equation 4). In a speaker clustering task this corresponds to adding a new speaker to the model or a new gaussian component to a speaker model. The capacity of generating new values is regulated by the value of α; if α is comparable to N there is a high probability of exploring new speaker (or speaker components) rather than clustering in previous ones.

According to expression (2) a DP can be interpreted as a mixture model with N fixed classes and one component responsible for creating new classes. In other words the model automatically adjusts the number of classes increasing progressively the number of components.

Another important representation of the Dirichlet Process is the *Stick breaking* construction ([10]). Let us assume two sequences of independent random variables v_i and y_i generated as:

$$(v_i)_{i=1}^{\infty} \sim Beta(1, \alpha) \quad (y_i)_{i=1}^{\infty} \sim G_0 \tag{5}$$

where $Beta(.)$ designates a Beta distribution. The Dirichlet Process $G \sim DP(\alpha, G_0)$ can be rewritten as:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{y_i} \quad \text{with} \quad \pi_i = v_i \prod_{j=1}^{i-1} (1 - v_j) \tag{6}$$

Representation (6) is called stick breaking representation (designated as $Stick(\alpha)$) . It is easy to verify that $\sum_{k=1}^{\infty} \pi_k = 1$ like in mixture models. Here an important inconvenient of G can be noticed: a measure drawn from a DP is discrete with probability one even if its base measure is continuous. In fact G is composed of an infinite sum of $\delta_a(b)$ functions which are equal to 1 only if $a = b$; in other words the support of G is discrete. We explain in section 3 how to overcome this problem.

Another useful representation is the infinite limit of finite mixture models. We will use it in the following for deriving the speaker clustering model. Let us consider a mixture model with L components i.e. $G = \sum_{i=1}^{L} \pi_i p_i$ where $\pi = \{\pi_1, ..., \pi_L\}$ are mixing proportions and p_i is a base measure (e.g. a gaussian distribution) . Let us set a symmetric prior over π as a Dirichlet distribution with hyperparameters $\{\alpha_0/L, ..., \alpha_0/L\}$. If the limit $L \to \infty$ is considered then the Dirichlet prior reduces to a Stick distribution $Stick(\alpha)$ (see [11] for details) and model $G = \sum_{i=1}^{L} \pi_i p_i$ coincides with model in equation (6). In other words in the limit case of an infinite number of mixture components, the model behaves like a DP.

3 Dirichlet process mixture

The discreteness of DP is a serious drawback if the model must handle continuous variables. It can be shown that a DP is discrete with probability one on the set of Borel probability measures even if G_0 is continuous [11]. A simple way for overcoming this problem is using the DP as non-parametric prior distribution in a hierarchical bayesian framework [3]. This is achieved by drawing a measure $G \sim DP(\alpha, G_0)$ and assuming that G is a prior distribution for model parameters θ_n i.e., $\theta_n \sim G$. In other words we add a level of hierarchy and we assume that model parameters have a distribution that follows a DP obtaining a continuous distribution. In mathematical terms, a data set Y_n is modeled as a DPM if :

$$\begin{aligned} G|\alpha_0, G_0 &\sim DP(\alpha, G_0) && (7)\\ \theta_n|G &\sim G && (8)\\ Y_n|\theta_n &\sim p(Y_n|\theta_n) && (9)\end{aligned}$$

If the distribution $p(.)$ is continuous, the model consists in a convolution of a degenerate density G with a continuous function $p(.)$ that is a continuous distribution. This model is referred as Dirichlet Process Mixture (DPM) [7]. As example, we can rewrite a common Gaussian Mixture Model with an infinite number of components in the same formalism of equations (7-9) using the Stick breaking representation of equation (5). Let us introduce an hidden variable X_n that designates which component emitted observation Y_n. Data follows the process:

$$\begin{aligned} V_i|\alpha &\sim Beta(1,\alpha),\ \ i = \{1,2,...,\infty\} && (10)\\ \theta_i|G_0 &\sim G_0\ \ i = \{1,2,...,\infty\} && (11)\\ & \quad \text{For the nth observation } Y_n && \\ X_n|\{V_1, V_2, ...\} &\sim Mult(\pi(V)) && (12)\\ Y_n|X_n &\sim p(y_n|\theta_{X_n}) && (13)\end{aligned}$$

where $Mult(\pi(V))$ designates a multinomial distribution with parameters $\pi(V(i))$ defined as in equation 6 and V_i can be interpretated as an hidden variable.

