VARIATIONAL BAYESIAN ADAPTATION FOR SPEAKER CLUSTERING

Fabio Valente, Christian Wellekens

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

ABSTRACT

In this paper we explore the use of Variational Bayesian (VB) learning for adaptation in a speaker clustering framework. Variational learning offers the interesting property of making model learning and model selection at the same time. We compare VB learning with a classical MAP/BIC (MAP for training, BIC for model selection) approach. Results on the NIST BN-96 HUB4 database show that VB learning can outperform the classical MAP-BIC method.

1. INTRODUCTION

A main task in many speech recognition and audio indexing systems consists in unsupervised speaker clustering . This task is also known as speaker diarization. Many different models have been proposed for achieving this purpose e.g. Hidden Markov Models (see [1]) or Self Organizing Map (see [3]). Clustering is generally done in a completely unsupervised fashion. A main problem with this task is that sometimes very few data per speaker are available and robust speaker models cannot be obtained. In order to overcome this problem, speaker model is generally obtained adapting a prior speaker model; adaptation is usually achieved using MAP (see [14],[12]).

In many situation the real speaker number is not known and it must be estimated from data. The most common model selection criterion used in speaker clustering is the *Bayesian Information Criterion* (BIC) used to penalize too complex models. So the speaker clustering procedure can be seen as a two step procedure in which at first a model is learned using Maximum Likelihood (ML) or MAP adaptation, and then the model is scored using BIC. Best model is the model with highest BIC score. A relatively new techniques for making model selection and parameter learning at the same time is the Variational Bayesian (VB) framework.

In [13] and [11] we investigated the use of Variational Bayesian learning for unsupervised speaker clustering when no information at all on speaker is available and compared results with a ML/BIC criterion. Results shows that VB can outperform ML/BIC on this task. In this paper we consider the case in which a background model for speakers is available and speaker model is obtained adapting the background model. We compare the speaker segmentation obtained using the VB framework against segmentation obtained using a MAP/BIC showing that even in this case VB can perform better.

This paper is organized as follows: in the next section our model for speaker clustering is presented, then MAP/BIC and VB model selection are introduced and finally experiments on NIST BN-96 HUB4 database evaluation set are discussed.

2. HMM FOR SPEAKER CLUSTERING

In this section we define the model we used for the speaker clustering task. A popular approach uses Ergodic Hidden Markov Models. This method introduced in [1] considers a fully connected HMM in which each state represents a speaker and the state emission probability is the emission probability for each speaker. In order to obtain a non-spare solution, we use a duration constraint of 100 consecutive frames as proposed in [3] in order to model each speaker in a robust way.

Let us designate α_{rj} the transition probability from state r to state j. We make here the assumption that the probability of transition to state j is the same regardless the initial state i.e. $\alpha_{rj} = \alpha_{r'j} \quad \forall r, r'$, where $j = 1, ..., S$ with S the total number of states; in other words, under this assumption we can model the ergodic HMM as a simple mixture model. 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 :

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

where S represent the number of states (that represent a speaker), M represents gaussian component that models each speaker, and $\{\beta_{ij}, \mu_{ij}, \Gamma_{ij}\}$ represent mixture model parameters (weights, means and gaussians) . When S is unknown, it must be estimated with a model selection criterion (for details about this model see [11]).

3. MODEL SELECTION AND PARAMETER LEARNING

Let us consider a data set $Y = \{y_1, ..., y_n\}$ and a model m. Each model m has a set of parameters θ with their distributions. The best model is the model that maximize $p(m|Y)$ i.e. applying bayes rule we obtain:

$$m = argmax_m\, p(m|Y) = argmax_m\, p(m)\, p(Y|m)/p(Y) \quad (2)$$

where $p(Y)$ is ignored because it does not depend on m. If $p(m)$ is considered uniform on the possible model space, the best model is the model that maximize the so called *marginal likelihood* $p(Y|m)$ and can be computed marginalizing w.r.t. model parameters i.e.

$$p(Y|m) = \int p(Y|\theta, m)\, p(\theta|m)\, d\theta \quad (3)$$

Marginal likelihood benefits from the so called *Occam's razor* properties (see e.g. [15]) i.e. simpler models are preferred to

more complex one. Two main difficulties arise in computing quantity (3). The first a reasonable choice for parameter distribution $p(\theta|m)$ must be found. Then integral is not always possible because in many models (HMM, GMM) it requires integration over latent variables and parameter distributions. Let us consider separately solutions to those two problems. The simplest choice to estimate parameter distributions is MAP posterior distributions that maximize the joint data and parameter probability *density* giving the following parameter estimate i.e.

