Variational Bayesian Speaker Clustering

Fabio Valente, Christian Wellekens

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

Abstract

In this paper we explore the use of Variational Bayesian (*VB*) learning in unsupervised speaker clustering. VB learning is a relatively new learning technique that has the capacity of doing at the same time parameter learning and model selection. We tested this approach on the NIST 1996 HUB-4 evaluation test for speaker clustering when the speaker number is a priori known and when it has to be estimated. VB shows a higher accuracy in terms of average cluster purity and average speaker purity compared to the Maximum Likelihood solution.

1. Introduction

An important task in many speech recognition applications consists in clustering speakers. A huge number of techniques for achieving robust speaker clustering have been proposed: they generally consist in vector quantizer [11], Hidden Markov Models (HMM) [2] and Self-Organizing Maps (SOM) [3]. A main issue in unsupervised learning is that the exact speaker number is not known; in order to determine a reasonable number of clusters a model selection method must be used; generally the BIC criterion is used or a "revisited" BIC [2].

In this paper we propose the use of a relatively new learning technique that allows simultaneous parameter learning and model selection; it is generally referred as Variational Bayesian Learning (or Ensemble Learning). Models like GMM and HMM can be learned using the VB framework (see [4],[10]). VB training has the advantage of using as optimization criterion an expression that can be also used for model selection. Even if VB learning is an approximated method, it has already been successfully applied in speech recognition problems for state clustering ([5]), dimension reduction ([6]), and GMM estimation ([7]). Here we apply VB methods to speaker clustering.

The paper is organized as follows: in section 2 we describe the baseline system featuring HMM trained using the EM algorithm, in section 3 we describe the general VB framework, in section 4 we describe speaker clustering system that uses VB learning and finally in section 5, we describe experiments and results.

2. HMM based speaker indexing

A popular approach to automatic speaker clustering uses Hidden Markov Model. This method introduced in [1] consider a fully connected HMM in which each state represent a speaker and the state emission probability is the emission probability for each speaker. In [2] an ergodic HMM with duration constraint is proposed. Duration constraint has the advantage of giving a non-sparse solution. Using a given number of consecutive frames gives in fact enough statistic to model a certain speaker in a robust way. In [3], it was shown that 100 consecutive frames are enough to build a speaker model.

Let us now details the model. An ergodic HMM is a fully connected HMM in which all possible transition are allowed. Let us define $P(O_t|s_j)$ the probability of observation O_t given state s_j at time t. Let us define 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.

$$\alpha_{rj} = \alpha_{r'j} \quad \forall r, r', j = 1, ..., S \tag{1}$$

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 and write the probability of an observation O_t as:

$$P(O_t) = \sum_{j=1}^{S} \alpha_j P(O_t|s_j) \tag{2}$$

In a model with duration constraint, the observation O_t is a group of D consecutive frames where D is the duration constraint. A possible way to model $P(O_t)$ is using a Gaussian Mixture Model with mixing coefficients β_{ij}, means μ_{ij} and variance Γ_{ij} where $i = 1, ..., M$ with M the number of Gaussians. It is then possible to write the whole model as:

$$P(O_t) = \sum_{j=1}^{S} \alpha_j \{ \prod_{p=1}^{D} \sum_{i=1}^{M} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij}) \} \tag{3}$$

It is so possible to write the log-likelihood of a given sequence O_t with $t = 1, ..., T$:

$$log\, P(O) = \sum_{t=1}^{T} log \sum_{j=1}^{S} \alpha_j \{ \prod_{p=1}^{D} \sum_{i=1}^{M} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij}) \tag{4}$$

In other words this model is a Hierarchical Mixture Model in which the first layer represents the S speakers and the second layer represents speaker models that is actually a GMM. α_j can be seen as the prior probability of the $j-th$ speaker.

Once the model is learned, observations can be assigned to a cluster (speaker) using a simple Viterbi decoding.

