Regroupement de modèles de locuteurs par méthode Bayesienne variationnelle

Applications à la sélection de modèles par méthode variationnelle pour l'indexation audio

Fabio Valente, Christian Wellekens

Institut Eurecom
2229 Routes des Crêtes - BP193
06904 Sophia-Antipolis, France
fabio.valente@eurecom.fr
christian.wellekens@eurecom.fr

RÉSUMÉ. *Dans cet article, on étudie l'utilisation des méthodes variationnelles Bayesiennes pour le regroupement de locuteurs. Les méthodes variationnelles Bayesiennes (aussi connues sous le nom d'apprentissage d'ensemble) sont des méthodes approximées qui offrent un cadre totalement Bayesien pour l'apprentissage des modèles. Des techniques classiques telles que le Maximum a posteriori (MAP) ou la Vraisemblance maximale (ML) peuvent être considérées comme cas particuliers des méthodes variationnelles Bayesiennes (VB). En outre VB permet l'apprentissage simultané des paramètres et la sélection des modèles. Des expériences sur une tâche de regroupement de locuteurs pour les Nouvelles Radiodiffusées (Broadcast News) montrent que l'apprentissage variationnel Bayesien surclasse les méthodes classiques dans cette application.*

ABSTRACT. *This paper aims at investigating the use of Variational Bayesian methods for speaker clustering purposes. Variational Bayesian methods (a.k.a. Ensemble learning) are approximated methods that allow a fully Bayesian framework for model learning. Classical techniques like Maximum a Posteriori (MAP) or Maximum Likelihood (ML) can be seen as special cases of VB methods. Furthermore VB allows parameter learning and model selection at the same time. Experiments on a speaker clustering task for Broadcast News show that Variational Bayesian Learning outperforms classical methods for this task.*

MOTS-CLÉS : *Variationnel, Bayesien, Apprentissage d'ensemble, Sélection de modèles, Regroupement, Estimation/Maximisation (EM)*

KEYWORDS: *Variational, Bayesian, Ensemble learning, Model selection, Clustering, Estimation-Maximization*

1. Introduction

[1] La sélection de modèles est un problème de première importance dans les applications d'apprentissage de machines. Dans différentes applications sur des données réelles avec des modèles de structure inconnue, on fait une hypothèse sur le modèle avant de commencer l'apprentissage proprement dit. Si le modèle choisi ne correspond pas à la structure des données expérimentales, l'efficacité de l'apprentissage est généralement sérieusement affectée. Il est donc nécessaire de mettre en oeuvre des techniques qui choisissent le modèle le mieux adapté aux données.

Le cadre probabiliste est très largement utilisé pour la sélection de modèles. On y considère les probabilités de différents modèles et on considère que le meilleur modèle est celui qui maximise la probabilité pour les données observées c'est à dire plus formellement qui maximise $P(m|D)$ pour un modèle m et un ensemble d'observations D. L'estimation des probabilités des modèles peut être paramétrique ou non ; dans la plupart des problèmes sur des données réelles, l'hypothèse paramétrique est souvent préférée pour sa facilité de mise en oeuvre. Dans ce cas, l'estimation de la probabilité du modèle peut être obtenue en marginalisant tous les paramètres du modèle. En fonction de la complexité du modèle, l'intégration peut ne pas être explicitable et l'on doit se résoudre à utiliser des méthodes approximatives. Les plus courantes (par exemple [SCH 78]) ne conviennent parfois pas aux applications considérées et requièrent une mise au point heuristique pour être efficaces. Dans cette communication, nous considérons un nouveau type de méthodes approximatives appelées Apprentissage Variationnel (connu aussi comme Apprentissage d'ensemble) qui offre une solution sous forme explicite mais approximative du problème de l'intégration des paramètres. La clé des méthodes variationnelles est le remplacement des distributions réelles mais inconnues des paramètres par des distributions approximées (distributions variationnelles) qui permettent de traiter la solution analytiquement. Evidemment l'efficacité de cette approche dépend de la qualité de ces distributions approximées.

Dans cette communication, nous étudions l'usage des techniques variationnelles dans une application d'indexation audio où la sélection des modèles est le problème principal. Le problème de l'indexation audio consiste à regrouper les parties d'un enregistrement possédant des caractéristiques semblables. En particulier, nous considérons ici le cas où les données provenant d'un même locuteur doivent être regroupées. Le problème de la sélection du modèle est central dans de telles applications car le nombre de groupes (locuteurs) n'est généralement pas connu a priori et doit être estimé à partir des données. L'approche la plus courante de ce problème utilise une approximation très grossière de l'intégrale pour la sélection du modèle ([CHE 98],[TRI 99]) qui n'est valable qu'asymptotiquement. Afin d'obtenir des résultats acceptables si le volume de données disponible est limité, on doit procéder à un ajustement heurisitique du critère de sélection du modèle. Cela cause souvent des problèmes sérieux de mise au point qui affectent profondément le résultat final. Les méthodes variation-

1. Fabio Valente est fi nancé par une bourse de thése MESR (Ministére de l'Enseignement Supérieur et de la Recherche)

nelles n'exigent aucune mise au point heuristique et ne sont pas restreintes à de très grandes bases de données : pour cette raison, elles sont plus efficaces que le BIC.

L'article est organisé comme suit : en section 2, nous décrivons la tâche d'indexation en locuteurs, en section 3 nous exposons quelques concepts généraux sur le problème de la sélection du modèle et l'approximation BIC, en section 4 nous introduisons l'apprentissage Bayesien variationnel (VB) et la sélection de modèles, en section 5 nous comparons les apprentissages VB, MAP et ML pour des modèles à mélange de Gaussiennes et finalement en section 6 nous décrivons des expériences sur la base de données Broadcast News.

2. Regroupement de locuteurs pour l'indexation audio

Le regroupement de locuteurs ou plus généralement le regroupement des données audio est une tâche de première importance dans les nombreux différents systèmes de traitement de l'audio ; par exemple, les systèmes de recherche d'information ou de reconnaissance de la parole tirent grand avantage d'une présegmentation de l'enregistrement audio en régions de caractéristiques homogènes. Plus particulièrement dans cet article, nous nous intéressons au regroupement de segments provenant d'un seul et même locuteur et subsidiairement à la séparation de la parole et de la non-parole. Ce problème a fait l'objet de nombreuses études dans le passé et la topologie la plus aboutie consiste à faire l'hypothèse que l'audio est produite par une modèle de Markov (HMM) ergodique (c'est à dire totalement connecté) dans lequel chaque état représente une classe acoustique (par example, un locuteur ou un type de bruit)(voir [OLS95]). Mais d'autres modèles comme les transformations autoorganisées (SOM)[LAP 03] ou la quantification vectorielle (VQ) furent utilisés avec succès. D'autre part, le nombre de classes acoustiques n'est généralement pas connu a priori et doit être estimé à partir des données. Généralement ce problème est formulé comme un problème de sélection de modèles, dans lequel différents modèles possèdent un nombre différent de classes. Le critère de sélection de modèles utilisé communément dans ces applications est le critère d'information Bayesien (BIC) [SCH 78]. BIC est un critère de sélection approximé de modèles raisonnables pour les problèmes de segmentation audio [CHE 98],[TRI 99] mais pour être efficace dans des applications réelles il doit être modifié de façon heuristique par l'ajustement d'un seuil. Une tentative d'élimination de cette dépendance au seuil est proposée par [AJM 02] mais la solution ne règle pas le problème pour les grandes bases de données ni celui du surentrainement. Nous verrons plus loin que les techniques variationnelles Bayesiennes constituent une solution raisonnable à ces deux problèmes.

Décrivons à présent la topologie utilisée dans cet article. Nous faisons l'hypothèse que le fichier audio peut être modélisé par S classes acoustiques. Le modèle considéré est un HMM à S états totalement connectés qui chacuns représentent un type d'audio. La probabilité d'émission de chaque état est un mélange de Gaussiennes. Le modèle est ensuite simplifié en faisant une hypothèse sur les probabilités de transition. Soit

α_{rj} la probabilité de transition de l'état r vers l'état j, alors nous faisons l'hypothèse que la probabilité ne dépend pas de l'état d'origine c'est à dire :

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

Sous cette hypothèse, le HMM ergodique avec des probabilités d'émission GMM peut être modélisé comme un mélange de Gaussiennes. Donc si nous définissons par $P(O_t|s_j)$ la probabilité d'observer O_t sur un état s_j à l'instant t, nous pouvons écrire $P(O_t) = \sum_{j=1}^{S} \alpha_j P(O_t|s_j)$ où α_j est la probabilité de transition indépendante de l'état initial. Afin d'obtenir une estimation robuste et d'éviter des solutions dispersées, nous imposons une contrainte de durée minimale de D trames sur chaque état comme en [LAP 03]. En [LAP 03], on montre qu'une valeur $D = 100$ est suffisante pour capturer les caractéristiques d'un locuteur et construire un modèle de locuteur robuste. Ainsi l'observation O_t est composée de D trames consécutives. Pour compléter notre modèle, définissons les paramètres GMM comme coefficients de mélange β_{ij}, moyennes μ_{ij} et variances Γ_{ij}, où $i = 1, ..., M$ avec M le nombre de Gaussiennes. Il est ainsi possible d'écrire la log-vraisemblance d'une séquence O_t avec $t = 1, ..., T$:

$$log\, P(O) = \sum_{t=1}^{T} log \sum_{j=1}^{S} \alpha_j \{ \prod_{p=1}^{D} \sum_{i=1}^{M} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij}) \} \tag{2}$$

Il est facile d'identifier dans le modèle (2) deux sortes de variables cachées : les unes désignent l'état (locuteur) et les autres la Gaussienne qui produit l'information. L'optimisation du modèle (2) peut être obtenue en utilisant l'algorithme Estimation-Maximisation (EM) (voir [DEM 77]). Nous donnerons plus loin des détails concernant le processus d'optimisation de l'EM classique avec le VB-EM. Dans cette topologie, nous supposons connu le nombre de classes S mais ceci est rarement le cas dans les applications réelles. Ainsi apparait la nécessité d'un critère de sélection des modèles permettant d'estimer le nombre exact de classes à partir des données ; dans la section suivante, nous discuterons le problème de la sélection des modèles d'un point de vue théorique.