$$\theta_{MAP} = argmax_\theta \; p(\theta)p(Y|\theta) \tag{4}$$

This solves the first problem (parameter distributions) but not the second (making the integral). Generally the most common solution is approximating the integral with a simplest function. The *Bayesian Information Criterion* (see [10]) is an asyntoptical approximation of marginal likelihood (3) i.e.

$$log\, p(Y|m) = log\, p(Y|m, \hat{\theta}) - \frac{\nu}{2} log\, N \tag{5}$$

where $\hat{\theta}$ is the estimation for model parameters θ (here we consider a MAP estimation), ν is the free parameter number and N is the observation number. In real data applications, penalty terms is generally multiplied by a threshold value λ heuristically determined. It is important to notice that in (5) there's no reference to distributions $p(\theta|m)$ but only on the number of free parameters. It can be shown that when $N \to \infty$, BIC converges to marginal likelihood.

In the next section we will show how Variational Bayesian learning solves the two problems with the same elegant solution.

4. VARIATIONAL BAYESIAN LEARNING

Variational Bayesian learning is an approximated method that allows computation of an upper bound of marginal log-likelihood (3). First of all let us suppose that the true parameter posterior distribution $p(\theta|m)$ can be approximated by some other distributions $q(\theta|Y, m)$ (referred as variational bayesian posterior distributions). Using Jensen inequality, it is possible to write:

$$\begin{aligned} log\, p(Y|m) &= log \int d\theta q(\theta|Y,m) \frac{p(\theta|m)p(Y,\theta|m)}{q(\theta|Y,m)} \\ &\geq \int d\theta q(\theta|Y,m) log \frac{p(Y,\theta|m)}{q(\theta|Y,m)} = F(\theta) \end{aligned} \tag{6}$$

$F(\theta)$ is called *Free Energy* and it is a strict lower bound on the log marginal-likelihood. It is easy to show that the difference between marginal log-likelihood and the free energy is:

$$KL(q(\theta|Y,m)||p(\theta|Y,m)) = -\int q(\theta|Y,m) log \frac{p(\theta|Y,m)}{q(\theta|Y,m)} \tag{7}$$

i.e. the KL divergence between approximated distributions and exact posterior distributions (that are actually unknown). If $q(\theta|Y, m) = p(\theta|Y, m)$ then expression (7) would be zero, and the bound will be equal to the true marginal log-likelihood.

Let us now manipulate the free energy $F(\theta)$ and write it in the following form:

$$F(\theta) = \int d\theta\, q(\theta|Y,m) log\, p(Y|\theta,m) - D(q(\theta|Y,m)||p(\theta|m)) \tag{8}$$

Variational Bayesian learning aims at optimizing $F(\theta)$ w.r.t. variational posterior distribution $q(\theta|Y, m)$.

Second term in expression (8) represents the KL divergence between variational posterior distributions and parameter prior distributions; it acts as a penalty term that becomes huger for more complex models. In this sense the free energy can be used as a model selection criterion (see section 4.2).

If the variational posterior distribution is constrained to be a delta distribution i.e. $q(\theta|Y, m) = \delta(\theta - \theta')$, the free energy reduces to the MAP estimator:

$$\begin{aligned} max_{Q(\theta)} F(\theta) &= max_{\theta'} \int \delta(\theta - \theta') log[p(Y|\theta)p(\theta)] d\theta \\ &= max_{\theta'} log[p(Y|\theta')p(\theta'))] \end{aligned} \tag{9}$$

where the term $\int q(\theta) log\, q(\theta) d\theta$ has been dropped because it is constant.

4.1. Variational Bayesian learning with hidden variables

Many popular models like HMM or GMM use hidden variables. Hidden variables make impossible the computation of (3) in closed form. In [4] the problem is solved using an independence assumption between hidden variables and parameters. This is actually the key for computing, even if in an approximated form marginal likelihood. Let us denote by X the hidden variables set. Variational posterior becomes $q(X, \theta|Y, m)$ and the simplification is assuming it factorizes as $q(X, \theta|Y, m) = q(X|Y, m)q(\theta|Y, m)$. Then the free energy to maximize is:

$$\begin{aligned} F(\theta, X) &= \int d\theta\, dX\, q(X|Y,m) q(\theta|Y,m) log[\frac{p(Y,X,\theta)}{q(X|Y,m)q(\theta|Y,m)}] \\ &= < log \frac{p(Y,X|\theta)}{q(X|Y,m)} >_{X,\theta} - D[q(\theta|Y,m)||p(\theta|m)] \end{aligned} \tag{10}$$

where $< . >_z$ means average w.r.t. z. Note that q is always understood to be conditioned on Y and m. To find the optimum $q(\theta)$ and $q(X)$ an EM-like algorithm is proposed in [4] based on the following steps:

$$q(X|Y,m) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \tag{11}$$

$$q(\theta|Y,m) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta|m) \tag{12}$$