This kind of model can be completely trained using classical EM algorithm. We have previously made the hypothesis that the speaker number is a priori known but this is not always the case. For this reason a model selection criterion must be used, if speaker number is not a priori known. In the next section we will consider the case of EM training when the speaker number is known.

2.1. EM learning

Model 4 is a latent variable models that can be learned using the well known Expectation-Maximization algorithm [12]. 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. For the Expectation step, it is easily demonstrated that:

$$\gamma_{x_t=j} = P(x_t = j|O_t) = \frac{\alpha_j P(O_t|s_j)}{\sum_j \alpha_j P(O_t|s_j)} \tag{5}$$

$$\gamma_{z_{tp}=i|x_t=j} = P(z_{tp} = i|x_t = j, O_{tp}) = \frac{\beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})}{\sum_{i=1}^{D} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})} \tag{6}$$

For the Maximization step, following reestimation formula can be derived.

$$\alpha_j = \sum_{t=1}^{T} \gamma_{x_t=j}/T \tag{7}$$

$$\beta_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}} \tag{8}$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j} O_{tp}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}} \tag{9}$$

$$\Gamma_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j} (O_{tp}-\mu_{ij})^T (O_{tp}-\mu_{ij})}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}} \tag{10}$$

3. Variational Bayesian Learning

In this section we introduce the Variational Bayesian learning. First, we consider a very general framework and then we show how to train the model (4).

3.1. Variational Bayesian Framework

Given a set of observed variables Y and some parameters θ, Bayesian learning aims at optimizing the so called *marginal likelihood* $p(Y)$, where parameters θ have been integrated out. From Bayes rule we have: $p(Y) = p(Y,\theta)/p(\theta|Y)$ and considering the log of both members it is possible to write: $log\, p(Y) = log\, p(Y,\theta) - log\, p(\theta|Y)$. Instead of integrating parameters θ w.r.t. their true unknown pdf, an approximation called variational posterior, and denoted as $q(\theta|Y)$, is used. Taking expectation w.r.t $q(\theta|Y)$, we obtain:

$$\begin{aligned} log\, p(Y) &= \int q(\theta|Y) log\, p(Y,\theta) d\theta - \int q(\theta|Y) log\, p(\theta|Y) d\theta \\ &= \int q(\theta|Y) log\, [p(Y,\theta)/q(\theta|Y)] d\theta + D(q(\theta|Y))||p(\theta|Y)) \end{aligned} \tag{11}$$

where $D(q(\theta|Y))||p(\theta|V)$ represents the Kullback-Leiber (KL) distance between the variational posteriors and the true posteriors. The term $\int q(\theta|Y) log\, [p(Y,\theta)/q(\theta|Y)] d\theta$ is often indicated as *negative free energy* $F(\theta)$. Because of the KL-distance property $D(a||b) \geq 0$ (with equality if $a = b$), $F(\theta)$ represent a lower bound on $log\, p(Y)$ i.e. $log\, p(Y) \geq F(\theta)$. Variational Bayesian learning aims at maximizing the lower bound $F(\theta)$ that can be rewritten as:

$$F(\theta) = \int q(\theta|Y) log\, p(Y|\theta) d\theta - D(q(\theta|Y)||p(\theta)) \tag{12}$$

The second term in eq. (12) represents the distance between the approximate posterior and the parameter prior and can be interpreted as a penalty term that penalizes more complex models. For this reason $F(\theta)$, can be used to determine the model that best fits to data in the same way the BIC criterion is used.