3. Sélection des modèles

Considérons un ensemble de données D et un ensemble de modèles $Model = \{m_j\}$. Dans un cadre probabiliste, le modèle qui s'accorde le mieux aux données est celui qui maximise $P(m|D)$ c'est à dire la probabilité du modèle pour un ensemble de données D. En appliquant la loi de Bayes, on trouve :

$$P(m|D) = \frac{P(D|m)P(m)}{P(D)} \tag{3}$$

où $P(D) = \sum_m P(D|m)P(m)$ ne dépend pas de m. Si la probabilité a priori du modèle $P(m)$ est uniforme (c'est à dire qu'aucune information a priori n'est disponible

pour le modèle), le meilleur modèle est alors celui qui maximise l'evidence des données c'est à dire $P(D|m)$. Considérons à présent un modèle paramétrique (quoique (3) reste valable pour une estimation non paramétrique) et désignons par θ l'ensemble des paramètres du modèle et par $p(\theta|m)$ les distributions des paramètres du modèle. Sous ces hypothèses et notations, l'évidence des données peut être obtenue en marginalisant les paramètres du modèle par rapport à leurs distributions, c'est à dire :

$$p(D|m) = \int p(D|\theta, m)\, p(\theta|m) d\theta \tag{4}$$

L'expression (4) est connue comme vraisemblance marginale. Selon le choix du modèle m, elle peut être calculée sous une forme explicite.

Cette vraisemblance marginalisée (4) présente la propriété de sélection de modèles en pénalisant les modèles qui contiennent des degrés de liberté excédentaires inutiles à la modélisation des données expérimentales. Cette propriété est aussi connue sous le nom de "rasoir d'Occam" (voir [MKAY 95]). L'idée est que les modèles comportant un plus grand nombre de paramètres peuvent modéliser un plus grand nombre de données avec en conséquence une densité de probabilité $p(\theta|m)$ plus répartie sur le domaine des paramètres. Pour mieux comprendre ce phénomène, considérons l'exemple simple décrit par [MKAY 95] d'un modèle uni-dimensionnel avec une probabilité a priori gaussienne sur le paramètre θ et une variance gaussienne $\sigma_{\theta|D}$ du paramètre θ. Si θ_{max} est l'estimation des paramètres qui maximise la densité des paramètres, nous pouvons approximer la vraisemblance marginale comme suit :

$$p(D|m) \approx p(D|\theta_{max}, m) \times p(\theta_{max}|m)\sigma_{\theta|D} \tag{5}$$

Le second facteur dans (5) est connu sous le nom de "facteur d'Occam" ; si on considère que $P(\theta|m)$ est uniforme sur une grande gamme de σ_θ, on peut écrire $P(\theta_{max}|m) = 1/\sigma_\theta$ et le facteur d'Occam devient $\sigma_{\theta|D}/\sigma_\theta$. Intuitivement ce facteur pénalise les modèles qui ont le plus de paramètres c'est à dire qui ont une grande valeur de σ_θ par rapport à ceux qui en ont peu. En ce sens, la vraisemblance marginalisée contient un terme de pénalité qui croit avec le nombre de paramètres et explique intuitivement pourquoi la vraisemblance marginalisée est une bonne mesure pour la sélection des modèles.

Malheureusement pour beaucoup de modèles utilisés actuellement comme les modèles de Markov cachés (HMM) ou les modèles à mélange de Gaussiennes (GMM), aucune forme explicite de la vraisemblance marginalisée ne peut être déduite à cause de l'usage de variables cachées. Un choix courant pour contourner cette difficulté consiste à ignorer l'intégrale dans (4) ; ainsi l'estimation des paramètres Maximum a Posteriori (MAP) classique peut être retrouvée, soit :

$$\theta_{MAP} = argmax_\theta\, p(D|\theta, m)p(\theta|m) \tag{6}$$

L'approche MAP est possible mais n'est pas totalement Bayesienne car elle est une approximation ponctuelle qui prend en compte les maxima des paramètres au lieu de leur complète distribution. En utilisant une métaphore de la mécanique , MAP considère

la “densité” et non la “masse” de la distribution. MAP devient une approximation raisonnable lorsque la distribution des paramètres est extrêmement pointue et la “masse” de la distribution concentrée autour du maximum mais de façon générale, on pourrait négliger d'importantes contributions à l'intégrale.

3.1. *Sélection approximative de modèles*

L'impossibilité de traiter (3) rend nécessaire l'usage de techniques d'approximation pour procéder à la sélection de modèles. Selon l'application, des techniques numériques d'estimation comme la simulation Monte Carlo peuvent fournir une solution mais lorsque les modèles consistent en milliers de paramètres libres, elles deviennent impraticables. C'est pourquoi de nombreuses approximations de l'intégrale ont été proposées ; dans cet article, nous considérons la plus répandue qui constitue l'état de l'art d'aujourd'hui dans de nombreux systèmes d'indexation audio : le Critère d'Information Bayesien (BIC).

BIC a été décrit pour la première fois par Schwartz [SCH 78] ; Il peut être dérivé d'une approximation Laplacienne d'une intégrale Bayesienne. L'approximation de Laplace est une approximation gaussienne locale autour de l'estimateur MAP $\hat{\theta}$ des paramètres pour le cas limite d'un grand nombre de données. Désignons par N, la cardinalité de l'ensemble de données D et par p le nombre de paramètres libres θ ; alors le BIC est :

$$ln\, p(D|m)_{BIC} = ln\, p(D|\hat{\theta}, m) - \frac{p}{2} ln\, N \qquad (7)$$

L'expression (7) a une justification intuitive : un modèle plus sophistiqué c'est à dire avec beaucoup de paramètres conduira à un terme de pénalité $\frac{p}{2} ln\, N$ plus élevé par rapport à un modèle comportant un plus petit nombre de paramètres libres. BIC est une approximation très grossière de l'intégrale Bayesienne mais présente beaucoup d'avantages pour le traitement du problème car il peut être calculé tant que l'estimation MAP du modèle est disponible. Comme indiqué plus haut, BIC est basé sur l'hypothèse d'une très grande base de données rarement atteinte dans les cas pratiques. Pour contourner cette restriction et rendre le critère plus efficace dans toutes les situations, le terme de pénalité est généralement multiplié par un “seuil” λ qui est déterminé de manière heuristique selon l'application. Par example, dans le cas de l'indexation audio, une amélioration importante dans la sélection des modèles est obtenue en modifiant manuellement le terme de pénalité [TRI 99] ou en utilisant des données de validation pour trouver le λ optimal pour un ensemble de données [DELA 00].

4. Apprentissage variationnel

L'apprentissage variationnel est une technique relativement nouvelle basée sur l'utilisation d'une distribution approximée en lieu et place de la distribution réelle afin de rendre la tâche calculable. Considérons la log-vraisemblance marginalisée. La méthode variationnelle fait l'hypothèse que la distribution a posteriori inconnue $p(\theta|m)$

peut être approximée par une autre distribution $q(\theta|D,m)$ qui est la distribution variationnelle a posteriori (plus simplement la distribution variationnelle) dérivée des données. En utilisant l'inégalité de Jensen, on peut écrire :

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

$F(\theta)$ est appelée énergie libre variationnelle ou énergie d'apprentissage d'ensemble et est une borne inférieure de la log-vraisemblance marginalisée ; l'apprentissage variationnel vise la maximisation de l'énergie libre par rapport aux distributions variationnelles au lieu de l'impraticable log-vraisemblance marginalisée. Un point clé dans ce cadre est le choix de la forme de la distribution $q(\theta|D,m)$ qui doit être suffisamment proche de la distribution réelle mais inconnue des paramètres et cependant de forme praticable. La différence entre la log-vraisemblance marginalisée et l'énergie libre est :

$$log\, p(D|m) - F(\theta) = KL(q(\theta|D,m)||p(\theta|D,m)) = -\int q(\theta|D,m) log \frac{p(\theta|D,m)}{q(\theta|D,m)} d\theta \tag{9}$$

L' équation (9) signifie que l'apprentissage variationnel minimise en réalité la distance entre la distribution vraie a posteriori et les distributions variationnelles a posteriori. Dans le cas limite, si $q(\theta|D,m) = p(\theta|D,m)$, l'énergie libre est égale à la log-vraisemblance marginalisée. On voit donc la nécessité de choisir $q(\theta|D,m)$ aussi proche que possible de $p(\theta|D,m)$.

4.1. *Apprentissage avec variables cachées*

Une propriété très attrayante de l'apprentissage variationnel est sa capacité de traiter des variables cachées. En fait les variables cachées peuvent être simplement vues comme d'autres variables stochastiques (comme les paramètres) avec leurs distributions propres. En pratique dans certains systèmes d'apprentissage variationnel, il n'y a pas de différence dans la manière de considérer les paramètres et les variables cachées (par exemple voir [BIS 03]). Définissons X l'ensemble des variables cachées ; il est possible d'introduire une distribution variationnelle jointe pour les variables cachées et les paramètres $q(X,\theta|D,m)$ et en appliquant l'inégalité de Jensen, on obtient :

$$log\, p(D|m) = \int p(D,X,\theta|m)\, d\theta\, dX \geq \int q(X,\theta|D,m) log\, \frac{p(D,X,\theta|m)}{q(X,\theta|D,m)} = F(\theta,X) \tag{10}$$

Cette fois, la différence entre la vraie log-vraisemblance marginalisée et l'énergie libre prendra en considération les variables cachées également et on peut écrire :

$$log\, p(D|m) - F(\theta,X) = KL(q(\theta,X|D,m)||p(\theta,X|D,m)) = -\int q(\theta,X|D,m) log \frac{p(\theta,X|D,m)}{q(\theta,X|D,m)} \tag{11}$$