By iteratively applying eq.(11) and eq.(12), it is possible to estimate variational posteriors for parameters and hidden variables. If $p(\theta|m)$ belongs to a conjugate family, posterior distribution $q(\theta|Y, m)$ will have the same form as $p(\theta|m)$.

Under this assumption it can be shown that free energy can be computed in a closed form for conjugate-exponential models (see [16]).

4.2. Model selection

An extremely interesting property of the Variational Bayesian learning is the possibility of selection models while training. As it was outlined in the previous section, the free energy (8) can be used as a model selection criterion because the KL distance between parameter posterior distributions and parameter prior distributions acts as a penalty term similar to the BIC criterion penalty. Let us introduce the model posterior probability $q(m)$ on a given model m. It can be shown (see [4]) that optimal $q(m)$ can be written as:

$$q(m) \propto exp\{F(\theta, X, m)\}\, p(m) \tag{13}$$

where $p(m)$ is the model prior. In absence of any prior information on model, $p(m)$ is uniform and optimal $q(m)$ will simply depend on the term $F(\theta, X, m)$ i.e. since higher free energies will result in higher $q(m)$, free energy can be used as model selection criterion.

To summarize, Variational Bayesian learning offers a solution to joint optimization of parameters and marginal likelihood bound at the same time, giving the possibility of selecting a model and learning parameters simultaneously.

5. DISCUSSION: VB VS. MAP/BIC

MAP and VB learning are achieved using iterative algorithms, the Expectation-Maximization algorithm (see [8]) for the MAP (see [14]) and an EM-like algorithm for the VB proposed [4]. As previously outlined MAP is a point estimate (it considers *densities*) while VB is a *mass* estimate in other words MAP optimizes parameters while VB optimizes distributions over parameters. Because of this, VB embeds the so called *Occam razor* property that allow the training to learn the best model avoiding overfitting (see [15],[16]). From a practical point of view both algorithms take benefits of the fact prior distributions over parameters are chosen in the conjugate family because posterior distributions have the same form of prior i.e. for both approaches we have

$$post(\theta) \ = p(\theta|Y) \ \times \ prior(\theta) \tag{14}$$

where $post$ and $prior$ are respectively the prior and the posterior distribution and have the same form. This will result in a similar M-step in the learning algorithm but in a different E-step ([4]). Another important difference consists in prediction on unseen data. MAP trained models can use parameters, but VB does not have optimal parameters but optimal distributions that must be integrated out. As outlined before, integration cannot be done exactly but approximation (6) must be considered again.

6. EXPERIMENTS

6.1. Occam razor principle

As previously discussed, VB learning is a full bayesian learning (contrarily to MAP) and benefits from the so called *Occam razor* properties i.e. fitting data with models that are not too big or too small but 'just right'. In order verify this principle we run the following experiment: let us consider a speaker background model GMM with $M = 256$ components and let adapt it with different amount of data. Let us consider the accumulator (a.k.a. gaussian statistics) of each gaussian component with its mean $(1/M)$ and its variance. In this case, variance can be seen as a measure of how hard clustering is done, high variance will mean hard clustering (data are clustered in few big clusters) while small variance will mean soft clustering (data are split over all clusters). Figure (1) plots accumulator variance as a function of the amount of data used for adaptation. It is easy to notice that for small amount of data, MAP clusters data harder than VB (giving overfitting problems), while when available amount of data increases, VB accumulator variance is higher than MAP. This can be seen as a consequence of the Occam razor principle that tries to find an 'average' solution for all possible amounts of data avoiding too simple models when few data are used or too complex models when large amount of data are used.

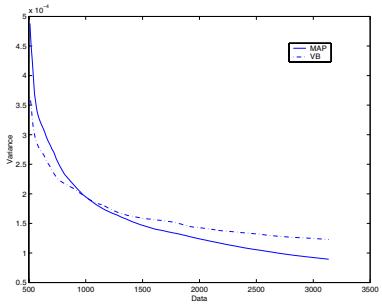

Fig. 1. Accumulator variance w.r.t. amount of data

6.2. Broadcast News speaker clustering

In order to compare the VB model selection and the MAP/BIC model selection we run experiments on the four files of the evaluation data set NIST-1996 HUB-4 . All files are processed in order to obtain 12 LPCC coefficients.

