Variational Bayesian methods for audio indexing

Fabio Valente and Christian Wellekens

Institut Eurecom
Sophia Antipolis, France
{fabio.valente,christian.wellekens}@eurecom.fr

Abstract. In this paper we aim to investigate the use of Variational Bayesian methods for audio indexing purposes. Variational Bayesian (VB) techniques are approximated techniques for fully Bayesian learning. Contrarily to non Bayesian methods (e.g. Maximum Likelihood) or partially Bayesian criterion (e.g. Maximum a Posteriori), VB benefits from important model selection properties. VB learning is based on the Free Energy optimization; Free Energy can be used at the same time as an objective function and as a model selection criterion allowing simultaneous model learning/model selection. Here we explore the use of VB learning and VB model selection in a speaker clustering task comparing results with classical learning techniques (ML and MAP) and classical model selection criteria (BIC). Experiments are run on the evaluation data set NIST-1996 HUB-4 and results show that VB can outperform classical methods.

1 Introduction

Model selection is a main issue in many machine learning problems. In different real data applications an hypothesis on the model is done before proceeding with the learning task. If the hypothesized model does not respect the structure of experimental data, the effectiveness of the learning is strongly affected. Here the need comes for techniques that can select the model that best fit to data.

The probabilistic framework is largely used for model selection. It considers probabilities over different models and assumes that the best model is the one that maximizes model probability given the observed data i.e. given a model m and an observation data set D , best model maximizes $P(m|D)$. Depending on the model complexity, $P(m|D)$ cannot always be obtained in close form and approximated techniques must be considered instead. The most used approximations (e.g. BIC [1]) are sometimes inappropriate according to the considered application and need heuristic tuning to be effective. In this paper we discuss a new type of approximated method called Variational Learning (a.k.a. Ensemble Learning) that allows an approximated close form solution to the model selection problem. The key of Variational methods is the replacement of real unknown parameter distributions with approximated distributions (Variational distributions) that permit an analytical tractability of the solution. Obviously the effectiveness of this approach depends on how close the approximated distributions are to real distributions.

We investigate here the use of Variational techniques in a speaker clustering task. This task often represents the first processing step in many audio indexing and speech recognition systems. Speaker clustering is formulated as a model selection problem in which the speaker number must be estimated. The most popular solution uses the BIC (see [2],[3]) that is actually true only asymptotically. In order to obtain reasonable results in the limited data case, an heuristic adjustment of the model selection criterion is done. It often gives serious tuning problems and final result is strongly affected. Variational methods are a finer approximation of the Bayesian integral and result more effective than BIC in many model selection tasks. Furthermore they allows simultaneous model learning and model selection in a fully Bayesian fashion.

The paper is organized as follows: in section 2 we discuss model selection problems, in section 3 we present the Variational Bayesian framework, in section 4 we present the speaker clustering model and experiments that compare VB and MAP/BIC, ML/BIC systems.

2 Model selection

Let us consider a data set D and model set $Model = \{m\}$. In a probabilistic framework, model that fits data in the best way is the model that maximizes $P(m|D)$ i.e. the model probability given the data. Applying Bayes rule, we obtain:

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)} \tag{1}$$

where $P(D) = \sum_m P(D|m)P(m)$ does not depend on m. If prior probability over model $P(m)$ is uniform (i.e. no prior information over the model is available), the best model is the model that maximizes data evidence i.e. $P(D|m)$. Let us designate with θ the model parameter set and with $p(\theta|m)$ parameter set distribution. The data evidence can be obtained marginalizing model parameters w.r.t. their distributions i.e.

$$p(D|m) = \int p(D|\theta, m)\, p(\theta|m) d\theta \tag{2}$$

Expression (2) is known as marginal-likelihood.

Marginal likelihood has the interesting property of penalizing models that have too many degree of freedom not necessary for modeling experimental data. This is also known as the Occam razor property (e.g. see [4]). The idea is that models with more free parameters can model a larger data set, resulting in a parameter probability $p(\theta|m)$ more spread over the parameter domain.