Ainsi, le même cadre élégant peut traiter les paramètres et les variables cachées. Maintenant, une approximation supplémentaire doit être faite afin de conserver un formalisme praticable : considérer la distribution jointe $q(\theta,X|D,m)$ peut être une tâche

prohibitivement lourde lorsque le nombre de variables cachées est très grand et conduit à de sérieux problèmes de mise en oeuvre. Pour cette raison, on supposera ces variables indépendantes afin de pouvoir factoriser la distribution jointe :$q(\theta, X|D, m) = q(\theta|D, m)q(X|D, m)$. Sous cette hypothèse, la distribution variationnelle a posteriori optimale qui maximise l'énergie libre peut être trouvée en utilisant un algorithme de type EM [ATT 00]. En dérivant simplement l'énergie libre par rapport à $q(\theta|D, m)$ et $q(X|D, m)$, on obtient la formule de mise à jour itérative qui convergera vers un maximum (local) de l'énergie libre. Le système d'équations consiste en un étape de type E :

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

et une étape de type M est :

$$q(\theta|D, m) \propto e^{<log\, p(D,X|\theta,m)>_X} p(\theta|m) \tag{13}$$

où $< \cdot >_z$ désigne la moyenne vis à vis de z. Les détails et la justification de cet algorithme peuvent être trouvés dans [ATT 00]. Cet algorithme de type EM n'estime pas les paramètres (contrairement au MAP) mais bien les distributions des paramètres. Sous l'hypothèse de factorisation il est possible de récrire l'énergie libre comme suit :

$$\begin{aligned} F(\theta, X) &= \int d\theta dX q(X|D, m)q(\theta|D, m)log[\frac{p(D, X, \theta|m)}{q(X|D, m)q(\theta|D, m)}] \\ &= < log\, \frac{p(D, X|\theta, m)}{q(X|D, m)} >_{X,\theta} - KL[q(\theta|D, m)||p(\theta|m)] \end{aligned} \tag{14}$$

L'énergie libre est composée ainsi de deux termes : le premier dépend des données et des distributions variationnelles (à la fois sur les paramètres et les variables cachées) et le second qui est la divergence Kullback-Leibler (KL) entre la distribution variationnelle des paramètres et la distribution a priori des paramètres $p(\theta|m)$. On sait que $KL[q(\theta|Y, m)||p(\theta|m)] \geq 0$ avec l'égalité si $q(\theta|Y, m) = p(\theta|m)$; ce terme agit comme une pénalité sur les modèles qui présentent le plus de paramètres. En fait les modèles qui contiennent beaucoup de paramètres vont avoir une divergence KL plus élevée et il n'est pas sans intérêt d'observer que cette pénalité ne dépend pas seulement du nombre de paramètres (comme dans le BIC) mais prend en compte la divergence entre les distributions a posteriori et a priori. On peut montrer que dans l'hypothèse limite d'un grand nombre de données, ce terme de pénalité converge vers la pénalité BIC [ATT 99]. Intuitivement l'énergie libre est une quantité intéressante pour la sélection de modèles comme nous le montrons de façon plus rigoureuse en section 4.2.

Les distributions variationnelles a posteriori sur les paramètres estimés en (13) sont le produit d'un facteur dépendant des données et de la distribution a priori. On sait que si la distribution a priori appartient à une famille conjuguée, alors la distribution a posteriori aura la même forme que la distribution a priori. Cette propriété suggère un choix intéressant pour la distribution variationnelle : le choix d'une distribution dans la famille conjuguée exponentielle simplifiera considérablement la complexité de l'étape M.

Maintenant qu'une solution efficace pour les variables cachées a été introduite, l'apprentissage totalement Bayesien dans de nombreux modèles auparavant impraticables est possible. Par exemple, l'apprentissage variationnel Bayesien des HMM a été introduit pour la première fois par [MKAY 97] et pour les mélanges de gaussiennes par [ATT 00]. L'application de l'algorithme Bayesien variationnel EM (VBEM) à un modèle général est étudiée par [BEAL 02]. L'algorithme VBEM peut être dérivé pour les modèles dits conjugués-exponentiels c'est à dire répondant aux deux conditions :

– La vraisemblance complète (variables cachées et paramètres) des données appartient à la famille exponentielle

– La distribution a priori des paramètres est conjuguée à la vraisemblance des données complète.

Un grand nombre de modèles bien connus satisfont á ces deux conditions : modèles à mélange de Gaussiennes, HMM, analyse factorielle, analyse par composantes principales (PCA), etc.

4.2. *Sélection des modèles en utilisant l'énergie libre*

Nous avons souligné ci-avant comment l'énergie libre variationnelle (14) contenait une sorte de pénalité qui charge les modèles plus complexes par rapport aux modèles simples. Dans cette section, nous formulons la sélection de modèles dans un un cadre plus rigoureux. Considérons donc la log-vraisemblance marginalisée obtenue en intégrant sur tous les modèles possibles $log\, p(D) = \sum_m p(D|m)p(m)$ et introduisons une probabilité variationnelle a posteriori sur les modèles $q(m)$. Appliquant à nouveau l'inégalité de Jensen, on peut écrire :

$$log\, p(D) = \sum_m p(D|m)p(m) \geq \sum_m q(m)[F_m + log\, \frac{p(m)}{q(m)}] \qquad (15)$$

où $p(m)$ est une probabilité a priori sur le modèle et F_m est l'énergie libre pour le modèle m. A nouveau une borne sur la log-vraisemblance marginalisée en résulte. Dérivant par rapport à $q(m)$ et explicitant, on trouve pour la distribution variationnelle optimale sur les modèles :

$$q(m) \propto exp\{F_m\}p(m) \qquad (16)$$

qui signifie que la probabilité a posteriori optimale est proportionnelle à l'exponentielle de l'énergie libre fois la probabilité a priori. Si cette dernière est uniforme, la distribution variationnelle a posteriori dépend seulement de l'énergie libre ce qui signifie que l'énergie libre peut être utilisée en lieu et place de la log-vraisemblance marginalisée réelle pour la sélection des modèles.

Cette solution est assurément plus intéressante que la solution BIC pour la sélection de modèles : lorsque BIC est utilisé, un terme de pénalité est ajouté au modèle entrainé avec un critère ignorant complètement le terme de pénalité alors qu'ici ce

terme est entrainé d'une certaine manière conjointement avec le modèle. Bien sûr, il faut mettre en évidence que la sélection de modèles VB est biaisée vers les modèles simples. Examinons à nouveau un exemple cité en [BEAL 03] : étant donné deux modèles gaussiens à mélange de Gaussiennes avec S et $S+1$ composantes, supposons que chaque composante contribue d'une quantité fixe KL_s à la différence entre la log-vraisemblance marginale $L(S)$ et l'énergie libre $F(S)$. Sous cette hypothèse, la différence entre L(S) et L(S+1) différera de la différence entre F(S) et F(S+1) soit :

$$\begin{aligned} L(S+1) - L(S) &= [F(S+1) + (S+1)KL_S] - [F(S) + SKL_S] \\ F(S+1) - F(S) + KL_S &\neq F(S+1) - F(S). \end{aligned} \quad (17)$$

En d'autres mots, utiliser l'énergie libre pour sélectionner un modèle est un estimateur biaisé vers les modèles les plus simples. Quoiqu'il en soit, nous trouverons expérimentalement que l'énergie libre est sans conteste plus efficace que le BIC pour la sélection de modèle.

4.3. *Probabilité prédictive*

Les méthodes variationnelles Bayesiennes peuvent aussi être utilisées pour faire de la prédiction sur des variables inobservées. Supposons en effet que nous souhaitions prédire des données non-observées U à partir de données observées O. Théoriquement nous devrions utiliser :

$$p(U|O) = \int p(\theta|O)p(U|\theta)d\theta \quad (18)$$

Cependant, selon le modèle, l'équation (18) n'est pas toujours explicitable et une fois encore, une approximation est nécessaire. Tout d'abord, les distributions a posteriori sur les paramètres θ peuvent être approximées par la distribution variationnelle a posteriori c'est à dire $p(\theta|S) \approx q(\theta|S)$. Cette hypothèse résoud le problème relatif à la distribution des paramètres mais laisse l'intégrale toujours impraticable. Dans ce cas, (18) peut être approximée à nouveau par l'inégalité de Jensen comme dans (8). Pour plus d'information sur les approximations possibles avec différents modèles, on consultera [BEAL 03]. Une autre solution très simple pour calculer la probabilité des données non-observées est d'utiliser des moments de distributions variationnelles comme paramètres du modèle [MKAY 97] ; même si c'est une approximation très grossière, les résultats sont cependant comparables avec ceux obtenus par les techniques d'estimation des paramètres MAP et ML.

5. VB, MAP et ML pour mélange de Gaussiennes

Dans cette section, nous passons en revue les techniques d'entrainement des modèles à mélange de Gaussiennes qui sont au coeur des méthodes d'indexation des

locuteurs par apprentissages VB, MAP et ML et nous mettons en évidence leurs similitudes et les différences. Considérons donc un GMM avec des coefficients du mélange β_i, moyennes μ_i et variances Γ_i où $i = 1, ..., M$ avec M le nombre de Gaussiennes. Supposons que l'ensemble d'entrainement consiste en $\{O_t\}$ observations. L'algorithme EM classique [DEM 77] consiste en une étape E au cours de laquelle les variables cachées sont estimées et une étape M dans laquelle on met à jour les paramètres. Désignons par $\{x_t\}$ l'ensemble des variables cachées, l'algorithme EM consiste en l'alternance des étapes E (19) et M (20) :

$$\gamma_{x_t=j} = P(x_t = j) = \frac{\beta_i N(O_t|\mu_i, \Gamma_i)}{\sum_i \beta_j N(O_t|\mu_j, \Gamma_j)} \quad (19)$$

$$\beta_j = \frac{\sum_{t=1}^{T} \gamma_{x_t=j}}{T} \quad \mu_i = \frac{\sum_{t=1}^{T} \gamma_{x_t=j}\, O_t}{\sum_{t=1}^{T} \gamma_{x_t=j}} \quad \Gamma_i = \frac{\sum_{t=1}^{T} \gamma_{x_t=j}\, (O_t - \mu_i)^T (O_t - \mu_i)}{\sum_{t=1}^{T} \gamma_{x_t=j}} \quad (20)$$

