Regroupement bayesien variationnel des locuteurs

Fabio Valente, Christian Wellekens

Institut Eurécom, BP 193 - F 06904 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 run experiments on the NIST 1996 HUB-4 evaluation test for speaker clustering. Two cases are considered : the speaker number is a priori known and it has to be estimated. We evaluate results in terms of average cluster purity and average speaker purity. VB shows a higher accuracy compared to the Maximum Likelihood solution.

1. INTRODUCTION

Une tache importante dans les applications de reconnaissance de la parole est le regroupement des locuteurs. Un nombre énorme de techniques de regroupement robuste des locuteurs a été proposé : elles consistent généralement en quantificateurs vectoriels [11] , modèles de Markov cachés (HMM)[2] et transformations auto-organisées (SOM) [3] . Un problème crucial en apprentissage non-supervisé est que le nombre exact de locuteurs est inconnu. Pour déterminer un nombre raisonnable de groupes, une méthode de sélection des modèles est indispensable ; généralement le critère BIC ou une forme " évoluée " du BIC (critère d'information bayesien) [2] sont utilisés. Dans cette communication, nous proposons l'usage d'une technique d'apprentissage relativement nouvelle qui permet d'estimer les paramètres et de sélectionner les modèles simultanément. Elle est généralement connue sous le nom d' *Apprentissage Bayesien Variationnel* (*VB*) (ou apprentissage d'ensemble). Des modèles comme les GMM (modèle à mélange de gaussiennes) ou les HMM peuvent être entrainés en utilisant le cadre bayesien variationnel (voir [4],[10])). L'entrainement VB a l'avantage d'utiliser comme critère d'optimisation une expression qui convient également comme critère de sélection des modèles. En dépit du fait que l'apprentissage VB est une méthode approximée, elle a déjà fait l'objet d 'applications fructueuses dans des problèmes de reconnaissance de la parole pour le regroupement d'états ([5]), la réduction de dimension ([6]) et l'estimation de GMM ([7]). Ici c'est au regroupement de locuteurs que nous appliquons la technique VB.

L'article est organisé comme suit : en section 2, nous décrivons le système de référence utilisant des HMM entrainés selon l'algorithme EM ; en section 3 nous décrivons le cadre général de l'apprentissage VB ; en section 4 nous décrivons le système de regroupement en locuteurs qui utilise l'apprentissage VB et finalement en section 5, nous décrivons nos expériences et en présentons les résultats.

2. INDEXATION HMM EN LOCUTEURS

Une méthode répandue de regroupement automatique en locuteurs utilise le HMM. Cette méthode introduite en [1] considère un HMM totalement connecté dans lequel chaque état représente un locuteur et la probabilité d'émission sur état est la probabilité d'émission pour chaque locuteur. En [2], un HMM ergodique muni de contraintes de durée est proposé. Les contraintes de durée ont l'avantage de fournir une solution non dispersée. L'utilisation d'un nombre donné de trames successives donne une statistique suffisante pour modéliser un locuteur particulier. En [3], il est montré que 100 trames successives sont suffisantes pour construire un modèle de locuteur. Examinons à présent les détails du modèle. Un HMM ergodique est un HMM totalement connecté dans lequel toutes les transitions même bidirectionnelles sont autorisées. Soit $P(O_t|s_j)$ la probabilité d'observation O_t à l'état s_j et l'instant t. Puisque toutes les transitions sont autorisées, on peut désigner par α_j la probabilité de transition vers l'état j quelque soit l'état initial où $j = 1, ..., S$ avec S désignant le nombre d'états. En d'autres mots, on peut modéliser un HMM ergodique par un simple mélange de densités. La probabilité d'observer O_t peut s'écrire :

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

Dans un modèle à contrainte de durée, l'observation O_t est un groupe de D trames consécutives où D désigne la durée minimale de présence d'un locuteur. Une façon de modéliser $P(O_t|s_j)$ est d'utiliser des GMM avec des pondérations β_{ij},des moyennes μ_{ij}et des variances Γ_{ij} où $i = 1, ..., M$, M est le nombre de gaussiennes par mélange. On peut alors écrire

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

et en conséquence l'expression de la log-vraisemblance d'une séquence complète O de T trames O_t est

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

En d'autres mots, ce modèle est un modèle de mélanges hiérarchiques dans lequel la première couche représente les S locuteurs et la seconde couche représente les modèles de locuteurs et qui forme en fait un GMM. La pondération α_j peut être interprétée comme la probabilité a priori du j-ème locuteur.