Unfortunately for many currently used models like Hidden Markov Models (HMM) or Gaussian Mixture Models (GMM), no close form solution is possible for marginal likelihood because of hidden variables. A common choice for overcoming this problem consists in simply ignoring the integral in (2); in this way the classical Maximum a Posteriori parameter estimation can be recovered i.e.

$$\theta_{MAP} = argmax_\theta\, p(D|\theta, m)p(\theta|m) \tag{3}$$

The MAP approach is tractable but not fully Bayesian because it is a point estimation that just considers parameters instead of distributions over parameters. Using a metaphor coming from physics, MAP just considers the 'density' instead of the 'mass' of the distribution. MAP becomes a reasonable approximation when parameter distribution is extremely peaked and the 'mass' of the distribution is concentrated around the maximum but in general cases it can neglect important contributions to the integral. On the other hand MAP criterion has no model selection properties and approximation of the integral mass must be considered. The most popular approximation technique is the Bayesian Information Criterion (BIC). BIC was first derived by Schwartz in [1]. It can be obtained from a Laplace approximation of the Bayesian integral (2). The Laplace approximation makes a local Gaussian approximation around the MAP parameter estimate $\hat{\theta}$ and is based on large data limit. Let us suppose that the cardinality of the data set D is N, and that the number of free parameters θ is p, in this case the BIC is:

$$log\, p(D|m)_{BIC} = log\, p(D|\hat{\theta}, m) - \frac{p}{2} ln\, N \tag{4}$$

Expression (4) has an intuitive explanation: a more complicated model i.e. a model with many free parameters p will result in a larger penalty term $\frac{p}{2} log\, N$ respect to a model with a smaller number of free parameters. BIC is a very rude approximation of the Bayesian integral but presents many tractability advantages because it can be easily computed as long as a MAP estimation of model parameter is available. As previously outlined BIC is based on a large data limit that is rarely meet in real data problems. To overcome this limitation and to make the criterion more effective in different situations, the penalty term is generally multiplied by a threshold λ that is heuristically determined depending on the application. For example in audio indexing problems, a huge gain in the model selection task is obtained manually modifying the penalty term (see [3]) or using some validation data to find the optimal λ for a given data set.

3 Variational Learning

Variational learning is a relatively new technique based on the use of approximated distributions instead of real distributions in order to obtain a tractable learning task. Variational methods assume that the unknown posterior distribution over parameter $p(\theta|D, m)$ can be approximated by another distribution $q(\theta|D, m)$ that is actually the variational posterior distribution (or simply variational distribution) derived from data. Considering Jensen inequality it is possible to write:

$$log\, p(D|m) = log \int d\theta q(\theta|D, m) \frac{p(\theta|m) p(D, \theta|m)}{q(\theta|D, m)} \geq \int d\theta q(\theta|D, m) log \frac{p(D, \theta|m)}{q(\theta|D, m)} = F(\theta) \tag{5}$$

$F(\theta)$ is called variational free energy or ensemble learning energy and is a lower bound on the marginal log likelihood; variational learning aims to optimize the free energy w.r.t. variational distributions instead of the intractable marginal log likelihood $log\, p(D|m)$. One of the key point in this framework is the choice of the form for distribution $q(\theta|D,m)$ that must be close enough to the real unknown parameter distribution $p(\theta|D,m)$ and still of a tractable form. The difference between marginal log-likelihood and free energy is:

$$log\, p(D|m) - F(\theta) = KL(q(\theta|D,m)||p(\theta|D,m)) = -\int q(\theta|D,m) log \frac{p(\theta|D,m)}{q(\theta|D,m)} \tag{6}$$

Equation (6) means that variational learning actually minimizes the distance between the true posterior distribution and the variational posterior distributions. In the limit case, if $q(\theta|D,m) = p(\theta|D,m)$ the free energy is equal to the log-marginal likelihood.