Appliquant itérativement (19) et (20), la log-vraisemblance des données convergera vers un maximum local. L'apprentissage ML est sérieusement dégradé par les problèmes de sur-entrainement et manque de robustesse lorsque la quantité de données est faible par rapport au nombre de paramètres. En fait si un vecteur seulement est associé à une composante Gaussienne, la matrice de covariance devient singulière. Pour cette raison, le déterminant de la matrice de covariance d'entrainement est généralement controlé ; dans ce cas, la mise à jour de cette composante est arrêtée. Dans les approches Bayesiennes, ce problème n'apparait pas car on utilise la distribution a priori. Considérons à présent l'apprentissage MAP de la même façon qu'il a été proposé par [GAU 94]. Tout d'abord un choix des distributions a priori sur les paramètres doit être fait. Un choix raisonnable est d'utiliser les distributions de la famille exponentielle conjuguée de sorte que les distributions a priori et a posteriori aient la même forme. Soit :

$$P(\beta_j) = Dir(\lambda_{\beta 0}) \quad P(\mu_j|\Gamma_j) = N(\rho_0, \xi_0 \Gamma_j) \quad P(\Gamma_j) = W(\nu_0, \Phi_0) \quad (21)$$

où Dir désigne une distribution de Dirichlet, N une distribution normale et W une distribution de Wishart et $\{\lambda_{\beta 0}, \rho_0, \xi_0, \nu_0, \Phi_0\}$ est l'ensemble des hyperparamètres. Nous comptons donc sur des distributions a posteriori MAP de forme semblable à celle des distributions a priori mais avec un ensemble d'hyperparamètres mis à jour $\{\lambda_{\beta j}, \rho_j, \xi_j, \nu_j, \Phi_j\}$. Ainsi pour un estimateur initial des paramètres, l'algorithme EM pour l'estimation MAP a une étape E semblable à (19). Dans l'étape M, on commence par mettre à jour les hyperparamètres a posteriori comme suit :

$$\lambda_{\beta_j} = N_j + \lambda_{\beta 0} \quad \xi_j = N_j + \xi_0 \quad \nu_j = N_j + \nu_0$$

$$\rho_j = \frac{N_j\, \mu_j + \xi_0\, \rho_0}{N_j + \rho_0} \quad \Phi_j = N_j\, \Gamma_j + \frac{N_j \xi_0 (\mu_j - \rho_0)(\mu_j - \rho_0)^T}{N_j + \rho_0} + \Phi_0 \quad (22)$$

où $N_j = \sum_t \gamma_{x(t)=j}$. Ensuite, on estime les paramètres MAP :

$$\beta_j = \frac{\lambda_{\beta_j} - 1}{\sum_j (\lambda_{\beta_j} - 1)} \quad \mu_j = \rho_j \quad \Gamma_j = (\nu_j - d)\Phi_j^{-1} \quad (23)$$

où d est la dimension des vecteurs acoustiques. Une fois de plus cet algorithme converge vers un maximum local de la fonction MAP objective. Cependant il faut attirer l'attention sur le fait que MAP n'est pas invariante à la re-paramétrisation (i.e. la représentation de distributions de paramètres) et donne une probabilité non-nulle a des modèles qui ne satisfont pas les contraintes structurelles, ce qui signifie que l'estimation des paramètres finals dépend de la forme initiale des distributions des paramètres. Par example si $\lambda_{\beta 0} \leq 1$, β_j sera ≤ 0. Généralement, pour éviter des composantes négatives, le paramètre a priori $\lambda_{\beta 0}$ est contraint par ≥ 1 même si des cas ≤ 1 ont également été étudiés [MKAY 95b]. En d'autres mots, l'estimation MAP des paramètres dépend toujours de la base [MKAY 98] ce qui est un inconvénient important puisque changer de base tout en conservant les mêmes données peut conduire à un estimateur radicalement différent. Un cas simple d'effet de re-paramétrisation est mis en évidence par [MKAY 98] où une paramétrisation softmax pour la distribution de Dirichlet est utilisée au lieu de la paramétrisation classique.

Considérons à présent l'algorithme de type EM utilisé pour l'apprentissage VB comme décrit par les équations (12) et (13). VBEM ne cherche pas à estimer les paramètres mais seulement les distributions variationnelles a posteriori des paramètres des modèles : c'est une différence fondamentale avec les autres critères d'estimation comme MAP et ML. Si des distributions a priori sont choisies dans la famille conjuguée c'est à dire comme en (21), nous obtenons des distributions variationnelles Bayesiennes a posteriori de même forme que les distributions a priori mais avec des hyperparamètres mis à jour (22). D'autre part, l'étape E est très différente de celle utilisée pour le MAP et le ML en ce sens qu'elle intègre sur les distributions des paramètres. Ainsi définissons les moments suivants pour les paramètres :

$$\begin{aligned} log\,\tilde{\beta}_j &= \Psi(\lambda_{\beta_j}) - \Psi(\sum_j \lambda_{\beta_j}) \qquad \bar{\Gamma}_j = \nu_{ij}\Phi_j^{-1} \\ log\,\tilde{\Gamma}_j &= \sum_{i=1}^{g} \Psi((\nu_j + 1 - i)/2) - log\,|\Phi_j| + d\,log 2 \end{aligned} \tag{24}$$

L'étape E aura la forme suivante :

$$\begin{aligned} \gamma^*_{x_t=j} &= \tilde{\beta}_j\,\tilde{\Gamma}_j^{1/2}\,exp\{-\frac{1}{2}(O_t - \rho_j)^T\bar{\Gamma}_{ij}(O_t - \rho_j)\}\,exp\{\frac{-d}{2\nu_j}\} \\ \gamma_{x_t=j} &= P(x_t = j) = \gamma^*_{x_t=j}/\sum_j \gamma^*_{x_t=j} \end{aligned} \tag{25}$$

où d est la dimension des vecteurs acoustiques. La conséquence de ceci sera discutée dans le cadre de l'élagage des paramètres. Considérons le cas particulier où la distribution variationnelle est une distribution de Dirac c'est à dire $q(\theta) = \delta(\theta^{'} - \theta)$. L'énergie libre se réduit à :

$$max_{Q(\theta)}F(\theta) = max_{\theta'}\int \delta(\theta - \theta^{'})log[p(Y|\theta)p(\theta)]d\theta = max_{\theta'}\,log[p(Y|\theta^{'})p(\theta^{'}))] \tag{26}$$

où le terme $\int q(\theta) log\, q(\theta) d\theta$ est négligé car il est constant. L'estimateur (26) est semblable à l'estimateur MAP mais il n'y a aucune garantie que la solution pour VB et MAP soient les mêmes parce que, comme rappelé plus haut, MAP n'est pas invariant à la paramétrisation de sorte qu'une autre paramétrisation peut conduire à des résultats complètement différents. D'autre part, ML tout comme VB est invariant vis à vis des paramètres. En d'autres termes, en adoptant une distribution a priori plate c'est à dire non informative on peut retrouver l'estimateur ML à partir de VB tandis que les résultats pour l'estimateur MAP dépendront de la paramétrisation choisie (par exemple le "-1" dans la formule de réestimation des poids Gaussiens). Choisissant une autre base pour la distribution de Dirichlet comme dans [MKAY 98] conduirait à un estimateur des paramètres complètement différent.

Considérons maintenant la distribution prédictive dérivée pour un mélange de Gaussiennes comme la distribution de données non-observées obtenue par marginalisation des paramètres GMM par rapport aux paramètres variationnels. On peut démontrer ([ATT 00]) que la distribution résultante est un mélange de distributions t-Student avec les paramètres

$$p(U|S) = \sum_{j=1}^{M} \bar{\pi}_j \; T(U|\bar{\omega}_j, \bar{\mu}_j, \bar{\Gamma}_j) \tag{27}$$

où $\bar{\pi}_j = \lambda_j / \sum_j \lambda_j$,$\bar{\omega}_j = \nu_j + 1 - d$ (degrés de liberté de la distribution t-Student), $\bar{\mu}_j = \mu_j$ et $\bar{\Gamma}_j = (\xi_j + 1)/(\xi_j \, \bar{\omega}_j) \, \Phi_j$.

5.1. *Elagage des caractéristiques*

Une autre propriété intéressante de l'apprentissage variationnel Bayesien est sa capacité d'éliminer des degrés de liberté en excès pendant l'entrainement. Comme montré plus haut dans le cadre des GMM, l'apprentissage ML crée une solution singulière lorsqu'un seul vecteur d'observation est assigné à une composante Gaussienne. Dans l'approche Bayesienne comme MAP ou VB une distribution a priori sur les paramètres est modifiée par les données ; quand la contribution des données est inexistante, la densité a posteriori est égale à la densité a priori sans problème de singularité. Une fois encore la faiblesse de l'approche MAP pour la paramétrisation doit être mise en évidence. En fait en cas d'absence de contribution des données à la distribution a posteriori, l'estimateur MAP des paramètres peut être très éloigné de la réalité si les hyperparamètres de la distribution a priori ne sont pas soigneusement choisis (c'est à dire $\lambda_{\beta 0} \geq 1$ dans le cas GMM). Le problème disparait dans le cadre VB parce qu'on n'y fait pas d'estimation explicite des paramètres mais seulement une estimation de la distribution des paramètres et dans ce cas, le résultat sera une distribution a posteriori égale à la distribution a priori (c'est à dire que les données n'apportent aucune contribution à la connaissance d'un composant résultant de la même distribution a priori). L'énergie libre (14) couvre ce cas particulier de façon très élégante : en fait, le terme $KL[q(\theta|Y,m)||p(\theta|m)]$ s'annulera simplement pour les facteurs dont la distribution a

posteriori $q(\theta|Y,m)$ est égale à la distribution a priori $p(\theta|m)$. Dans le cadre GMM, les distributions des moyennes et des covariances vont disparaitre du calcul du terme de pénalité et la seule contribution qui se maintiendra sera celle de la distribution de Dirichlet. Cette propriété remarquable permet d'éviter le sur-entrainement d'un modèle donné lorsque seulement peu de données sont disponibles. Un exemple d'élagage" automatique des caractéristiques utilisant l'apprentissage VB pour les HMM est donné en [MKAY 97]. Il faut cependant indiquer que l'élimination de degrés de liberté excédentaires (qui décrivent une région "non explorée" par les données expérimentales) peut engendrer une perte de généralité très gênante du modèle et parfois un élagage définitivement faux (voir exemple dans [MKAY 01]).