Maximum a Posteriori can be seen as a special case of VB learning. In fact, if $q(\theta|Y) = \delta(\theta - \theta')$, finding the maximum of equation (12) means:

$$\begin{aligned} max_{Q(\theta)} F(\theta) &= max_{\theta'} \int \delta(\theta - \theta') log[p(Y|\theta)p(\theta)] d\theta \\ &= max_{\theta'} log[p(Y|\theta')p(\theta'))] \end{aligned} \tag{13}$$

where the term $\int q(\theta) log\, q(\theta) d\theta$ has been dropped because it is constant. Expression (13) corresponds to the classical MAP criterion. It is important to notice that the VB approach carries information about the uncertainty on parameters θ while MAP does not. In fact in MAP, parameter learning is done punctually ($max\, log\, [p(Y|\theta')p(\theta'))]$) while in VB, parameters are integrated out, even if they are integrated w.r.t. variational posterior ($max \int q(\theta|Y)\, log[p(Y|\theta)p(\theta)] d\theta$). Furthermore VB allows model comparison: free energy value gives information on the model quality, while MAP only gives best parameters for an imposed model. The price to pay is that the free energy is only a lower bound and not an exact value.

3.2. Variational Bayesian learning with hidden variables

Variational Bayesian learning can be extended to the incomplete data case. In many machine learning problems, algorithms must take care of hidden variables X as well as of parameters θ (see [4]). In the hidden variable case, the variational posterior becomes $q(X,\theta|Y)$ and a further simplification is assumed considering it factorizes as $q(X,\theta|Y) = q(X|Y)q(\theta|Y)$. Then the free energy to

maximize is:

$$F(\theta, X) = \int d\theta dX q(X)q(\theta)log[p(Y, X, \theta)/q(X)q(\theta)]$$
$$= < log \frac{p(Y, X|\theta)}{q(X)} >_{X,\theta} - D[q(\theta)||p(\theta)] \quad (14)$$

where $< . >_z$ means average w.r.t. z. Note that q is always understood to be conditioned on Y. It can be shown that when $N \rightarrow \infty$ the penalty term reduce to $(|\theta_0|/2)log\ N$ where θ_0 is the number of parameters i.e. the free energy becomes the Bayesian Information Criterion (BIC). To find the optimum $q(\theta)$ and $q(X)$ an EM-like algorithm is proposed in [4] based on the following steps:

$$q(X) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \quad (15)$$
$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta) \quad (16)$$

Iteratively applying eq.(15) and eq.(16) it is possible to estimate variational posteriors for parameters and hidden variables. If $p(\theta)$ belongs to a conjugate family, posterior distribution $q(\theta)$ will have the same form as $p(\theta)$.

An interesting property of VB learning is that extra degrees of freedom are not used i.e. the model prunes itself. There are two possible opinions about the correctness of the model self pruning: on the one hand it is not satisfactory because prediction will not take into account uncertainty that models with extra parameters can provide (see [8]), on the other hand it can be used to find the optimal model while learning the model itself, initializing it with a lot of parameters and letting the model prune parameters that are not used.

4. Speaker clustering using VB

In this section we derive formulas that can be used to estimate parameters in model (4). Before applying the EM-like algorithm previously described, we have to define prior probabilities on parameters. So let us define the following probabilities that belong to the conjugate family.

$$P(\alpha_j) = Dir(\lambda_{\alpha 0}) \quad P(\beta_{ij}) = Dir(\lambda_{\beta 0})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0\Gamma_{ij}) \quad P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \quad (17)$$

where Dir designates a Dirichlet distribution, N a Normal distribution and W a Wishart distribution. The advantage in using probability functions that belong to the conjugate family is that posterior probability will have the same analytical form as priors. So let us introduce the parameters posterior probabilities.

$$P(\alpha_j) = Dir(\lambda_{\alpha j}) \quad P(\beta_{ij}) = Dir(\lambda_{\beta ij})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho_{ij}, \xi_{ij}\Gamma_{ij}) \quad P(\Gamma_{ij}) = W(\nu_{ij}, \Phi_{ij}) \quad (18)$$

Figure 1 shows a direct graph that represent the model.

It is now possible to apply the EM-like algorithms that consists in iteratively applying equations (15) and (16).