Cette sorte de modèles peut être entièrement entrainée en utilisant l'algorithme classique EM (estimer-maximiser). Nous avons fait précédemment l'hypothèse que le nombre

de locuteurs est connu a priori mais cela n'est pas toujours le cas. C'est pour cette raison qu'un critère de sélection des modèles doit être utilisé lorsque le nombre de locuteurs n'est pas connu. Dans la section suivante, nous considérons le cas de l'entrainement EM lorsque le nombre de locuteurs est connu.

2.1. Entrainement EM

Le modèle (3) est un modèle à variables cachées latentes qui peut être entrainé en utilisant le célèbre algorithme EM [12]. Deux types de variables latentes x et z sont considérés ici : une variable x qui désigne le locuteur (ou l'état associé) et z (conditionnée par x) qui désigne la composante gaussienne qui a émis l'observation. Pour le pas "Estimer" de l'algorithme, on démontre facilement que

$$\gamma_{x_t=j} = P(x_t = j|O_t) = \frac{\alpha_j P(O_t|s_j)}{\sum_j \alpha_j P(O_t|s_j)} \quad (4)$$

$$\gamma_{z_{tp}=i|x_t=j} = P(z_{tp} = i|x_t = j, O_{tp}) = \frac{\beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})}{\sum_{i=1}^{D} \beta_{ij} N(O_{tp}, \mu_{ij}, \Gamma_{ij})} \quad (5)$$

Pour le pas Maximiser, les formules de réestimation suivantes sont utilisées :

$$\alpha_j = \sum_{t=1}^{T} \gamma_{x_t=j}/T, \quad \beta_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}} \quad (6)$$

$$\mu_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j} O_{tp}}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}} \quad (7)$$

$$\Gamma_{ij} = \frac{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}(O_{tp} - \mu_{ij})^T(O_{tp} - \mu_{ij})}{\sum_{t=1}^{T}\sum_{p=1}^{D} \gamma_{x_t=j}\gamma_{z_{tp}=i|x_t=j}} \quad (8)$$

3. APPRENTISSAGE VB

Dans un premier temps, nous considérons le cadre général de l'apprentissage VB et ensuite nous montrons comment l'utiliser pour l'entrainement du modèle (3).

3.1. Cadre bayesien variationnel

Un ensemble de variables observées Y et un jeu de paramètres θ étant donnés, l'apprentissage bayesien s'attache à l'optimisation de la *vraisemblance marginale* $p(Y)$, où les paramètres θ ont été éliminés par intégration. Par la règle de Bayes, nous avons : $p(Y) = p(Y, \theta)/p(\theta|Y)$ et considérant le logarithme des deux membres, on peut écrire : $log\, p(Y) = log\, p(Y, \theta) - log\, p(\theta|Y)$. Au lieu d 'intégrer les paramètres θ vis à vis de leur densité de probabilité véritable mais inconnue, on utilise une approximation connue sous le nom de probabilité a posteriori variationnelle et notéee $q(\theta|Y)$. En calculant l'espérance vis à vis de cette densité approximée $q(\theta|Y)$, on obtient :

$$log\, p(Y) = \int q(\theta|Y) log\, p(Y, \theta) d\theta - \int q(\theta|Y) log\, p(\theta|Y) d\theta$$
$$= \int q(\theta|Y) log\, [p(Y, \theta)/q(\theta|Y)] d\theta + D(q(\theta|Y))||p(\theta|Y)) \quad (9)$$

où $D(q(\theta|Y))||p(\theta|V)$ représente la divergence Kullback-Leibler (KL) entre les densités a posteriori approximées et les véritables densités a posteriori . Le terme $\int q(\theta|Y) log\, [p(Y, \theta)/q(\theta|Y)] d\theta$ est souvent désigné sous le vocable d'énergie libre négative $F(\theta)$. Puisque la distance KL est positive par construction et nulle ssi les densités comparées sont identiques ($D(a||b) \geq 0$), $F(\theta)$ représente une borne inférieure de la log-vraisemblance $log\, p(Y)$ c'est à dire que $log\, p(Y) \geq F(\theta)$. L'apprentissage VB revient à maximiser cette borne inférieure $F(\theta)$ qui peut être réécrite

$$F(\theta) = \int q(\theta|Y) log\, p(Y|\theta) d\theta - D(q(\theta|Y)||p(\theta)) \quad (10)$$