3.1 Learning with hidden variables

A very appealing property of variational learning is its capacity of handling hidden variables. In fact hidden variables can be simply seen as stochastic variables (as parameters) with their own distributions. In some variational learning systems there is no difference between the way parameters and hidden variables are considered (e.g. see [6]). Let us define X the hidden variable set, it is possible to introduce a joint variational distribution over hidden variables and parameters $q(X,\theta|D,m)$ and applying again Jensen inequality we obtain:

$$log\, p(D|m) = \int p(D,X,\theta|m)\, d\theta\, dX \geq \int q(X,\theta|D,m) log\, \frac{p(D,X,\theta|m)}{q(X,\theta|D,m)} = F(\theta, X) \tag{7}$$

At this point another further approximation must be done in order to obtain a tractable form: in fact considering the joint variational distribution $q(\theta,X|D,m)$ of parameters and hidden variables can be a prohibitive task when the number of hidden variables is large. For this reason the independence between hidden variables and parameters is assumed i.e. $q(\theta,X|D,m) = q(\theta|D,m)q(X|D,m)$. Under this hypothesis, optimal variational posterior distributions that maximizes the free energy can be found using an EM-like algorithm (see [8]) also known as VBEM algorithm. In fact simply deriving the free energy w.r.t. $q(\theta|D,m)$ and $q(X|D,m)$, it is possible to obtain an iterative update equation system that will converge in a local maximum of the free energy. The equation system consists of an E-like step :

$$q(X|D,m) \propto e^{<log\, p(D,X|\theta,m)>_{\theta}} \tag{8}$$

and an M-like step:

$$q(\theta|D,m) \propto e^{<log\, p(D,X|\theta,m)>_{X}} p(\theta|m) \tag{9}$$

where $< . >_z$ means average w.r.t. z. This EM-like algorithm does not estimate any parameter (contrarily to MAP) but just parameter distributions. Under the factorization hypothesis it is possible to rewrite the free energy as follows:

$$F(\theta, X) = \int d\theta dX q(X|D,m)q(\theta|D,m) log[\frac{p(D,X,\theta|m)}{q(X|D,m)q(\theta|D,m)}]$$
$$= < log \frac{p(D,X|\theta,m)}{q(X|D,m)} >_{X,\theta} - KL[q(\theta|D,m)||p(\theta|m)] \quad (10)$$

The free energy can be seen as composed of two terms: a first term depending on data and variational distributions (over both parameters and hidden variables) and a second term that is the KL divergence between the variational distribution over parameters $q(\theta|D,m)$ and the prior distribution over parameters $p(\theta|m)$. By definition we have $KL[q(\theta|Y,m)||p(\theta|m)] \geq 0$ with equality when $q(\theta|Y,m) = p(\theta|m)$; this term acts like a sort of penalty term that penalizes models with more parameters. In fact models with many parameters will result in a sum of KL divergence term for each parameter. It is very interesting to notice that VB does not simply consider the number of parameters (like in the BIC) but it explicitly considers the divergence between posterior distributions and prior distributions. It can be shown that in large data limit this penalty term converges to the BIC penalty term (see [9]). Intuitively free energy is an interesting quantity for doing model selection as long as it approximates the Bayesian integral, we will consider a more rigorous framework in section 3.2.

Now that an efficient solution for handling hidden variables has been introduced, fully Bayesian learning is possible in many models previously intractable . For example variational learning in Hidden Markov Models is first introduced in [5] and in Gaussian Mixture Models is first introduced in [8]. The applicability of Variational Bayesian EM (VBEM) algorithm to a general model is studied in [7]. VBEM algorithm can be derived for conjugate-exponential models i.e. models that meet the following two conditions: 1) The complete data likelihood is in the exponential family; 2) The parameter prior is conjugate to the complete data likelihood. Many well known models satisfy those two conditions: Gaussian mixture models, Hidden Markov models, Factor Analyzer, Principal Component Analysis, etc.

3.2 Model selection using Free Energy

In this section we define a more rigorous framework for model selection using free energy. Let us consider the log marginal likelihood obtained integrating over all possible models $p(D) = \sum_m p(D|m)p(m)$ and let us introduce a variational posterior probability over models $q(m)$. Applying one more time Jensen inequality, it is possible to have:

$$log\, p(D) = log[\sum_m p(D|m)p(m)] \geq \sum_m q(m)[F_m + log \frac{p(m)}{q(m)}] \quad (11)$$

where $p(m)$ is a prior probability over the model and F_m is the free energy for model m. Again a bound on the log marginalized likelihood is derived. Deriving w.r.t. $q(m)$ and solving we obtain for the optimal variational distribution over models:

$$q(m) \propto exp\{F_m\}p(m) \tag{12}$$

that means that optimal posterior over model is proportional to the exponential of the free energy times the prior probability. If prior probability over models is uniform, $q(m)$ will depend on free energy only. It means that the free energy can be used for doing model selection instead of the real log marginal likelihood.