Again in the E step we have to consider two kinds of latent variables, x and z that respectively designate the cluster (i.e. the speaker) and the gaussian component:

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

Developing 19, it is possible to derive formulas similar to formulas (5) and(6) but computed on the base of parameter expected values instead of parameter values. We will designate them with $\tilde{\gamma}_{x_t=j}$ and $\tilde{\gamma}_{z_{tp}=i|x_t=j}$.

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

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

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

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

where g is the dimension of acoustic vectors.

Parameters expected values can be computed as follows:

$$log\, \tilde{\alpha}_j = \Psi(\lambda_{\alpha_j}) - \Psi(\sum_j \lambda_{\alpha_j}) \quad (24)$$

$$log\, \tilde{\beta}_{ij} = \Psi(\lambda_{\beta_{ij}}) - \Psi(\sum_j \lambda_{\beta_{ij}}) \quad (25)$$

$$log\, \tilde{\Gamma}_{ij} = \sum_{i=1}^{g} \Psi((\nu_{ij} + 1 - i)/2) - log\, |\Phi_{ij}| + glog2 \quad (26)$$

$$\bar{\Gamma}_{ij} = \nu_{ij}\Phi_{ij}^{-1} \quad (27)$$

where Ψ is the digamma function.

In the M step, we know that posterior distributions will have the same form of prior distributions. Reestimation formulas for parameters are given by:

$$\alpha_j = \frac{\sum_{t=1}^{T} \tilde{\gamma}_{x_t=j}}{T} \quad (28)$$

$$\beta_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}\tilde{\gamma}_{z_{tp}=i|x_t=j}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}} \quad (29)$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}\tilde{\gamma}_{z_{tp}=i|x_t=j} O_{tp}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}\tilde{\gamma}_{z_{tp}=i|x_t=j}} \quad (30)$$

$$\Gamma_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}\tilde{\gamma}_{z_{tp}=i|x_t=j}(O_{tp} - \mu_{ij})^T(O_{tp} - \mu_{ij})}{\sum_{t=1}^{T}\sum_{p=1}^{D} \tilde{\gamma}_{x_t=j}\tilde{\gamma}_{z_{tp}=i|x_t=j}} \quad (31)$$

and hyperparameter reestimation formulas are given by:

$$\lambda_{\alpha_j} = \sum_{t=1}^{T} N_j + \lambda_{\alpha 0} \quad (32)$$

$$\lambda_{\beta_{ij}} = N_{ij} + \lambda_{\beta 0} \quad (33)$$

$$\rho_{ij} = \frac{N_{ij}\ \mu_{ij} + \xi_0\ \rho_0}{N_{ij} + \rho_0} \quad (34)$$

$$\xi_{ij} = N_{ij} + \xi_0 \quad (35)$$

$$\Phi_{ij} = N_{ij}\,\Gamma_{ij} + \frac{N_{ij}\xi_0(\mu_{ij} - \rho_0)(\mu_{ij} - \rho_0)^T}{N_{ij} + \rho_0} + \Phi_0 \quad (36)$$

$$\nu_{ij} = N_{ij} + \nu_0 \quad (37)$$

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 $N_j = \sum_{t=1}^{T} \tilde{\gamma}_{x_t=j}$.

4.1. Model selection using VB

An extremely interesting property of the Variational Bayesian learning is the possibility of doing model selection while training the model. As it was outlined in the previous section, the free energy (12) 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. We will now consider a more rigorous framework for model selection.

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:

$$q(m) \propto exp\{F(\Theta, X, m)\}\, p(m) \tag{38}$$

where $p(m)$ is the model priors. 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. free energy can be used as model selection criterion.

In other words free energy can be used exactly as a model selection criterion. An important advantage is that no threshold must be manually set (as for example in the BIC criterion). For the model considered here, it is possible to obtain a closed form for the free energy (12) (see Appendix A).