Le second terme de (10) représente la distance entre la densité a posteriori approximée et la densité a priori des paramètres et peut être interprétée comme un terme qui pénalise les modèles plus complexes. C'est pour cette raison que $F(\theta)$ peut être utilisé pour déterminer le modèle qui s'ajuste le mieux aux données comme le fait le critère BIC.
Le Maximum a posteriori peut être vu comme cas particulier de l'apprentissage VB. En fait si $q(\theta|Y) = \delta(\theta - \theta')$, trouver le maximum de (10) revient à

$$max_{Q(\theta)} F(\theta) = max_{\theta'} \int \delta(\theta - \theta') log[p(Y|\theta)p(\theta)] d\theta$$
$$= max_{\theta'} log[p(Y|\theta')p(\theta'))] \quad (11)$$

où le terme $\int q(\theta) log\, q(\theta) d\theta$ a été écarté car constante. L'expression (11) correspond au critère MAP (Maximum a posteriori) classique. Il est important de remarquer que l'approche VB apporte de l'information sur l'incertitude sur les paramètres contrairement à MAP. En fait l'apprentissage MAP des paramètres est réalisé de façon locale ($max\, log\, [p(Y|\theta')p(\theta'))]$) tandis qu'en utilisant VB, les paramètres sont marginalisés même si l'intégration se fait par rapport à la densité a posteriori variationnelle ($max \int q(\theta|Y)\, log[p(Y|\theta)p(\theta)] d\theta$). En outre, VB permet la comparaison des modèles : la valeur de l'énergie libre donne de l'information sur la qualité du modèle alors que MAP fournit simplement les meilleurs paramètres pour un modèle imposé. Le prix à payer est que l 'énergie libre n'est qu'une borne inférieure et non une valeur exacte.

3.2. Apprentissage VB avec variables cachées

L' apprentissage VB peut être étendu au cas de données incomplètes. Dans beaucoup de problèmes d'apprentissage de machines, les algorithmes doivent traiter les variables cachées X comme les paramètres θ (voir [4]). Dans le cas des variables cachées, la densité a posteriori variationnelle devient $q(X, \theta|Y)$ et une simplification supplémentaire consiste en l'hypothèse d' índépendance conditionelle $q(X, \theta|Y) = q(X|Y)q(\theta|Y)$. Dans ce cas, l'énergie libre à maximiser devient :

$$F(\theta, X) = \int d\theta q(X) q(\theta) log[p(Y, X, \theta)/q(X)q(\theta)]$$
$$= < log\, \frac{p(Y, X|\theta)}{q(X)} >_{X,\theta} - D[q(\theta)||p(\theta)] \quad (12)$$

où $< \, . \, >_z$ signifie la moyenne par rapport à z. Il est important de noter que q est toujours interprété comme conditionné par Y et où le conditionement de q par rapport a Y n'est pas explicité pour alléger la notation. On peut montrer que lorsque $N \rightarrow \infty$ le terme de pénalité se réduit à $(|\theta_0|/2) log\, N$ où θ_0 est le nombre de paramètres c'est-à-dire que l'énergie libre devient le critère BIC. Pour trouver les densités $q(\theta)$ et $q(X)$, un algorithme EM est proposé en [4]. Il est basé sur les étapes suivantes.

$$q(X) \propto e^{<log\, p(Y,X|\Theta)>_\theta} \quad (13)$$

$$q(\theta) \propto e^{<log\, p(Y,X|\theta)>_X} p(\theta) \quad (14)$$

En appliquant itérativement (13-14), il est possible d'estimer les densités a posteriori variationnelles pour les paramètres et les variables cachées. Si $p(\theta)$ appartient à une famille conjuguée, la distribution $q(\theta)$ aura la même forme que $p(\theta)$. Une propriété intéressante de l'apprentissage VB est qu'il ne requiert pas de degrés de liberté supplémentaires mais que le modèle se réduit automatiquement. On peut apprécier cette auto-réduction de deux façons contradictoires : d'une part, elle n'est pas satisfaisante car la prédiction ne prendra pas en compte l'incertitude que

les modèles à paramètres additionnels peut apporter (voir [8]) mais d'autre part, elle peut être utilisée pour trouver le modèle optimal en entrainant un modèle initialisé avec un grand nombre de paramètres et en le laissant se réduire par élimination des paramètres non-utilisés.