Afin de mieux comprendre comment fonctionne le regroupement VB comparé au MAP, considérons un exemple intuitif très simple. Afin de négliger le problème de la paramétrisation du MAP, considérons ici le seul problème de l'estimation des variables cachées pour une observation x_t : nous aurons respectivement pour le MAP et le VB :

$$E_{MAP} \propto N(x_t|\theta_{MAP}) \tag{28}$$

$$E_{VB} \propto exp(\int log\, N(x_t|\theta) Q(\theta)_{VB} d\theta \tag{29}$$

Récrivons l'étape E du VB en considérant explicitement l'intégrale comme une contribution de α :

$$E_{VB} = exp(\alpha\, log\, N(x_t|\theta_{MAP})) = exp(log\, N(x_t|\theta_{MAP})^{\alpha}) = N(x_t|\theta_{MAP})^{\alpha}. \tag{30}$$

La valeur du rapport α dépendra de l'étroitesse de la distribution $Q(\theta)_{VB}$ et peut être définie comme :

$$\alpha = \frac{\int log\, N(x_t|\theta) Q(\theta)_{VB} d\theta}{N(x_t|\theta_{MAP})}. \tag{31}$$

Dans le cas extrême $Q(\theta) = \delta(\theta - \hat{\theta})$, l'étape E VB se réduit à l'étape E MAP (faisant toujours l'hypothèse que MAP n'est pas invariante pour la paramétrisation), α est égal à 1 et les étapes E MAP et VB coincident.

D'autre part, lorsque $Q(\theta)$ est très plat (par exemple distribution a priori plate et très peu de données d'entrainement), $\alpha \to 0$. Considérons un simple GMM à deux composantes. Supposons que pour une initialisation donnée, l'étape E MAP fournisse $\{a, b\}$ comme estimation des variables cachées non-normalisées. Considérons la discussion précédente, l'étape VB-E fournit $\{a^{\alpha}, b^{\beta}\}$ où α, β sont des rapports comme en (31). $\{a, b\}$ doit être normalisé :

$$\pi_{MAP} = \frac{a}{a+b} = \frac{1}{1+\frac{b}{a}} \qquad \pi_{VB} = \frac{a^{\alpha}}{a^{\alpha}+b^{\beta}} = \frac{1}{1+\frac{b^{\beta}}{a^{\alpha}}} \tag{32}$$

La valeur de π_{VB} différera de celle de π_{MAP} dépendant des rapports des distributions simples α et β. Plus étroites sont ces distributions, plus proches de 1 seront α et β et plus proche sera la solution finale de la solution MAP. Dans tous les autres cas, des

solutions intermédiaires seront trouvées dans lesquelles VB peut assigner une probabilité plus ou moins élevée à un regoupement donné. Dans les cas extrêmes comme les très grands modèles calculés sur très peu de données, l'ensemble des rapports "α" peut amener le système à élaguer les degrés de liberté qui ne sont pas utilisés.

6. Expérimentations

Nous décrivons à présent le cadre expérimental pour l'évaluation de la méthode Variationnelle Bayesienne.

Il est important de signaler que nous expérimentons la tâche de regroupement en locuteurs mais que nous ne sommes pas intéressés à comparer différentes topologies mais des critères différents pour une même topologie. Dans ce sens, le modèle que nous utilisons ici peut être entrainé en utilisant de pures méthodes ML/MAP/VB et convient pour comparer différents critères.

La base de données utilisée pour les tests sur données réelles est NIST Broadcast News (BN) 1996 HUB-4 qui consiste en 4 fichiers de près d'une demi-heure. Le premier fichier consiste en 7 locuteurs, le deuxième 13, le troisième 15 et le quatrième de 21. Un problème important de la base BN est que les fichiers contiennent une grande part de non-parole telle que de la musique ou du bruit. En plus, la parole n'est pas toujours propre mais est plus généralement bruitée (musique, parole de fond, etc). La parole peut être à bande étroite ou large. Ces problèmes rendent généralement la tâche de regroupement très difficile.

6.1. *Critère d'évaluation*

Afin d'évaluer la qualité du regoupement, nous utilisons les concepts de pureté du regroupement et pureté de locuteur introduits respectivement en [SOLO 98] et [LAP 03]. Nous considérons dans tous nos tests un regroupement additionnel pour les évènements non-parole. En fait, parole et non-parole ne sont pas explicitement séparés dans le prétraitement mais la séparation est obtenue comme conséquence du regroupement lui-même. Utilisant la notation de [LAP 03], définissons

– R : nombre de locuteurs

– S : nombre de regroupements

– n_{ij} : nombre de trames dans le regoupement i prononcée par le locuteur j

– $n_{.j}$: nombre de trames prononcées par le locuteur j, $j = 0$ désigne les trames de non-parole.

– $n_{i.}$: nombre de trames dans le regroupement i

– N : nombre de trames dans le fichier

– N_s : nombre de trames parole dans le fichier

On définit alors la pureté de regroupement p_i et la pureté de locuteur q_j :

$$p_i = \sum_{j=0}^{R} \frac{n_{ij}^2}{n_{i.}^2} \quad q_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_{.j}^2} \tag{33}$$

Les définitions de pureté moyenne de regroupement (*acp*)et de pureté moyenne du locuteur (*asp*) s'ensuivent :

$$acp = \frac{1}{N} \sum_{i=0}^{S} p_i \, n_{i.} \quad asp = \frac{1}{N_s} \sum_{j=1}^{R} q_j \, n_{.j} \tag{34}$$

La moyenne géométrique :

$$K = \sqrt{asp \cdot acp} \tag{35}$$

est utilisée comme mesure globale prenant en compte à la fois *acp* et *asp*. La valeur de K et utilisée afin d'éviter des propriétés indésirables introduites par le seul usage de *acp* ou de *asp*. En fait, pour obtenir une grande valeur pour *acp*, il suffit de considérer un grand nombre de regroupements ; d'autre part, pour obtenir un *asp* élevé, il suffit de considérer un petit nombre de regroupements. Evidemment dans le cas idéal nous devrions avoir $acp = asp = K = 1$ mais généralement un score de $0 \leq K \leq 1$ est obtenu.

6.2. *Cadre expérimental*

Le but de notre tâche est d'estimer le meilleur nombre de regroupements et la meilleure segmentation en même temps. En section 4.2 nous avons défini la topologie des modèles que nous utiliserons. Considérons à présent l'algorithme pour la recherche du meilleur nombre de regroupements. Nous faisons l'hypothèse que le fichier peut être "sur-regroupé" en utilisant un nombre de regroupements $S_{initial}$ supérieur au nombre réel. Les paramètres sont appris en utilisant un des critères considérés (ML, MAP ou VB). Ensuite, le nombre de regroupements est progressivement réduit de $S_{initial}$ à 1 et les modèles intermédiaires sont entrainés. Une fois que tous les modèles sont obtenus, les scores pour la sélection des modèles sont calculés par BIC ou l'énergie libre VB. Evidemment le modèle choisi aura le score le plus élevé. Même si ce choix est la décision finale sur le modèle, il est également intéressant de considérer certains autres scores qui éclairent la différence entre les algorithmes. Dans nos expériences nous considérerons aussi le score de la segmentation optimale pour tous les critères d'apprentissage (le choix est fait sur base de l'étiquette) et le score de segmentation pour le système initialisé avec le nombre réel de regroupements (qui est généralement inconnu). De façon surprenante, le meilleur score de segmentation n'est pas obtenu pour un système initialisé avec le nombre réel de regroupements.

Dans [VAL 04a] et [VAL 04b], nous présentions une comparaison préliminaire entre les systèmes ML/BIC et VB. Dans cet article, nous étudions l'impact de la distribution a priori sur le système VB mettant en évidence les avantages et désavantages de différentes distributions a priori.

6.3. *Résultats sur le fichier Broadcast News*

Dans cette section, nous présentons les résultats obtenus sur les quatre fichiers Broadcast News décrits précédemment. Les caractéristiques utilisées sont 12 LPCC calculés sur des trames de 30 ms décalées de 10ms. Le système est initialisé avec $S_{initial} = 30$ et 15 composantes pour chaque GMM utilisé. Contrairement à nos travaux précédents [VAL 04a] et [VAL 04b] où nous utilisions une contrainte de durée courte, ici la contrainte est de 200 trames soit 2 secondes. A première vue, cette contrainte peut sembler très sévère mais dans les fichiers BN l'alternance des locuteurs est assez uniforme et c'est pourquoi 200 trames donnent des résultats très intéressants. Augmenter la durée au delà de 2 secondes n'apporte aucun avantage au regroupement. Les formules utilisées pour l'estimation ML, MAP et VB sont détaillées en annexe I tandis que les détails pour le calcul de l'énergie libre se trouvent en annexe II.

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
ML (connu)	8	0.61	0.89	0.74	14	0.76	0.66	0.71
ML (meilleur)	10	0.80	0.88	0.84	15	0.83	0.69	0.76
ML/BIC $\lambda = 1$	28	0.91	0.50	0.67	30	0.89	0.45	0.64
ML/BIC $\lambda = 2$	19	0.90	0.67	0.77	20	0.86	0.57	0.70
ML/BIC $\lambda = 3$	10	0.80	0.88	0.84	15	0.83	0.69	0.76
File	File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K
ML (connu)	16	0.76	0.77	0.77	21	0.72	0.66	0.69
ML (meilleur)	15	0.79	0.84	0.82	12	0.64	0.82	0.72
ML/BIC $\lambda = 1$	30	0.87	0.55	0.69	30	0.81	0.57	0.68
ML/BIC $\lambda = 2$	15	0.79	0.84	0.82	21	0.71	0.65	0.68
ML/BIC $\lambda = 3$	15	0.79	0.84	0.82	14	0.66	0.74	0.70

Tableau 1. *Résultats des tests de regroupement de locuteurs sur des données NIST 1996 HUB-4 pour ML/BIC*

La table 1 montre les résultats pour le système ML/BIC à la fois pour le cas théorique de $\lambda = 1$ mais aussi pour la méthode ajustée pour $\lambda = 2, \lambda = 3$.