As previously outlined, another interesting point in using Variational Bayesian learning is the capacity of pruning extra freedom degrees. It means that it is possible to initialize the system with a high number of clusters and with a high number of gaussians per speaker and let the system eliminate clusters and gaussians that are not used. In gaussian based model the capacity of pruning extra parameters is somehow regulated by the prior parameter Φ_0 that seems to be the more sensitive parameter w.r.t. clustering result (see e.g. [6]). In other words large values of Φ_0 will result in smaller number of final clusters or smaller number of final gaussians.

In [13] an important point is outlined: when different speakers speak for a different amount of time, it is reasonable to model them with different models. Authors propose to use a BIC criterion to determine the best model between a GMM (that performs better when a lot of training data are available) and a VQ (that performs better when few training data are available). The use of Variational Bayesian learning allows a somehow similar effect; if we initialize speaker models with an initial high gaussian number, VB automatically prunes together with the cluster number, the best gaussian model at the same time, resulting in smaller models where few observations are available and in bigger models where more observations are available.

5. Experiments

We run two type of experiments: first set of experiments is on synthetic data, second set is on real data. The database we used for tests on real data is the NIST 1996 HUB-4 evaluation dataset that consists of 4 files of almost half an hour. The first file consists of 7 speakers, the second of 13, the third of 15 speakers and finally the forth of 21 speakers. In files 1, 3, 4 there are large part of non-speech events while file 3 is almost pure speech.

5.1. Evaluation criterion

In order to evaluate the quality of clustering we use concepts of cluster purity and speaker purity introduced respectively in [9] and [3]. We consider in all our tests an additional cluster for non-speech events. Using the same notation of [3], 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:

$$p_i = \sum_{j=0}^{R} \frac{n_{ij}^2}{n_i^2} \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2} \tag{39}$$

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

$$acp = \frac{1}{N} \sum_{i=0}^{S} p_i\, n_i \quad asp = \frac{1}{N_s} \sum_{j=1}^{R} q_j\, n_j \tag{40}$$

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

$$K = \sqrt{asp \cdot acp} \tag{41}$$

5.2. Results on synthetic data

The importance of experiments on synthetic data consists in verifying if in ideal condition the system can achieve very high values of *acp* and *asp*. For this purpose we generated two files in which we simulated a conversation between four speakers each of them modeled with a 3 components gaussian mixture. In the first file (we will refer as

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
Baseline	4	1	1	1	4	0.89	0.99	0.94
VB system I	4	1	1	1	4	0.75	1	0.86
VB system II	4	1	1	1	4	1	1	1

Table 1: Clustering results on synthetic data: baseline vs. Variational Bayesian system I (a priori known cluster number) vs. Variational Bayesian system II (initialized with 30 clusters)

File	File 1				File 2				File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	N_c	K	acp	asp	K
Baseline	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
VB system I	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
VB system II	16	0.81	0.88	0.85	14	0.84	0.81	0.82	14	0.75	0.90	0.82	9	0.53	0.81	0.66

Table 2: Clustering results baseline vs. Variational Bayesian system I (a priori known cluster number) vs. Variational Bayesian system II (initialized with 30 clusters)

file 1) we considered the case where the four 'speakers' pronounce the same amount of 'speech' i.e. 10000 observations each; in second file (we will refer as file 2) we tried to simulate a situation that actually we often found in real conversation in which some speakers pronounce a big amount of data while others just speak for few seconds. File 2 consists of two 'speakers' who pronounce 10000 observations, one who pronounces 5000 observation and one who pronounce just 1000 observations.

We run clustering with three different systems; system 1 (we will refer as our baseline) is initialized with the right number of clusters and the right number of gaussian per cluster and learning is done using EM/ML. System 2 (we will refer as VB system I) is initialized like the baseline system but learning is done using Variational Bayesian method and finally System 3 (we will refer as VB system II) is initialized with a huge number of clusters (30) and each GMM is initialized with 10 gaussian components, learning is done using VB. Results are shown in table 1.