4 Variational Bayesian Speaker Clustering

The most popular approach to speaker clustering task consists in the use of an ergodic HMM [10] in which each state represents a speaker. Our system is based as well on a fully connected HMM with emission probabilities modeled by GMMs. In order to obtain a non-spare solution a duration constraint of D frames on the emission probabilities is imposed. Furthermore we assume that the probability of transition to state j is the same regardless the initial state i.e. $\alpha_{r\,j} = \alpha_{r'\,j}\ \forall r, r^{'}$, where $j = \{1, ..., S\}$ with S the total number of states. To summarize let us designate $[O_1, ..., O_T]$ a sequence of T blocks of D consecutive frames $[O_{t1}, ..., O_{tD}]$ where D is the duration constraint. It is then possible to write the log-likelihood :

$$log\, P(O|\theta, m) = \sum_{t=1}^{T} log\, [\sum_{j=1}^{S} \alpha_j \{\prod_{p=1}^{D} \sum_{i=1}^{M} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})\}] \tag{13}$$

where S represents the number of states (that represent speakers), M Gaussian component that models each speaker, and $\theta = \{\beta_{ij}, \mu_{ij}, \Gamma_{ij}\}$ represents mixture model parameters (weights, means and Gaussians). If the state number S is known, model (13) can be learned using the Expectation-Maximization algorithm for both MAP and ML criteria.

As long as the number of speakers (i.e. states) is unknown, it must be estimated using a model selection criterion. Generally the BIC criterion is used. We consider here the Variational Bayesian framework for both model learning and model selection at the same time.

In VB methods prior distributions over parameters must be chosen; according to the discussion of section 3 we set those distributions as belonging to the conjugate exponential family. Let us consider now model in expression (13) and let us define following probability distributions over parameters:

$$\begin{gathered} P(\alpha_j) = Dir(\lambda_{\alpha\, 0})\ \ P(\beta_{ij}) = Dir(\lambda_{\beta\, 0}) \\ P(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0 \Gamma_{ij})\ \ P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \end{gathered} \tag{14}$$

where $Dir()$, $N()$, $W()$ are respectively Dirichlet, Normal, Wishart distributions and $\{\lambda_{\alpha\, 0}, \lambda_{\beta\, 0}, \rho_0, \xi_0, \nu_0, \Phi_0\}$ are hyperparameters. We assume here a fully tied prior distributions i.e. $\lambda_{\alpha\, 0} = \lambda_{\beta\, 0} = \xi_0 = \nu_0 = \tau$, $\Phi_0 = \tau \times I$ where I is an identity matrix and $\rho = \bar{y}$ where $\bar{y}$ is the average of all file observations. Parameter τ is also known as relevance factor.

Model (13) with prior distribution (14) belongs to the conjugate-exponential family so the EM-like algorithm can be applied. Variational posterior distributions have the same form of prior distributions with updated hyperparameters i.e.:

$$\begin{aligned} P(\alpha_j) = Dir(\lambda_{\alpha\, j}) \;\; P(\beta_{ij}) = Dir(\lambda_{\beta\, ij}) \\ P(\mu_{ij}|\Gamma_{ij}) = N(\rho_{ij}, \xi_0 \Gamma_{ij}) \;\; P(\Gamma_{ij}) = W(\nu_{ij}, \Phi_{ij}) \end{aligned} \tag{15}$$

Once variational posterior are estimated, a close form for the free energy can be obtained and used for model selection purposes. The EM-like algorithm for model (13) is derived in the appendix.