An infinite sequence of parameters θ_i is drawn from the base measure G_0 (expression (11)) together with probabilities V_i (expression (10)). For each observation Y_n an hidden variable X_n is drawn from a multinomial distribution defined by the set of V_i (expression (12)). Finally the parametric likelihood $p(y_n|\theta_{x_n})$ is computed according to the parameter θ_{x_n} (expression (13)). In case of a GMM, distribution $p(.)$ is a gaussian distribution with parameters $\theta_i = \{\mu_i, \Sigma_i\}$ where μ_i and Σ_i are mean vector and covariance matrix. In bayesian terms this model can be interpreted as a model with an a priori distribution composed of an infinite number of components. Model grows according to statistical properties and amount of data.

Likelihood of an observation Y_n conditioned on X_n can be written as:

$$p(Y_n|X_n, \{\theta_i\}) = \prod_{i=1}^{\infty} p(Y_n|\theta_{X_n})^{1[X_n=i]} \tag{14}$$

where i represent the component number (out of the possible infinite components) and the function $1[a = b]$ is equal to 1 if $a = b$ and zero otherwise.

4 Inference in a DPM

Even if DPM defines an infinite prior model, processing of finite amount of data will produce a finite posterior model. In fact, if N is the data set size, posterior model will have a maximum number of components equal to N, i.e. one component per sample. The clustering algorithm should learn posterior probability over DPM together with model complexity i.e. the number of components.

Monte Carlo Markov Chains sampling methods are probably the most popular method for making inference in models based on DP and DPM. Anyway sampling methods are generally slow and prohibitive when the amount of data is large like in applications that involve the processing of many hours of speech . For this reason we consider here a deterministic approximation based on a Variational Bayesian method as proposed in [12].

Variational Bayesian (VB) methods are suitable in those cases in which the complexity of the model must be determined (e.g. the number of speaker in a file and the number of components per speaker in the model) (see [16]). The well known Bayesian Information Criterion is a special case of VB model selection. In our previous work ([14]) we investigated the use of VB learning in the case of a model with fixed a priori number of components. The use of DPM as non-parametric prior extends somehow the flexibility of the model that use infinite prior.

Variational Bayesian methods approximate real posterior distributions over parameters θ and hidden variables X with a distribution $q(X, \theta)$(see [12]). A simplified form for $q(X, \theta)$ is chosen on the basis of a mean-field approximation that considers independence assumption between elements of X and θ i.e.

$$q(X, \theta) = \prod_{i=1}^{I} q(X_i) \prod_{j=1}^{J} q(\theta_j) \tag{15}$$

where I is the number of hidden variables and J is the number of parameters. If the parametric form for $q(.)$ belongs to the conjugate-exponential family, a coordinate ascent algorithm can be derived for iteratively optimizing $q(X, \theta)$ (for details see [12]). Actually, the condition on the form of $q(.)$ is not particularly restrictive, in fact a large variety of models satisfy this condition (e.g. HMM, mixture models, state space models, etc.).

For example, in model described by equations (10-13) posterior distributions over parameters θ and hidden variables X and V will factorize as:

$$q(V, X, \theta) = \prod_{i=1}^{\infty} q(V_i) \prod_{n=1}^{N} q(X_n) \prod_{n=0}^{N} q(\theta_n) \tag{16}$$

where $q(V_i)$ is a Beta distribution, $q(X_n)$ is a multinomial distribution and $q(\theta_n)$ is a function of the exponential family. In expression (16) the product $\prod_{i=0}^{\infty} q(V_i)$ has an infinite number of terms. To handle this infinite series we choose the approximation proposed in [12] in which the infinite posterior is truncated up to a certain value T i.e., the posterior distribution $q(V_i) = 0$ for $i > T$. Some important remarks must be done on this approximation. First of all, the prior distribution is still infinite, only the posterior distribution is truncated i.e. the model can grow up only to number of components equal to T. We are *not* imposing the number of components of the model but just its maximum number. This choice is not that difficult if the amount of training observations N is known: the model cannot have a number of components larger than N. A reasonable choice for T is N, but all values $T > N$ will result in a maximum of N components.

It is interesting to investigate what happens when the number of components is exactly equal to N; if we suppose the base measure G_0 to be gaussian, the model results in the sum of N gaussians; this is equivalent to the Parzen window estimation method which consists of estimating probability of unseen data with a sum of gaussian kernels equal in number to the training data.