In File 1, unsurprisingly all three systems achieve values of acp and asp of 1. It is interesting to notice that in VB system II, not only the right number of clusters is inferred (4), but also the right number of gaussian components per clusters is inferred (3). In File 2 the baseline and VB system I do not achieve a value of K equal to 1; in fact it looks like the difference of size between clusters play a serious role in the clustering. Of course the speaker who is less represented is the one who is worst modeled. On the other side VB system II is able to achieve a value of K equal to 1; anyway it must be pointed out that the final number of gaussian components is not equal to the number used for the generation of synthetic data in this case.

5.3. Results on real data

In this section we describe experiments on NIST 1996 HUB-4 evaluation dataset.

Features used consist in 12 LPCC calculated every 30 ms frames with a 10 ms frame rate. We found that using a minimum duration constraint of 100 frames (i.e. 1 second) and 15 gaussians per state is enough to ensure robust speaker identification. For the Variational Bayesian system we used the following hyperparamters: $\lambda_{\alpha 0} = \lambda_{\beta 0} = 1$, $\rho_0 = \bar{O}$, $\xi_0 = 1$, $\Phi_0 = 200$ and $\nu_0 = g$ where g is the acoustic vector dimension and $\bar{O}$ is the observation mean values.

In the first set of experiments we fixed the cluster number to the real speaker number plus one cluster for modeling non-speech events as in [3], then we trained the system using EM/ML and VB learning. Results are shown in the first and second row of table 2. On the first two files VB outperforms classical EM/ML while on the last two they gives almost the same results.

We think that there are basically two reasons to explain the best performance of the VB learning towards the classical EM/ML. On one hand final parameters take advantage of a regularization effect coming from prior distributions. On the other hand VB system converges for each speaker to a final GMM model that may have a component number smaller than the original one (15 gaussians in this case) depending on the speech utterances coming from the speaker in consideration. It generally results in higher speaker purity as it can be noted from table (2).

In the second set of experiments we initialized the model with a high cluster number (30) and we let the VB learning prune clusters converging to a final cluster number smaller than 30. Actually in Hub-4 files there are a lot of non-speech events that radically influence the clustering. VB system II performances are shown in third

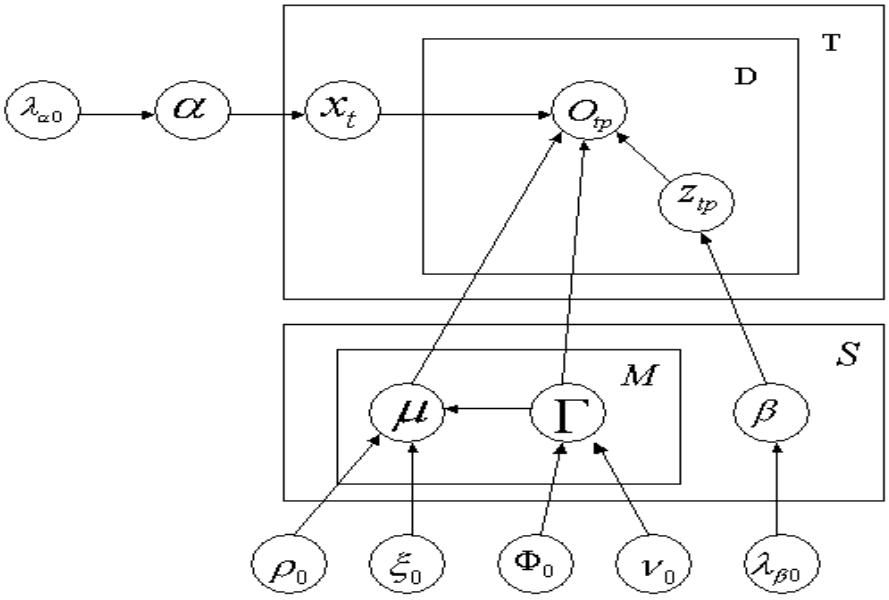

Figure 1: Direct graph that represent the Bayesian model for speaker clustering; x_t and z_{tp} are hidden variables. The box indicates that the elements inside must be repeated a number of times equal to the value in the high-right corner.