4.1 Experiments

In this section we compare experimentally VB system, ML/BIC system and MAP/BIC system in a speaker clustering task on the evaluation data set NIST-1996 HUB-4. Acoustic features consist of 12 MFCC coefficients. The training procedure uses the following algorithm: the system is initialized with a large speaker number $M_{initial}$ then optimal parameters are learned using three criteria (VB,ML,MAP). Initial speaker number is then reduced progressively from $M_{initial}$ to 1 and parameter learning is done for each new initial speaker number. Optimal speaker number is estimated scoring different models with VB free energy (that was used as objective function in the training step) and with the BIC (for MAP and VB). It is important to outline that when $M_{initial}$ is big VB prunes models to a smaller number of final speaker.

In order to evaluate the quality of clustering we use concepts of cluster purity and speaker purity introduced respectively in [11] and [12]. We consider in all our tests an additional cluster for non-speech events. Using the same notation of [12], let us define:

- R: number of speakers
- S: number of clusters
- n_{ij}: total number of frames in cluster i spoken by speaker j
- $n_{j.}$: total number of frames spoken by speaker j, $j = 0$ means non-speech frames
- $n_{.i}$: total number of frames in cluster i
- N: total number of frames in the file
- N_s: total number of speech frames

It is now possible to define the cluster purity p_i and the speaker purity q_j:

$$p_i = \sum_{j=0}^{R} \frac{n_{ij}^2}{n_{.i}^2} \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_{j.}^2} \tag{16}$$

Definitions of *acp (average cluster purity)* and *asp (average speaker purity)* follow:

$$acp = \frac{1}{N} \sum_{i=0}^{S} p_i \, n_{.i} \quad asp = \frac{1}{N_s} \sum_{j=1}^{R} q_j \, n_{.j} \tag{17}$$

In order to define a criterion that takes care of both *asp* and *acp*, the geometrical mean is used:

$$K = \sqrt{asp \cdot acp} \tag{18}$$

We present three different scores for each system: a score obtained initializing the system with the real speaker number obtained from labels (designed as (known)), the best score obtained (designed as (best))and the score of the selected model given by BIC or VB score (designed as (selected)).

File	File 1				File 2				File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K
(a) ML (known)	8	0.60	0.84	0.71	14	0.76	0.67	0.72	16	0.75	0.74	0.75	21	0.72	0.65	0.68
(b) ML (best)	10	0.80	0.86	0.83	9	0.72	0.77	0.74	15	0.77	0.83	0.80	12	0.63	0.80	0.71
(c) ML (selected)	13	0.80	0.86	0.83	16	0.84	0.63	0.73	15	0.77	0.83	0.80	21	0.76	0.60	0.68
(d) VB (known)	8	0.70	0.91	0.80	14	0.75	0.82	0.78	16	0.68	0.86	0.76	21	0.60	0.80	0.69
(e) VB (best)	12	0.85	0.89	0.87	14	0.84	0.81	0.82	14	0.75	0.90	0.82	13	0.63	0.80	0.71
(f) VB (selected)	15	0.85	0.89	0.87	14	0.84	0.81	0.82	14	0.75	0.90	0.82	13	0.64	0.72	0.68

Table 1. Results on NIST 1996 HUB-4 evaluation test for speaker clustering: ML/BIC vs. VB with non-informative priors

Let us consider at first results of ML/BIC and VB. In order to compare them in fairest way VB priors are initialized as non-informative priors (i.e. small relevance factor τ that brings no information). The system is initialized with $M_{initial} = 35$ speakers modeled by a 15 components GMM. Results are shown in table 1. First of all, VB baseline and best results (lines d-e) are higher than the ML/BIC results (lines a-b) on the first three files while they are almost similar on the last one. It is very important to notice that on the first three files the VB selected model corresponds to the best model; this shows the fact that the VB bound is a very effective metrics for performing model selection. Results in table 1 for ML/BIC refers to values selected using $\lambda = 2$: for this threshold value, BIC selected models are near to the best ML model (even if their K score are lower compared to VB scores). Anyway results in the ML/BIC system are extremely sensitive to the value of λ. In File 1 inferred speaker number is far away from the real speaker number probably because of the fact that a big part of the file is non-speech events that are clustered in many different clusters: anyway final K is high. In File 2 and File 3 inferred speaker number is near to real speaker number (File 2 contains very few non-speech parts). Finally in File 4 BIC infers the right cluster number while VB does not: anyway final K score is the same for BIC and ML. As we outlined in section 3.2, VB should infer the best Gaussian component number per cluster together with the best speaker number. Figure 1 plots on a double Y axis graph final Gaussian components (left Y axis) and observation number assigned to a cluster (right Y axis). It is easy to notice that small amount of data assigned to a cluster results in a smaller number of final Gaussian components; on the other hand a large amount of data results in a model that keeps all Gaussian components (15 in our case). In Figure 2 free energy and score K are plotted on the same graph w.r.t. number of speakers for file 1. The free energy follows closely the score K for all considered number of speakers resulting in an extremely useful criterion for inferring the best system (similar graphs can be obtained for the other 3 files). As final remark we can notice that the best score never corresponds to the score obtained initializing the system with the real speaker number.