5 Speaker clustering based on DPM

In this section we present a DPM based model for speaker clustering purposes as limit case of a finite model. The most popular approach for speaker clustering uses an ergodic HMM with emission probability modeled by a GMM. As long as the current speaker number is not known, it must be estimated from data using a model selection criterion (e.g. BIC, Bayesian integral, etc.). This is sometimes achieved introducing a large number of initial states (i.e. speakers) and merging them successively. The approach we use here is completely different. The prior model is infinite i.e. an infinite number of state (speakers) that emit according to a mixture model with an infinite number or components. Without loss of generality let us define a finite model for the ergodic HMM/GMM as :

$$P(O) = \prod_{i=1}^{S} a_i \prod_{j=1}^{M} c_{ij} P(O|\mu_{ij}, \Sigma_{ij}) \tag{17}$$

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
(a) ML/BIC	10	0.80	0.86	0.83	9	0.72	0.77	0.74
(b) MAP/BIC	10	0.68	0.71	0.87	9	0.70	0.78	0.74
(c) VB	12	0.85	0.89	0.87	14	0.84	0.81	0.82
(d) DPM	18	0.87	0.91	0.89	14	0.87	0.92	0.89
File	File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K
(a) ML/BIC	15	0.77	0.83	0.80	12	0.63	0.80	0.71
(b) MAP/BIC	15	0.76	0.83	0.80	21	0.75	0.64	0.69
(c) VB	14	0.75	0.90	0.82	13	0.63	0.80	0.71
(d) DPM	16	0.74	0.91	0.82	19	0.63	0.85	0.76

Table 1: Results on NIST 1996 HUB-4 evaluation test for speaker clustering

where O is an observation a_i is the weight for speaker i, c_{ij} is the weight for gaussian component j in speaker i, μ_{ij} and Σ_{ij} are mean and covariance matrix. S and M are fixed number of speakers and number of gaussian mixture components per speaker. We assume here that the transition from one speaker to another is not regulated by a Markov process; this reduce the HMM to a mixture model and allows the model written as in equation (17). Let us now impose a prior distribution over parameters in (17). We use Dirichlet distributions for a_i and c_{ij} and Normal-Wishart distributions for joint distribution of μ_{ij} and Σ_{ij} i.e.

$$\begin{aligned} P(a_j) = Dir(\lambda_{a\,0}/S) \;\; P(c_{ij}) = Dir(\lambda_{c\,0}/M) \\ P(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0\Gamma_{ij}) \;\; P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \end{aligned} \tag{18}$$

where $Dir()$, $N()$, $W()$ are respectively Dirichlet, Normal, Wishart distributions and $\{\lambda_{a\,0}, \lambda_{c\,0}, \rho_0, \xi_0, \nu_0, \Phi_0\}$ are hyperparameters as in [17]. The correspondent DPM model with infinite number of components is obtained taking the limit $S \to \infty$ and $M \to \infty$. In the limit case Dirichlet distributions $Dir(\lambda_{a\,0}/S)$ and $Dir(\lambda_{c\,0}/M)$ become Stick breaking distributions $Stick(\lambda_{a\,0})$ and $Stick(\lambda_{c\,0})$ over an infinite number of speakers and components per speaker. Learning can be done applying the variational algorithm for truncated posterior distributions briefly described in section 4 (for details see [12]).

6 Experiments

We compare the infinite DPM model with three different systems: a classical ML/BIC system (Maximum Likelihood for the training, BIC for the model selection system) referred as System I, a MAP/BIC system (Maximum a Posteriori for the training, BIC for the model selection system) referred as System II, a Variational Bayesian system which simultaneously performs the training and the model selection (see [14] for a description) referred as system III. We run experiments on the evaluation data set NIST-1996 HUB-4. It consists of 4 recordings of half an hour long in which speech and non-speech events occur together (music, noise, etc.). All files are processed in order to obtain 12 LPCC coefficients. In those files amount of speech provided by different speakers is very heterogeneous making unsupervised clustering difficult; here comes the need for a flexible model.

The training procedure uses the following algorithm: the system is initialized with a large number of speakers $M_{initial}$ then optimal parameters are learned using criteria VB, ML and MAP. Initial speaker number is then reduced progressively from $M_{initial}$ to 1 and parameter learning is done for each intermediate number of speakers. Optimal number of speakers is estimated scoring the different models with VB free energy for system III and with BIC criterion for systems I and II. On the other hand in the DPM based system we just have to provide the truncation order T for the posterior distribution: the model will grow automatically up to the maximum number of components imposed by

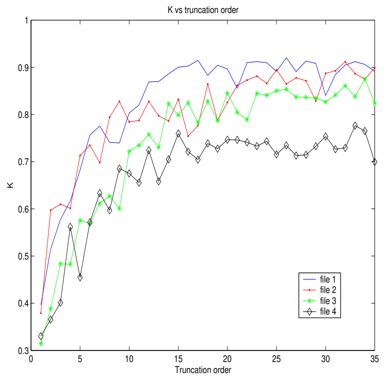

Figure 1: Speaker clustering K function of the truncation order.