row of table 2. For the first three files the system outperforms the baseline while performance on the last files are almost identical to the baseline. Let us analyze results file by file:

- File 1: final cluster number is higher than real speaker number; anyway extra clusters are clusters where different non-speech events are organized. Final clustering results in higher acp and asp respect to the baseline system.
- File 2: final cluster number is close to real speaker number. This file is almost composed of speech events. Final asp and acp are very high.
- File 3: again final cluster number is close to real speaker number and final score is very high (asp = 0.9)
- File 4: in this case many speakers are clustered into the same cluster. This is probably due to the fact that our pruning factor Φ_0 is too high and to the fact that in this file there are many speakers that speak only for a short time. Anyway final system performance in term of acp and asp is comparable to the baseline system.

6. Conclusion and future works

In this paper we applied successfully Variational Bayesian Learning to unsupervised speaker clustering. Tests on the NIST 1996 HUB-4 evaluation data show that VB learning can outperform the baseline system. Actually some considerations must be done. First of all the system would definitely benefit from a preliminary discrimination between speech and non-speech because non-speech events often disturb clustering. Then in the approach we followed we let the system prunes itself up to the right cluster number; this technique is actually very sensitive to local maxima. A possible solution consists in overclustering the data (using a low value of Φ_0 that will keep more clusters) and then trying to merge clusters using Variational Bayesian bound (i.e. expression 14) as a measure.

7. 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", Advances in Neural Information Processing 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] O.-W. Kwon, T.-W. Lee, K. Chan, "Application of variational Bayesian PCA for speech feature extraction," Proc. ICASSP 2002, Orlando, FL, pp. I-825–I-828, May 2002.

[7] Somervuo P., "Speech modeling using Variational Bayesian mixture of gaussians",Proc ICSLP 2002.

[8] MacKay D.J.C. "Local Minima, symmetry breaking and model pruning in variational free energy minimization"

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

[10] MacKay D.J.C., "Ensemble Learning for Hidden Markov Models"

[11] 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

[12] 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

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

[14] Penny W.,"Kullback-Liebler divergences of Normal,Gaussian, Dirichlet and Wishart densities",Wellcome Department of Cognitive Neurology, 2001

A. Free energy

We show in this section that is possible to derive a close form for the variational free energy (14) when we consider a model like (4). The importance of a close form for the free energy expression consists, as previously described, in the fact that it is equivalent to a model selection criterion that can be used instead of other model selection criterion (e.g. BIC, MML, etc.).

Let us re-write expression (14) for the model we are considering:

$$F(\theta,\gamma) = \int d\theta d\gamma q(\gamma)q(\theta)log[p(O,\gamma,\theta)/q(\gamma)q(\theta)] = < log\ \frac{p(O,\gamma|\theta)}{q(\gamma)} >_{\gamma,\theta} - D[q(\theta)||p(\theta)] \quad (42)$$

where accordingly to our previous discussion, hidden variable set $\gamma = \{\gamma_{z_{tp}|x_t}\ ,\ \gamma_{x_t}\}$ consists of two variables: one referred to the cluster and the other referred to the component.

Considering the factorization $p(O,\gamma|\theta) = p(O|\gamma,\theta)p(\gamma|\theta)$ we can rewrite (42) as sum of three different terms:

$$F(\theta,\gamma) = \int d\theta d\gamma q(\gamma)q(\theta)[log(p(O|\gamma,\theta)) + log(p(\gamma|\theta))] + - \int d\theta d\gamma q(\gamma)q(\theta)logq(\gamma) - D[q(\theta)||p(\theta)] \quad (43)$$

Considering the fact that $q(\gamma_{z_{tp}} = i, \gamma_{x_t} = j) = q(\gamma_{x_t} = j)\, q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j)$ and considering the same notation as before:

$$\gamma_{z_{tp}=i|x_t=j} = q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j) \quad (44)$$

$$\gamma_{x_t=j} = q(\gamma_{x_t} = j) \quad (45)$$

Hidden variables are actually discrete variables. Coming back to expression (43) we will consider separately the three terms.