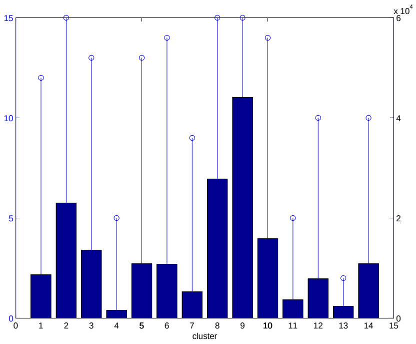

Fig. 1. Thick line (left Y axis): final Gaussian components vs. cluster number; big line (right Y axis): observation number assigned to a cluster vs. cluster number.

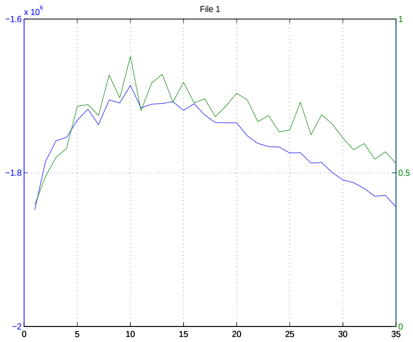

Fig. 2. Blue line (right Y axis): free energy vs. number of clusters (states); Green line (left Y axis): K vs. number of clusters; Free energy follows the clustering score.

Another interesting result can be obtained comparing the VB method with another partially Bayesian method like the MAP. In fact MAP needs as well prior distributions but does not produce any posterior distributions contrarily to VB. MAP and VB can be initialized with the same prior distributions that can be obtained from some previous knowledge on the data. This is the idea of all adaptation methods that initialize prior distributions with a general speaker model called Universal Background Model (UBM). We want to compare here the adaptation obtained with the MAP criterion with the adaptation obtained with the VB criterion. In the MAP system the model selection is the BIC while in the VB system the free energy is used. The UBM used is a 32 component GMM estimated from the BN96 HUB4 training data set. The system is initialized with $M_{initial} = 35$ speakers and results are again provided in terms of average cluster purity, average speaker purity and $K = \sqrt{acp \cdot asp}$.

Table 2 shows results on the four files. Line (a) shows MAP results when the speaker number is a priori known, line (b) shows the best score obtained by

File	File 1				File 2				File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K
(a) MAP (known)	8	0.52	0.72	0.62	14	0.68	0.78	0.73	16	0.71	0.77	0.74	18	0.65	0.69	0.67
(b) MAP (best)	20	0.81	0.84	0.83	22	0.84	0.80	0.82	29	0.78	0.74	0.76	18	0.65	0.69	0.67
(c) MAP (selected)	15	0.80	0.81	0.81	18	0.78	0.85	0.81	16	0.69	0.77	0.73	20	0.63	0.64	0.64
(d) VB (known)	8	0.68	0.88	0.77	14	0.69	0.80	0.74	16	0.74	0.83	0.78	21	0.67	0.73	0.70
(e) VB (best)	22	0.83	0.85	0.84	18	0.85	0.87	0.86	22	0.82	0.82	0.82	20	0.69	0.72	0.70
(f) VB (selected)	22	0.83	0.85	0.84	19	0.87	0.80	0.83	16	0.78	0.79	0.79	19	0.67	0.73	0.70

Table 2. Results on NIST 1996 HUB-4 evaluation test for speaker clustering: MAP/BIC vs. VB with informative priors

the MAP system changing speaker number from $M_{initial} = 35$. Line (c) shows results for MAP system with BIC selection. Lines (d),(e) and (f) are analogous to lines (a), (b) and (c) but model learning and model selection is done using VB learning. We actually present in line (c) the best results obtained with an empirical threshold set to $\lambda = 0.4$.