4. REGROUPEMENT DES LOCUTEURS EN UTILISANT VB

Dans cette section, nous établissons les formules nécessaires à l'estimation des paramètres du modèle (3). Avant d'appliquer l'algorithme de type EM décrit ci-dessus, nous devons définir les probabilités a priori des paramètres. Définissons les probabilités suivantes qui appartiennent à une famille conjuguée (c'est à dire que les densités a priori ont la même forme que les densités a posteriori) :

$$P(\alpha_j) = Dir(\lambda_{\alpha 0}) \quad P(\beta_{ij}) = Dir(\lambda_{\beta 0})$$
$$P(\mu_{ij}|\Gamma_{ij}) = N(\rho^0, \xi^0 \Gamma_{ij}) \quad P(\Gamma_{ij}) = W(\nu_0, \Phi_0) \tag{15}$$

où Dir désigne une distribution de Dirichlet, N une distribution normale et W une distribution de Wishart et $\{\lambda_{\alpha 0}, \lambda_{\beta 0}, \rho^0, \xi^0, \nu_0, \Phi_0\}$ sont les hyperparamètres du modèle. Ensuite, il est possible d'appliquer l'algorithme de type EM qui consiste à alterner itérativement (13) et (14). En développant (14), il est possible de trouver des formules semblables à (4) et (5) qu' on désignera $\tilde{\gamma}_{x_t=j}$ et $\tilde{\gamma}_{z_{tp}=i|x_t=j}$. Dans le pas M, nous savons que les distribution a posteriori sont de forme semblable aux distributions a priori. Les formules de réestimation des paramètres sont données par les formules (6)-(8) où on utilise $\tilde{\gamma}$ a la place de γ. Les formules de réestimation des hyperparamètres sont données par (voir [4]) :

$$\lambda_{\alpha_j} = \sum_{t=1}^{T} N_j + \lambda_{\alpha 0} \;\; \nu_{ij} = N_{ij} + \nu_0 \;\; \lambda_{\beta_{ij}} = N_{ij} + \lambda_{\beta 0} \tag{16}$$

$$\rho_{ij} = \frac{N_{ij}\,\mu_{ij} + \xi_0\,\rho_0}{N_{ij} + \rho_0} \qquad \xi_{ij} = N_{ij} + xi_0 \tag{17}$$

$$\Phi_{ij} = N_{ij}\,\Gamma_{ij} + \frac{N_{ij} xi_0 (\mu_{ij} - \rho_0)(\mu_{ij} - \rho_0)^T}{N_{ij} + \rho_0} + \Phi_0 \tag{18}$$

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

4.1. Sélection des modèles en utilisant VB

Une propriété particulièrement intéressante de l'apprentissage bayesien variationnel est la possibilité de sélectionner les modèles durant l'entrainement. Comme exposé à la section précédente, l'énergie libre (10) peut être utilisée comme critère de sélection des modèles parce que la distance KL entre les distributions des paramètres a priori et a posteriori agit comme une pénalité comme BIC. Considérons maintenant ce problème de façon plus précise. Soit la densité a posteriori $q(m)$ pour un modèle m. On peut montrer (voir [4]) que la densité optimale $q(m)$ peut être écrite :