La ligne ML(connu) fournit les résultats du regoupement lorsque le système est initialisé avec le nombre réel de groupes (qui est connu par l'étiquetage). La ligne ML(meilleur) montre les meilleurs résultats obtenus c'est à dire les solutions à K élevé. Tout d'abord il faut remarquer que le meilleur résultat est rarement obtenu avec le nombre de groupes connu. Le système ($\lambda = 1$) fournit de très mauvais résultats avec un *acp* très élevé, un *asp* très faible et une estimation élevée du nombre de groupes. Ceci montre de façon évidente les limitations de l'approximation BIC. Lorsque le système est ajusté manuellement en modifiant la valeur de λ, on peut obtenir un score meilleur. Cependant la valeur optimale de λ diffère pour chaque fichier par exemple le fichier 3 présente son meilleur score pour $\lambda = 2$ tandis que pour les fichiers 1 et

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
VB (connu)	8	0.72	0.92	0.81	14	0.69	0.64	0.67
VB (meilleur)	10	0.84	0.90	0.87	9	0.70	0.78	0.74
VB (selectionne)	10	0.84	0.90	0.87	9	0.70	0.78	0.74
File	File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K
VB (connu)	16	0.76	0.84	0.80	21	0.71	0.65	0.68
VB (meilleur)	14	0.76	0.84	0.80	12	0.69	0.79	0.74
VB (selectionne)	14	0.76	0.84	0.80	12	0.69	0.79	0.74

Tableau 2. *Résultats des tests de regroupement de locuteurs sur des données NIST 1996 HUB-4 pour la méthode VB avec distributions a priori plates*

File	File 1				File 2			
	N_c	acp	asp	K	N_c	acp	asp	K
VB (connu)	8	0.68	0.92	0.79	14	0.69	0.84	0.76
VB (meilleur)	19	0.90	0.93	0.92	17	0.87	0.91	0.89
VB (selectionne)	16	0.83	0.92	0.88	14	0.71	0.91	0.81
File	File 3				File 4			
	N_c	acp	asp	K	N_c	acp	asp	K
VB (connu)	16	0.80	0.88	0.84	21	0.74	0.78	0.76
VB (meilleur)	20	0.81	0.90	0.85	22	0.74	0.84	0.79
VB (selectionne)	24	0.86	0.80	0.83	21	0.71	0.79	0.75

Tableau 3. *Résultats des tests de regroupement de locuteurs sur des données NIST 1996 HUB-4 par la méthode VB avec distributions a priori non-plates*

2, un choix de $\lambda = 3$ est optimal. C'est un problème typique du BIC que de devoir réestimer le seuil pour chaque cas différent.

La figure 2 montre sur le même graphe les résultats des regoupements (en vert) et les scores BIC (en rouge et en bleu) pour le sysème ML/BIC pour des valeurs de $\lambda = 1$ et $\lambda = 3$. La courbe $\lambda = 1$ est très différente de la courbe des scores de regroupement tandis que pour $\lambda = 3$ ces courbes sont assez proches. La dépendance du résultat BIC vis à vis de λ est évidente : une valeur erronée de λ produit une courbe complètement différente.

Considérons à présent la sélection de modèle par apprentissage VB. En table 2, on trouve les résultats pour l'apprentissage VB. Les distributions VB a priori ont été initialisées comme distributions suffisamment plates pour ne pas apporter d'information et donc sans influence sur le résultat final. Elles peuvent être simulées en utilisant de très petits hyperparamètres dans les distributions. Dans l'expérience, nous avons utilisé $\lambda_{\beta 0} = \rho_0 = \xi_0 = 10^{-3}$, $\nu_0 = d - 1 + 10^{-3}$ où d est la dimension des vec-

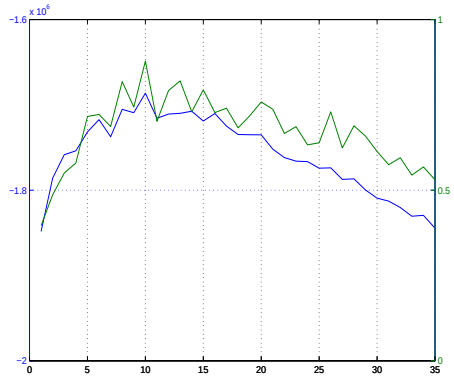

Figure 1. Score variationelle (ligne blue - raxe Y a droite) versus score du regroupment (ligne verte - axe Y a gauche) avec prior plat pour le fi chier 1

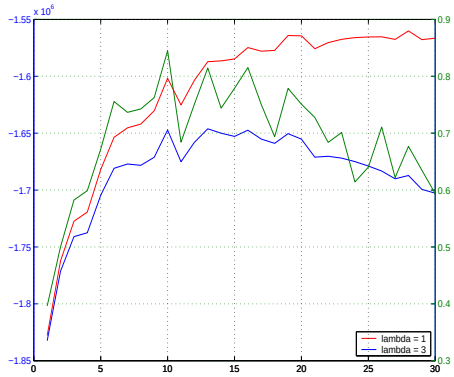

Figure 2. score BIC avec $\lambda = 1,3$(lignes rouge et blue - axe Y a droite) versus score du regroupment (ligne verte - axe Y a gauche)

teurs acoustiques et $\Phi_0 = \phi I$ où I est la matrice unité et $\phi = 10^{-3}$. Comparons à présent le résultat du meilleur système pour ML/BIC avec le VB avec distribution a priori plate. Sur les fichiers 1 et 4, le meilleur résultat VB est supérieur au meilleur résultat ML/BIC tandis que pour les fichiers 2 et 3, ML/BIC fournit des résultats supérieurs. Cependant, ML/BIC et VB avec distribution a priori plate fournissent des résultats fort semblables (la différence ne dépasse pas 2 à 3% en valeur absolue). Un point intéressant est que VB choisit *toujours* le système présentant le meilleur score. La figure 1 montre le score K et l'énergie libre variationnelle pour le fichier 1 : il est facile de remarquer que l'énergie libre suit de près le score de regroupement et que le maximum des courbes du score de regroupement coincide avec le maximum de la courbe de l'énergie libre variationnelle. Un comportement semblable est observé pour

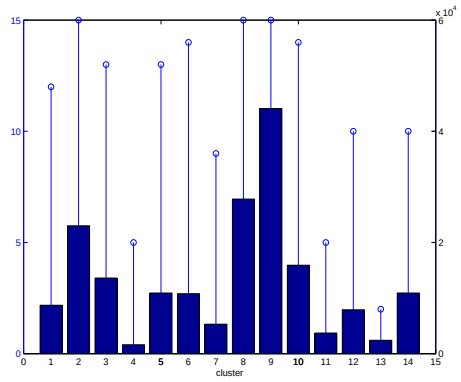

Figure 3. ligne fi ne (axe Y a gauche) : nombre fi nale de composantes gaussiennes vs. nombre de regroupments ; ligne grosse (Y axe a droite) : nombre de trames assigneé a un regropument vs. nombre de regroupments.

les quatre fichiers. Tant que la distribution a priori n'apporte aucune information, un comportement semblable est observé.

D'autre part, pour des distributions a priori non-plates, les résultats du regroupement sont affectés par les distributions. Evidemment des distributions plates ne sont pas optimales du point de vue du score de regroupement. Différentes approches peuvent être utilisées pour contourner le problème du choix initial. Dans la modélisation hiérarchique, les hyperparamètres a priori peuvent être considérés comme paramètres eux-mêmes avec leurs distributions propres régulées par d'autres hyperparamètres qui à leur tour possèdent leur distibution propre et ainsi de suite. Le but du regroupement hiérarchique est de rendre l'algorithme moins sensible au choix arbitraire de certains hyperparamètres parce que trop éloignés des données dans le modèle hiérarchique. D'autre part, les distributions a priori peuvent être simplement optimisées en trouvant les hyperparamètres optimaux compte tenu des données. Dans ce cas l'algorithme démarrerait avec une estimation itérative des paramètres et hyperparamètres. En général, l'optimisation des hyperparamètres conduit à un système d'équations non-linéaires.

En regroupement de locuteurs VB, les performances peuvent être significativement améliorées en utilisant une distribution a priori appropriée. En table 3, on présente les résultats pour des distributions a priori déterminées de façon heuristique en adoptant simplement des grandes valeurs pour les hyperparamètres c'est à dire $\lambda_{\beta 0} = \rho_0 = \xi_0 = 300$,$\nu_0 = d-1+300$ où d est la dimension des vecteurs acoustiques et $\Phi_0 = \phi I$ où I est la matrice unité et $\phi = 300$.

Les scores pour le meilleur système en table 3 sont régulièment supérieurs à ceux des tables 1 et 2. Et pourtant le système sélectionné n'est pas le meilleur ce qui signifie que lorsque les distributions a priori ne sont pas plates, la correspondance entre le score de regroupement et l'énergie libre variationnelle n'est plus aussi étroite que

précédemment. Même si le système sélectionné n'est pas le meilleur, les scores des systèmes sélectionnés sont à nouveau meilleurs que ML/BIC et VB avec distribution a priori plate. Ce résultat n'est pas du tout surprenant car utiliser des distributions étroites signifie l'ajout d'information dans le processus. Il faut remarquer que celles-ci ne sont pas des distributions a priori à partir des données (comme en [GAU 94]) mais simplement des distributions "fortes" qui décident de la force de la distribution. D'une part, si les distributions a priori non plates doivent être déterminées de façon heuristique, l'apprentissage VB perd la propriété d'être indépendante du réglage.

En outre, les distributions a priori sont à l'origine d'une autre propriété de l'apprentissage VB : les degrés de liberté excédentaires sont élagués. En fait le modèle initial avec S groupes et M composantes Gaussiennes par mélange est rarement conservé à l'issue de l'entrainement. Le nombre de groupes et de composantes Gaussiennes est réduit à un nombre plus petit dépendant généralement de la quantité de données. Cette propriété est très utile parce que dans certains cas le modèle original n'est pas adapté à la quantité de données dans le groupe : au lieu de considérer toutes les composantes Gaussiennes initiales, l'apprentissage VB simplifie lui-même le modèle en nombre réduit de composantes Gaussiennes. Notre système est initialisé avec 15 Gaussiennes par état (locuteur) mais en fin d'apprentissage moins de 15 composantes restent en piste. Le nombre final de composantes Gaussiennes et les données assignées à chaque groupe sont représentées en figure 3 : aux groupes plus riches en données sont assignées plus de composantes Gaussiennes. Cette propriété est très intéressante pour éviter le surentrainement ; en fait MAP et ML partagent généralement les données entre toutes les composantes disponibles tandis que VB réduit ses composantes.