the order of the truncation. The DPM is learned as well with a Variational Bayesian approximation as system III. However system III has a fixed dimension prior model while DPM has an infinite distribution as prior model. Details about estimation formula for the ML and VB learning applied to model (17) can be found in [15]. Results are shown in table 1 provided in terms of estimated number of speaker N_c, average cluster purity (acp), average speaker purity (asp) and $K = \sqrt{acp \cdot asp}$ (for details see [15]). Choice of prior distributions is done heuristically. In order to make a fair comparison, we initialized the three Bayesian approaches (MAP,VB and DPM) with the same prior distributions over parameters (this does not mean that the number of prior components is the same, in fact in the DPM case it is infinite). The Bayesian Information Criterion need a tuning factor λ in order to be really effective; we set this tuning factor in order to obtain the best possible performance in order to compare with the best possible BIC system.

The ML/BIC baseline is poor compared to other bayesian approaches. This is probably due to the regularization effect of the prior distribution. On the other hand DPM is the bayesian approach that performs better both in terms of acp and asp. Considering that MAP is a special case of VB (see [14]) and that DPM is an extension of VB to a more flexible model with an infinite number of prior distributions, results are not surprising. On file 3 all the systems perform almost the same while the largest improvements are obtained on files 2 (15% relative) and 4 (17% relative). File 4 is the file with the lowest amount of data per speaker (22 speakers for half an hour) and very heterogeneous distribution (some speakers talk just few utterances); DPM system provides the best performance probably because there is no priori information on the number of components per speaker which are automatically inferred by the system allowing more model flexibility. We verify in the DPM system that number of gaussian component is proportional to the amount of data provided per speaker.

The robustness of the system to the level of truncation is investiagated as well; figure 1 plots speaker clustering score K function of the truncation level T in the 4 files. If the truncation level is large enough (more than 25), the clustering score does not change significantly. Small fluctuation are seen due to different local minima. This means that in real data problems with finite amount of data the truncation algorithm can be efficiently applied.

7 Conclusion and Discussions

In this paper we have presented and discussed a first system for speaker clustering based on a Mixture of Dirichlet Process as flexible model with an unbounded number of prior components. Theoretical bases of DPM model were presented and discussed. Experiments on Broadcast news data show interesting improvements . In future work we would like to consider different learning approaches for the DPM models like Expectation-Propagation and comparing results with the current variational truncated method. Robustness with respect to prior distribution must be addressed as well.

References

[1] Olsen J. O., “Separation of speaker in audio data”, EUROSPEECH 1995, pp. 355-358.

[2] Ferguson T., “A Bayesian analysis of some nonparametric problems”, The Annals of Statistic 1:209-230.

[3] Antoniak C., ”Mixtures of Dirichlet process with applications to Bayesian nonparametric problems. The Annals of Statistic 2(6): 1152-1174.

[4] Attias, H., “A Variational Bayesian framework for graphical models”, Adv. in Neural Inf. Proc. Systems 12, MIT Press,Cambridge, 2000.

[5] Watanabe S. et al. “Application of the Variational Bayesian approach to speech recognition” NIPS’02. MIT Press.

[6] , Chen S. and Gopalakrishnan P., ”Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion”,Proceedings of the DARPA Workshop, 1998,

[7] Teh Y.W., Jordan M.I., Beal M.J., Blei D.M., “Hierarchical Dirichlet process”, Technical Report 653, UC Berkeley Statistics, 2004.

[8] Blackwell D. and MacQueen “Ferguson Distributions via Polya Urn Schemes”, Annals of Statistic, 1, pp.353-355,1973.

[9] Solomonoff A., Mielke A., Schmidt, Gish H.,” Clustering speakers by their voices”, ICASSP 98, pp. 557-560

[10] Sethuraman J., “A constructive Definition of Dirichlet Priors”, Statistica Sinica, 4, pp. 639-650

[11] Ishwaran J., and James L. “ Gibbs sampling methods for stick-breaking priors”, Journal of the American Statistical Associations 96: 161-174.

[12] Blei D.M. , ”Variational Inference for Dirichlet process mixtures”. Bayesian Analysis, 1, 2004

[13] Nishida M. et Kawahara T. “Unsupervised speaker indexing using speaker model selection based on bayesian information criterion” Proc. ICASSP 2003

[14] Valente F., Wellekens C. “Variational Inference for Dirichlet Process mixtures”, Proc. of ICSLP, 2004

[15] Valente F., Wellekens C. “Variational Bayesian Speaker Clustering”, Proc. Odyssey 2004

[16] MacKay D. J. C., ”Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks”, Network:Comput. Neural Syst. 6, 469–505, 1995

[17] Gauvain J.L. and Lee C.H. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions SAP, 2:291–298, 1994.