- the first term is:

$$\int d\theta d\gamma q(\gamma)q(\theta)[log(p(O|\gamma,\theta)) + log(p(\gamma|\theta))] \quad (46)$$

Because of the fact hidden variables are actually discrete variables, integral w.r.t. γ becomes a sum over states and mixtures. Let us explicit expression (46) w.r.t T,D and hidden variables:

$$\sum_{t=1}^{T}\sum_{p=1}^{D}\sum_{j=1}^{S}\sum_{i=1}^{M} \gamma_{z_{tp}=i|x_t=j}\gamma_{x_t=j} \int d\theta q(\theta)[log(p(O_{tp}|,\theta\gamma_{z_{tp}=i,x_t=j})) + log(p(\gamma_{z_{tp}=i,x_t=j})|\theta))] \quad (47)$$

Considering now the factorization $p(\gamma_{z_{tp}=i,x_t=j}|\theta) = p(\gamma_{x_t=j}|\theta)p(z_{tp} = i|x_t = j,\theta) = \alpha_j\beta_{ij}$ and using the previously defined quantity $\tilde{\gamma}^*_{z_{tp}=i|x_t=j}$, it is possible to rewrite (47):

$$\sum_{t=1}^{T}\sum_{j=1}^{S}\gamma_{x_t=j}[log\ \tilde{\alpha}_j + \sum_{p=1}^{D}\sum_{i=1}^{M}\gamma_{z_{tp}=i|x_t=j}\ log\ \tilde{\gamma}^*_{z_{tp}=i|x_t=j}] \quad (48)$$

All elements in (48) are explicit and known.

- Let us now consider the second term in (43):

$$\int d\theta d\gamma q(\gamma)q(\theta)logq(\gamma) = \int q(\gamma)logq(\gamma)d\gamma \quad (49)$$

Let us explicit this expression w.r.t. time, duration and hidden variables. The result is:

$$\sum_{t=1}^{T}\sum_{p=1}^{D}\sum_{j=1}^{S}\sum_{i=1}^{M}\gamma_{x_t=j}\gamma_{z_{tp}=i,x_t=j}\ log\,[\gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}] = = \sum_{t=1}^{T}\sum_{j=1}^{S}\{\gamma_{x_t=j}[log\ \gamma_{x_t=j} + \sum_{p=1}^{D}\sum_{i=1}^{M}\gamma_{z_{tp}=i|x_t=j}log\ \gamma_{z_{tp}=i|x_t=j}]\} \quad (50)$$

Again in (50) all terms are explicit and known.

- The last term to consider is the KL divergence between posterior distributions and prior distributions. Parameter distributions are Dirichlet distribution, Normal distribution and Wishart distribution defined in (17-18). Because of independence between parameter distributions, it is possible to write:

$$
\begin{aligned}
&D[q(\theta)||p(\theta)] = D(Dir(\lambda_{\alpha j})||Dir(\lambda_{\alpha 0})) \\
&+ \sum_j D(Dir(\lambda_{\beta ij})||Dir(\lambda_{\beta 0})) \\
&+ \sum_j \sum_i D(N(\rho_{ij}, \xi_{ij}\nu_{ij}\Phi_{ij}^{-1})||N(\rho_0, \xi_0\nu_{ij}\Phi_{ij}^{-1})) \\
&+ \sum_i \sum_j D(W(\nu_{ij}, \Phi_{ij})||W(\nu_0, \Phi_0)) \qquad (51)
\end{aligned}
$$

  A close form for all KL divergence in 51 can be found (a useful summary can be found in [14]).

In this appendix we finally show how it is possible to compute a close form for the variational free energy.