Contrarily to what we have done in [13] where we have used heuristic priors, priors to speaker model is this time informative i.e. a background speaker model is provided and is adapted using MAP or VB procedure. We find out that a 32 component GMM is appropriate to this task. The model is trained with data contained in training set labeled as regions F0, F1, F2.

Prior distributions are fixed belonging to a conjugate family. Let us define following probability distributions over parameters in model (1):

$$P(\alpha_j) = Dir(\lambda_{\alpha\,0j}) \ \ P(\beta_{ij}) = Dir(\lambda_{\beta\,0j})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho_{0j}, \xi_{0j}\Gamma_{ij}) \ \ P(\Gamma_{ij}) = W(\nu_{0j}, \Phi_{0j})$$

where $Dir(), N(), W()$ are respectively Dirichlet, Normal, Wishart distributions and $\{\lambda_{\alpha\,0j}, \lambda_{\beta\,0j}, \rho_{0j}, \xi_{0j}, \nu_{0j}, \Phi_{0j}\}$ are hyperparameters that come from a background model.

The training procedure uses the following algorithm: the system is initialized with a huge speaker number $M_{initial}$ then optimal parameters are learned using adaptation based on VB and on 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 the different models with VB free energy (that was used as objective function in the training step) and with BIC criterion. The system is initialized with $M_{initial} = 35$ speakers modeled by a 32 components GMM with duration constraint of 100 frames (1 second). Adaptation using the MAP framework uses the same EM procedure described in [14]. Details about estimation formula for VB learning applied to model (1) can be found in [11]. Actually only distributions on means and gaussian weights are considered for adaptation, while covariance matrices coming from background model are not modified.

Results are provided in terms of average cluster purity (*acp*) and average speaker purity (*asp*) and $K = \sqrt{acp \cdot asp}$ (for details see [11]).

Table 1 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 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

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

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 it is the training itself that has a higher score. Results with informative priors are still comparable to that with heuristic prior described in [13].

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 heuristic priors described in [13]. It can easily explained considering the fact that models are adapted from a background model giving origin to some small spurious cluster 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 for *acp* and *asp* is high showing a good clustering; this is because 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 in a bigger cluster: that explains the fact in our previous heuristic prior system (see [13]), 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 prior step of speech/non-speech discrimination.

7. CONCLUSION

In this paper we compare on a speech clustering task two framework the first one based on MAP learning and BIC selection and the second one based on Variational Bayesian learning that allows model selection and parameter learning at the same time. Both systems use prior information constituted by a background model trained on the BN train set. Performance provided in terms of cluster purity and speaker purity shows that VB approach to speaker clustering can outperform MAP/BIC approach. This confirm what we have investigated in ([13]), where we compared VB approach with heuristic priors against ML/BIC coming to the same conclusions.

8. REFERENCES

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

[2] Ajmera J., "Unknown-multiple speaker clustering using HMM",ICSLP 2002.

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

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

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

[6] MacKay D.J.C., "Ensemble Learning for Hidden Markov Models", http://www.inference.phy.cam.ac.uk/mackay/

[7] Cohen A. et Lapidus V. " Unsupervised text independent speaker classification", Proc. of the Eigtheenth Convention of Electrical and Electronics Engineers in Israel 1995, pp. 3.2.2 1-5

[8] Dempster A.P. , Laird N.M. , and Rubin D.B. , "Maximum Likelihood from Incomplete Data via theEM algorithm". Journal of the Royal statistical Society, Series B, 39(1): 1-38, 1977

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

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

[11] Valente F., Wellekens C. "Variational Bayesian Speaker Clustering", Proc. Odyssey 2004

[12] Reynolds D.,Quatieri T., and Dunn R., "Speaker verification using adapted Gaussian mixture models" Digital Signal Processing, vol. 10, no. 1-3, 2000.

[13] Valente F., Wellekens C. "Scoring unknown speaker clustering: VB vs. BIC" Proceedings of ICSLP 2004, Korea

[14] Gauvain, J.-L.; Chin-Hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains",Speech and Audio Processing, IEEE Transactions on , Volume: 2 , Issue: 2 , April 1994

[15] McKay D.J.C. "Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks", Network:Computation in Neural System 6,1995

[16] Beal, M.J., Ghahramani, Z. "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structure " , Bayesian Statistics 7, Oxford University Press, 2003