First of all we can notice that on the three considered situation VB always outperforms the MAP/BIC framework. Probably the most interesting result comes from best results obtained from the two approaches (lines (b) and (e)) that shows that VB does not simply make selection better than MAP but can also adapt a model that holds a higher score. Results with informative priors are still comparable to results with non-informative priors described in table 1.

Inferred cluster number is near to real speaker number for file 3 and file 4 while it is definitely far from reality in file 1 and file 2. Actually final cluster number obtained with informative priors is always higher than the one obtained using non-informative priors described in table 1. It can easily explained considering the fact that models are adapted from a background model giving origin to some small spurious clusters that are not merged together. For instance in file 1 the real cluster number is 8 while the inferred cluster number is 22, anyway values of *acp* and *asp* are high showing a good clustering; this is probably due to the fact there are many small clusters of speech and non-speech that are not merged together.

The use of informative priors (i.e. a background model) for speaker clustering presents the advantage that robust models can be obtained with small amount of data. Sometimes a speaker does not provide enough speech to generate a model and in systems without prior information it is simply clustered together with other speakers: that explains the fact in our previous non-informative prior system (see table 1), inferred cluster number is smaller. Anyway a drawback comes from the quality of the background model: if for any reason it is not a good prior model for the current speech, the same speaker may be split in more clusters. This is a very important issue in Broadcast news segmentation because speech is often corrupted by many noise sources (e.g. music, background speech, various noises) that are obviously unpredictable by the background model; in those cases an absence of prior information may be more efficient (for clustering) than a wrong prior information. For this reason the system would definitively benefits of a preliminary step of speech/non-speech discrimination.

5 Conclusions

In this paper we have studied a speaker clustering system based on the VB framework for model learning and model selection with non-informative and informative (i.e. an UBM model is used for prior distributions initialization). We compared results with a ML/BIC system and a MAP/BIC system. VB outperforms both approaches both in model learning and in model selection.

References

[1] Schwartz G., "Estimation of the dimension of a model",Annals of Statistics, 6, 1978.

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

[3] Tritschler A. and Gopinath R,"Improved speaker segmentation and segments clustering using the Bayesian information criterion",Proceedings of Eurospeech'99,679–682,1999.

[4] 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.

[5] MacKay D. J. C. "Ensemble Learning for Hidden Markov Models", /www.inference.phy.cam.ac.uk/mackay/abstracts/ensemblePaper.html1997.

[6] Bishop C. M. and Winn J.,"Structured variational distributions in VIBES",Proceedings Artificial Intelligence and Statistics,C. M. Bishop and B. Frey editors,Society for Artificial Intelligence and Statistics,2003.

[7] Beal M.,"Variational Algorithm for Approximate Bayesian Inference",2003,PhD thesis, The Gatsby Computational Neuroscience Unit, University College London

[8] Attias H."A variational Bayesian framework for graphical models",Advances in Neural Information Processing Systems,12,2000,209–215.

[9] Attias H.,"Inferring parameters and structures of latent variable models by Variational Bayes",Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence,21–30,1999.

[10] Olsen J. O., "Separation of speakers in audio data", EUROSPEECH 1995, pp. 355-358.

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

[12] Lapidot I. "SOM as Likelihood Estimator for Speaker Clustering",EUROSPEECH 2003.