Pour résumer nos résultats, nous pouvons considérer les trois systèmes décrits précédemment : ML/BIC, VB avec distribution a priori plate et VB avec une distribution a priori heuristique. Lorsque VB avec distribution a priori plate est utilisée, le meilleur système est très proche du système ML mais la sélection est faite sans aucun réglage heuristique. Les résultats ML/BIC sont fortement affectés par le seuil (c'est à dire le choix de λ) tandis que VB détecte toujours le meilleur regroupement avec le terme de pénalité entrainé avec le système. D'autre part, lorsqu'on fait l'hypothèse d'une distribution a priori forte et déterminée heuristiquement, VB surclasse clairement à la fois ML/BIC et le VB avec distribution a priori plate. Cependant la détermination de la distribution a priori rend le système dépendant d'une sorte d'ajustement même si la distribution a priori peut être estimée à partir des données.

Les résultats surpassent légèrement ceux obtenus en [LAP 03] dans lequel le modèle était basé sur des mises en correspondance auto-organisées et le critère de sélection du modèle était le critère d'indexation Bayesien et ceux obtenus en [AJM 02] dans lesquels le modèle est basé sur les mélanges de Gaussiennes et un critère BIC modifié est utilisé.

7. Discussion et Conclusion

Dans ce papier, nous avons décrit un nouveau système de regroupement de locuteurs basé sur une sélection de modèles jointe à un apprentissage Variationnel Bayesien. Le système est comparé à un système classique de regroupement de locuteurs basé sur l'apprentissage ML et la sélection BIC. La fonction objective variationnelle Bayesienne aussi connue sous le nom d'énergie libre inclut un terme de pénalité comme effet direct de l'intégration Bayesienne : ce terme est entrainé avec le modèle et très utile pour la sélection de modèles. Sous l'hypothèse d'indépendance des distributions sur les variables cachées et sur les paramètres, un algorithme de type EM pour l'apprentissage variationnel (VBEM) peut être développé. L'accroissement de la charge de calcul par rapport à l'EM classique n'est pas significative. Cette méthode peut être appliquée à une grande catégorie de modèles et parmi ceux-ci les HMM et les GMM. Nous avons dérivé les formules d'estimation VB pour un regroupement en locuteurs basé sur un HMM ergodique dont les émissions sur état sont modélisées par des GMM. Les résultats expérimentaux sur les fichiers BN montrent que contrairement au BIC l'apprentissage VB avec distributions a priori plates peut sélectionner efficacement le meilleur regroupement. Les performances peuvent également être améliorées en utilisant des distributions a priori déterminées de façon heuristique.

Notre discussion laisse cependant une série de problèmes ouverts. Tout d'abord, il serait intéressant d'optimiser les distributions a priori afin d'accroitre le score de regroupement et de choisir le meilleur modèle. Cette tâche peut être réalisée en utilisant différentes techniques discutées précédemment comme les distributions a priori hiérarchiques et l'optimisation numérique des hyperparamètres. Nous n'avons pas encore expérimenté ces voies. Un autre problème ouvert est que nous avons décidé de comparer VB et ML sur un système donné alors qu én regroupement de locuteurs un très grand nombre de méthodes et de topologies ont été proposées. Une comparaison plus correcte prendrait aussi en considération différents systèmes. Parfois un choix courant pour compenser l'absence de données suffisantes pour un locuteur est de construire un modèle à partir d'un Modèle de référence universel (UBM) entrainé préalablement. Le modèle final est typiquement obtenu par adaptation MAP. Nous avons discuté largement la similitude et les différences entre VB et l'apprentissage MAP dans les sections précédentes et les problèmes de reparamétrisation avec MAP. L'adaptation VB peut être obtenue simplement en utilisant des distributions a priori Bayesiennes empiriques telles que les distributions a priori VB. En ce sens, il serait vraiment intéressant de comparer notre système VB avec le système ML/BIC. Cependant pour réaliser cela de la manière la plus correcte certains problèmes sur les dépendances de reparamétrisation MAP doivent être considérées. Ceci permettrait de saisir comment les différents apprentissages VB, ML et MAP fonctionnent sur une tâche de regroupement de locuteurs.

Pour conclure, nous attirons l'attention sur le fait que ces résultats devraient être validés sur des bases de données plus volumineuses comme celles fournies par NIST ou ESTER contenant plus de signaux hétérogènes et plus de variantes de propriétés acoustiques des locuteurs.

8. Bibliographie

[SCH 78] SCHWARTZ G., « Estimation of the dimension of a model », Annals of Statistics, 6, 1978.

[CHE 98] CHEN S., GOPALAKRISHNAN P., « Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion », Proceedings of the DARPA Workshop, 1998.

[TRI 99] TRITSCHLER A. , GOPINATH R.,« Improved speaker segmentation and segments clustering using the bayesian information criterion », In EUROSPEECH'99, 679-682.

[MKAY 95] MACKAY D.J.C. « Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural network », Network :Computation in Neural Systems,6 :469-505,1995

[ATT 00] ATTIAS, H., « A Variational Bayesian framework for graphical models », , Adv. in Neural Inf. Proc. Systems 12, MIT Press,Cambridge, 2000.

[MKAY 97] MACKAY D.J.C., « Ensemble Learning for Hidden Markov Models », , Technical Report, Cavedish Laboratory, University of Cambridge 1997

[BEAL 02] BEAL, M.J. , GHAHRAMANI, Z. « The Variational Bayesian EM Algorithm for Incomplete Data : with Application to Scoring Graphical Model Structures », , In Bayesian Statistics 7 :453-464, Oxford University Press, 2003

[BEAL 03] BEAL, M.J. « Variational algorithms for approximate bayesian inference », , PhD thesis, University College London.

[OLS95] OLSEN J. O., « Separation of speaker in audio data », , In EUROSPEECH 1995, pp. 355-358.

[AJM 02] AJMERA J. , AL « Unknown-multiple speaker clustering using HMM », , In ICSLP 2002.

[LAP 03] LAPIDOT I. « SOM as Likelihood Estimator for Speaker Clustering », , In EUROSPEECH 2003.

[DEM 77] DEMPSTER A.P. , LAIRD N.M. , , RUBIN D.B. , « Maximum Likelihood from Incomplete Data via theEM algorithm », , Journal of the Royal Statistical Society, Series B, 39(1) : 1-38, 1977

[GAU 94] 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

[MKAY 95b] MACKAY D.J.C. , PETO L.C., « A hierarchical Dirichlet language model », , Natural Language Engineering 1995

[MKAY 98] MACKAY D.J.C. « Choice of basis for Laplace approximation », , In Machine Learning 33 1998

[MKAY 01] MACKAY D.J.C. « A problem with variational free energy minimization », , Technical Report, Cavedish Laboratory, University of Cambridge 2001

[SOLO 98] SOLOMONOFF A. MIELKE A., SCHMIDT, , GISH H.,« Clustering speakers by their voices », , In ICASSP 98, pp. 557-560

[DELA 00] DELACOURT P. , WELLEKENS C. « Distbic : a speaker based segmentation for audio data indexing », , In Speech Communication Vol.32 Septembre 2000.

[BIS 03] BISHOP, C. M. , WINN J. « Structured variational distributions in VIBES. », , In C. M. Bishop and B. Frey (Eds.), Proceedings Artifi cial Intelligence and Statistics, Key West, Florida. Society for Artifi cial Intelligence and Statistics.

[ATT 99] H ATTIAS. « Inferring parameters and structure of latent variable models by variational Bayes. », , In Proceedings of the 15th Conference on Uncertainty in Artifi cial Intelligence, 21-30, 1999.

[VAL 04a] VALENTE F. , WELLEKENS C. « Variational Bayesian Speaker clustering », , In Proceedings of Odyssey 2004.

[VAL 04b] VALENTE F. , WELLEKENS C. « Regroupment variationnelle du locuteur », , In Proceedings of Jep 2004.

9. Annexe I

Dans cette section nous dérivons les formules de réestimation pour notre modèle (2) pour les critères d'apprentissage ML, MAP et VB. L'ensemble des paramètres consiste en probabilités de transition α_j d'état à état, les poids de Gaussiennes β_{ij}, les moyennes μ_{ij} et les variances Γ_{ij} où $i = 1, ..., M$ avec M le nombre de composantes Gaussiennes et $j = 1, ..., S$ avec S le nombre de groupes. Deux sortes de variables latentes x et z doivent être considérées ici : une variable x qui désigne le locuteur (ou l'état equivalent) et z (conditionné par x) qui désigne la composante Gaussienne qui a émis l'observation. Considérons donc le critère ML, l'étape E a la forme suivante :

$$\gamma_{x_t=j} = P(x_t = j|O_t) = \frac{\alpha_j P(O_t|s_j)}{\sum_j \alpha_j P(O_t|s_j)} \quad (36)$$

$$\gamma_{z_{tp}=i|x_t=j} = P(z_{tp} = i|x_t = j, O_{tp}) = \frac{\beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})}{\sum_{i=1}^{D} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})} \quad (37)$$

Pour l'étape de maximisation, les formules suivantes de réestimation sont déduites :

$$\alpha_j = \sum_{t=1}^{T} \gamma_{x_t=j}/T \quad (38)$$

$$\beta_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j} \gamma_{z_{tp}=i|x_t=j}}{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j}} \quad (39)$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j} \gamma_{z_{tp}=i|x_t=j} O_{tp}}{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j} \gamma_{z_{tp}=i|x_t=j}} \quad (40)$$

$$\Gamma_{ij} = \frac{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j} \gamma_{z_{tp}=i|x_t=j} (O_{tp} - \mu_{ij})^T (O_{tp} - \mu_{ij})}{\sum_{t=1}^{T} \sum_{p=1}^{D} \gamma_{x_t=j} \gamma_{z_{tp}=i|x_t=j}} \quad (41)$$