Appendix

In this appendix we give details the EM-like algorithm for variational Bayesian learning for model 13. Two kinds of latent variables x and z must be considered here: a variable x that designate the speaker (or equivalent state) that is speaking, and z (conditioned to x) that designate the Gaussian component that has emitted the observation. We assume prior distributions as (14). The E-like step i.e. expression (8) for hidden variables x and z can be written as:

$$q(x_t, z_{tp}|O_{tp}) = q(z_{tp}|O_{tp}, x_t)q(x_t|O_{tp}) \propto exp\{< log\alpha_{x_t} > + < log\beta_{x_t, z_{tp}} > + < log\, P(O_{tp}|x_t, z_{tp}) >\} \tag{19}$$

Developing (19), it is possible to derive $\tilde{\gamma}_{x_t=j} = q(x_t|O_{tp})$ and $\tilde{\gamma}_{z_{tp}=i|x_t=j} = q(z_{tp}|O_{tp}, x_t)$ where j is the hidden state and i is the hidden Gaussian.

$$\tilde{\gamma}^*_{z_{tp}=i|x_t=j} = \tilde{\beta}_{ij}\, \tilde{\varGamma}_{ij}^{1/2}\, exp\{-E\}\, exp\{\frac{-g}{2\nu_{ij}}\}\ \ with\ \ E \quad = \frac{1}{2}(O_{tp} - \rho_{tp})^T \bar{\varGamma}_{ij}(O_{tp} - \rho_{tp}) \tag{20}$$

$$\tilde{\gamma}_{z_{tp}=i|x_t=j} = q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j) = \frac{\tilde{\gamma}^*_{z_{tp}=i|x_t=j}}{\sum_i \tilde{\gamma}^*_{z_{tp}=i|x_t=j}} \tag{21}$$

$$\tilde{\gamma}^*_{x_t=j} = \tilde{\alpha}_j \prod_{p=1}^{D} \sum_{i=1}^{M} \tilde{\gamma}^*_{z_{tp}=i|x_t=j} \tag{22}$$

$$\tilde{\gamma}_{x_t=j} = q(\gamma_{x_t} = j) = \frac{\tilde{\gamma}^*_{x_t=j}}{\sum_j \tilde{\gamma}^*_{x_t=j}} \tag{23}$$

where g is the dimension of acoustic vectors. Parameters expected values can be computed as follows:

$$log\, \tilde{\alpha}_j = \varPsi(\lambda_{\alpha_j}) - \varPsi(\sum_j \lambda_{\alpha_j}); \qquad log\, \tilde{\beta}_{ij} = \varPsi(\lambda_{\beta_{ij}}) - \varPsi(\sum_j \lambda_{\beta_{ij}}); \tag{24}$$

$$log\, \tilde{\varGamma}_{ij} = \sum_{i=1}^{g} \varPsi((\nu_{ij} + 1 - i)/2) - log\, |\varPhi_{ij}| + glog2; \qquad \bar{\varGamma}_{ij} = \nu_{ij} \varPhi_{ij}^{-1}; \tag{25}$$

where $\varPsi$ is the digamma function. In the M step, we know that posterior distributions will have the same form of prior distributions i.e. distributions (15). Re-estimation formulas for parameters are given by:

$$\alpha_j = \frac{\sum_{t=1}^{T} \tilde{\gamma}_{x_t=j}}{T} \tag{26}$$

$$\beta_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j}}{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}} \tag{27}$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j} O_{tp}}{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j}} \tag{28}$$

$$\varGamma_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j} (O_{tp} - \mu_{ij})^T (O_{tp} - \mu_{ij})}{\sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j}} \tag{29}$$

and hyperparameter re-estimation formulas are given by:

$$\lambda_{\alpha_j} = \sum_{t=1}^{T} N_j + \lambda_{\alpha 0}; \qquad \lambda_{\beta_{ij}} = N_{ij} + \lambda_{\beta 0}; \qquad \rho_{ij} = \frac{N_{ij}\, \mu_{ij} + \xi_0\, \rho_0}{N_{ij} + \rho_0}; \tag{30}$$

$$\varPhi_{ij} = N_{ij}\, \varGamma_{ij} + \frac{N_{ij} \xi_0 (\mu_{ij} - \rho_0)(\mu_{ij} - \rho_0)^T}{N_{ij} + \rho_0} + \varPhi_0; \tag{31}$$

$$\nu_{ij} = N_{ij} + \nu_0; \qquad \xi_{ij} = N_{ij} + \xi_0; \tag{32}$$

$$\tag{33}$$

where $N_{ij} = \sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j}$ and $N_j = \sum_{t=1}^{T} \tilde{\gamma}_{x_t=j}$.