Dans l'apprentissage MAP, la distribution a priori sur les paramètres doit être considérée. Ainsi définissons les distributions a priori suivantes et les distributions a posteriori correspondantes :

$$P(\alpha_j) = Dir(\lambda_{\alpha 0}) \quad P(\beta_{ij}) = Dir(\lambda_{\beta 0})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho_0, \xi_0 \Gamma_{ij}) \quad P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \quad (42)$$

$$P(\alpha_j) = Dir(\lambda_{\alpha j}) \quad P(\beta_{ij}) = Dir(\lambda_{\beta ij})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho_{ij}, \xi_{ij} \Gamma_{ij}) \quad P(\Gamma_{ij}) = W(\nu_{ij}, \Phi_{ij}) \quad (43)$$

où Dir désigne une distribution de Dirichlet, N est une distribution normale et W une distribution de Wishart. L'étape E pour l'apprentissage MAP a la même forme que l'étape E pour l'apprentissage ML soit (37). Dans l'étape M, on commence par mettre à jour les distributions a posteriori comme suit :

$$\lambda_{\alpha_j} = \sum_{t=1}^{T} N_j + \lambda_{\alpha 0} \quad \lambda_{\beta_{ij}} = N_{ij} + \lambda_{\beta 0} \tag{44}$$

$$\rho_{ij} = \frac{N_{ij}\,\mu_{ij} + \xi_0\,\rho_0}{N_{ij} + \rho_0} \tag{45}$$

$$\xi_{ij} = N_{ij} + \xi_0 \quad \nu_{ij} = N_{ij} + \nu_0 \tag{46}$$

$$\Phi_{ij} = N_{ij}\,\Gamma_{ij} + \frac{N_{ij}\,\xi_0\,(\mu_{ij} - \rho_0)(\mu_{ij} - \rho_0)^T}{N_{ij} + \rho_0} + \Phi_0 \tag{47}$$

où $N_{ij} = \sum_{t=1}^{T} \sum_{p=1}^{D} \tilde{\gamma}_{x_t=j} \tilde{\gamma}_{z_{tp}=i|x_t=j}$ and $N_j = \sum_{t=1}^{T} \tilde{\gamma}_{x_t=j}$. Finalement les paramètres optimaux MAP sont estimés :

$$\alpha_i = \frac{\lambda_{\alpha_i} - 1}{\sum_i (\lambda_{\alpha_i} - 1)} \quad \beta_{ij} = \frac{\lambda_{\beta_{ij}} - 1}{\sum_j (\lambda_{\beta_{ij}} - 1)} \quad \mu_{ij} = \rho_{ij} \quad \Gamma_{ij} = (\nu_{ij} - d)\Phi_{ij}^{-1} \tag{48}$$

Pour l'apprentissage VB, il n'y a pas d'estimation de paramètres mais bien l'estimation de leurs distributions. Tant que nous faisons l'hypothèse de distributions a priori semblables (21), l'étape M sera identique à celui de l'apprentissage MAP comme discuté précédemment. D'autre part, l'étape E considèrera l'intégration sur les distributions des paramètres :

$$\tilde{\gamma}^*_{z_{tp}=i|x_t=j} = \tilde{\beta}_{ij}\,\tilde{\Gamma}_{ij}^{1/2}\,exp\{-E\}\,exp\{\tfrac{-g}{2\nu_{ij}}\}$$

$$avec\ E = \tfrac{1}{2}(O_{tp} - \rho_{tp})^T \bar{\Gamma}_{ij}(O_{tp} - \rho_{tp}) \tag{49}$$

$$\tilde{\gamma}_{z_{tp}=i|x_t=j} = q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j) = \frac{\bar{\gamma}^*_{z_{tp}=i|x_t=j}}{\sum_i \bar{\gamma}^*_{z_{tp}=i|x_t=j}} \tag{50}$$

$$\tilde{\gamma}^*_{x_t=j} = \tilde{\alpha}_j \prod_{p=1}^{D} \sum_{i=1}^{M} \tilde{\gamma}^*_{z_{tp}=i|x_t=j} \tag{51}$$

$$\tilde{\gamma}_{x_t=j} = q(\gamma_{x_t} = j) = \frac{\tilde{\gamma}^*_{x_t=j}}{\sum_j \tilde{\gamma}^*_{x_t=j}} \tag{52}$$

10. Annexe II

Dans cette section nous dérivons la forme explicite de l'énergie libre variationnelle (14) lorsque nous considérons un modèle comme (2). L'importance de disposer d'une forme explicite pour l'énergie libre consiste comme exposé plus haut, dans le fait qu'elle est très efficace comme critère de sélection de modèle qui peut être substituée à d'autres critères (par exemple BIC,MML,...). Réécrivons l'expression (14) pour le modèle considéré

$$\begin{aligned} F(\theta,\gamma) &= \int d\theta d\gamma q(\gamma) q(\theta) log[p(O,\gamma,\theta)/q(\gamma)q(\theta)] \\ &= < log\,\frac{p(O,\gamma|\theta)}{q(\gamma)} >_{\gamma,\theta} - D[q(\theta)||p(\theta)] \end{aligned} \tag{53}$$

où conformément à nos discussions précédentes, l'ensemble des variables cachées $\gamma = \{\gamma_{z_{tp}|x_t}\ ,\ \gamma_{x_t}\}$ consiste en deux variables : une désigne l'état (le locuteur) soit x_t et l'autre désigne la composante z_{tp} tandis que $q(\gamma_{z_{tp}} = i,\ \gamma_{x_t} = j)$ est la probabilité que le locuteur j parle et que la composante i de l'état j produise l'observation.

Considérons la factorisation $p(O, \gamma|\theta) = p(O|\gamma, \theta)p(\gamma|\theta)$, nous pouvons écrire (53) comme somme de trois termes différents :

$$\begin{aligned} F(\theta, \gamma) &= \int d\theta d\gamma q(\gamma) q(\theta) [log(p(O|\gamma, \theta)) + log(p(\gamma|\theta))] + \\ &- \int d\theta d\gamma q(\gamma) q(\theta) log q(\gamma) - D[q(\theta)||p(\theta)] \end{aligned} \tag{54}$$

Considérons le fait que $q(\gamma_{z_{tp}} = i, \gamma_{x_t} = j) = q(\gamma_{x_t} = j)\, q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j)$ et utilisant la même notation que précédemment nous définissons $\gamma_{z_{tp}=i|x_t=j} = q(\gamma_{z_{tp}} = i|\gamma_{x_t} = j)$ et $\gamma_{x_t=j} = q(\gamma_{x_t} = j)$. Revenant à (54) nous examinons chacun des trois termes.

– le premier terme est :

$$\int d\theta d\gamma q(\gamma) q(\theta) [log(p(O|\gamma, \theta)) + log(p(\gamma|\theta))] \tag{55}$$

Parce que les variables cachées sont discrètes, l'intégrale vis à vis de γ se transforme en une somme sur les états et les mélanges de Gaussiennes. Explicitons (55) par rapport à T,D et les variables cachées :

$$\sum_{t=1}^{T} \sum_{p=1}^{D} \sum_{j=1}^{S} \sum_{i=1}^{M} \gamma_{z_{tp}=i|x_t=j} \gamma_{x_t=j} \int d\theta q(\theta) [log(p(O_{tp}|, \theta \gamma_{z_{tp}=i, x_t=j})) + log(p(\gamma_{z_{tp}=i, x_t=j})|\theta))] \tag{56}$$

Considérons à présent la factorisation $p(\gamma_{z_{tp}=i, x_t=j}|\theta) = p(\gamma_{x_t=j}|\theta) p(z_{tp} = i|x_t = j, \theta) = \alpha_j \beta_{ij}$ et utilisant l'expression définie plus haut $\tilde{\gamma}^*_{z_{tp}=i|x_t=j}$, il est possible de réécrire (56) :

$$\sum_{t=1}^{T} \sum_{j=1}^{S} \gamma_{x_t=j} [log\, \tilde{\alpha}_j + \sum_{p=1}^{D} \sum_{i=1}^{M} \gamma_{z_{tp}=i|x_t=j}\, log\, \tilde{\gamma}^*_{z_{tp}=i|x_t=j}] \tag{57}$$

Tous les éléments de (57) sont explicites et connus.

– Considérons à présent le deuxième terme de (54) :

$$\int d\theta d\gamma q(\gamma) q(\theta) log q(\gamma) = \int q(\gamma) log q(\gamma) d\gamma \tag{58}$$

Explicitons cette expression vis à vis du temps, de la durée et des variables cachées. Le résultat est :

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

$$= \sum_{t=1}^{T}\sum_{j=1}^{S}\{\gamma_{x_t=j}[log\,\gamma_{x_t=j} + \sum_{p=1}^{D}\sum_{i=1}^{M}\gamma_{z_{tp}=i|x_t=j} log\,\gamma_{z_{tp}=i|x_t=j}]\} \qquad (59)$$

A nouveau dans (59) tous les termes sont explicites et connus.

– Le dernier terme à considérer est la divergence KL entre les distributions a posteriori et a priori. Les distributions de paramètres sont de Dirichlet, Normale et de Wishart définies en (21). Grâce à l'indépendance entre les distributions de paramètres, on peut écrire :

$$\begin{aligned} D[q(\theta)||p(\theta)] &= D(Dir(\lambda_{\alpha j})||Dir(\lambda_{\alpha 0})) + \sum_j D(Dir(\lambda_{\beta ij})||Dir(\lambda_{\beta 0})) \\ &+ \sum_j\sum_i D(N(\rho_{ij}, \xi_{ij}\nu_{ij}\Phi_{ij}^{-1})||N(\rho_0, \xi_0\nu_{ij}\Phi_{ij}^{-1})) \\ &+ \sum_i\sum_j D(W(\nu_{ij}, \Phi_{ij})||W(\nu_0, \Phi_0)) \qquad (60) \end{aligned}$$

Une forme explicite pour toutes les divergences KL en (60) peut être trouvée.

Nous avons donc montré qu'il est possible d'expliciter l'énergie libre variationnelle.