$$q(m) \propto exp\{F(\Theta, X, m)\}\, p(m) \tag{19}$$

où $p(m)$ est la densité a priori du modèle. En l'absence de toute information a priori sur le modèle, $p(m)$ est uniforme et la densité optimale $q(m)$ dépendra simplement du facteur $F(\Theta, X, m)$ c'est à dire que l'énergie libre peut être utilisée comme critère de sélection. Un avantage important est qu'aucun seuil ne doit être choisi manuellement (comme par exemple pour BIC). Pour le modèle considéré ici, il est possible d'obtenir une forme explicite de l'énergie libre (10). Comme décrit plus haut, un autre point intéressant dans l'utilisation de l'apprentissage VB est la capacité de réduire les degrés de liberté excédentaires. Cela signifie qu'il est possible d'initialiser le système avec un grand nombre de groupes et un grand nombre de gaussiennes par locuteur et de laisser le système éliminer groupes et gaussiennes non utilisés. Dans les modèles basés sur les gaussiennes, cette capacité d'éliminer des groupes et des gaussiennes est sous le contrôle du paramètre a priori Φ_0 qui semble être le paramètre le plus sensible pour le résultat du regroupement (voir [6]). En d 'autres mots, de grandes valeurs de Φ_0 conduiront à un nombre plus petit de groupes et de gaussiennes. Le besoin de différentes tailles de modèles de mélanges de gaussiennes a été décrit en [13] : le critère BIC est appliqué deux fois afin de choisir le meilleur nombre de groupes et de modèle par locuteur. En fait les locuteurs qui sont plus représentés dans l'enregistrement bénéficient d'un nombre plus élevé de composantes gaussiennes tandis que les locuteurs qui n'ont qu'un faible nombre de segments bénéficieront d'un plus petit nombre de composantes gaussiennes. VB réduit automatiquement et simultanément les groupes et le nombre de gaussiennes par modèle de sorte qu'il en résulte des modèles plus petits lorsque peu d'observations sont disponibles et des modèles plus sophistiqués lorsque le volume d'observations le permet.

5. EXPÉRIENCES

La base de données utilisée pour nos tests est l'ensemble de données d'évaluation NIST 1996 HUB-4 qui consiste en 4 fichiers d'environ une demi-heure. On utilise des coéfficients acousitques LPCC. Le premier fichier contient 7 locuteurs, le second 13, le troisième 15 et le quatrième 21. Dans les fichiers 1,3 et 4 il y a une part importante d'événements non-parlés tandis que le fichier 2 est de la parole presque pure.

5.1. Critère d'évaluation

Afin d'évaluer la qualité du regroupement, nous utilisons les concepts de pureté de groupe et de pureté par locuteur introduites respectivement en [9] et [3]. Nous considérons dans notre test un groupe supplémentaire pour les événements non-parole. Utilisant la même notation qu'en [3], définissons :

- R : Nombre de locuteurs
- S : Nombre de groupes
- n_{ij} : Nombre de trames dans le groupe i prononcée par le locuteur j
- $n_{.j}$: Nombre total de trames prononcées par le locuteur j, ($j = 0$ signifie trames non-parole)
- $n_{i.}$: Nombre total de trames dans le groupe i
- N : Nombre total de trames dans l'enregistrement
- N_s : Nombre total de trams parole dans l'enregistrement.

Il est possible de définir la pureté de groupe i p_i et la pureté en locuteur j p_j

$$p_i = \sum_{j=0}^{R} \frac{n_{ij}^2}{n_i^2} \quad p_j = \sum_{i=0}^{S} \frac{n_{ij}^2}{n_j^2} \tag{20}$$

Suivent les définitions de la pureté de groupe moyenne (acp : average cluster purity) et la pureté moyenne par locuteur (asp : average speaker purity) et leur moyenne géométrique K :

$$acp = \frac{1}{N} \sum_{i=0}^{S} p_i\, n_i, \quad asp = \frac{1}{N_s} \sum_{j=1}^{R} p_j\, n_j, \quad K = \sqrt{asp \cdot acp} \tag{21}$$

TAB. 1: Résultats du regroupement : système base vs. système VB I (nombre de groupes connu à priori) vs. système VB II (initialisé avec 30 groupes)

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

5.2. Résultats

Dans cette section nous décrivons notre premier ensemble d'expériences dans lequel le nombre de locuteurs est connu a priori. Nous trouvons qu'imposer une contrainte de durée de 100 trames (soit 1 sec) et 15 gaussiennes par état suffit à garantir une identification robuste des locuteurs. Pour le système VB, nous utilisons les paramètres suivants : $\lambda_{\alpha 0} = \lambda_{\beta 0} = 1$, $\rho_0 = \bar{O}$, $\xi_0 = 1$, $\Phi_0 = 200$ et $\nu_0 = g$ où g est la dimension, de l'espace acoustique et $\bar{O}$ est la valeur moyenne des observations.

Dans un premier ensemble d'expériences, nous fixons le nombre de groupes égal au nombre de locuteurs plus 1 pour tenir compte des trames non-parole comme en [3], et ensuite nous entraînons le système en utilisant les apprentissages EM/ML et VB. Les résultats sont présentés aux première et seconde lignes du tableau 1. Sur les deux premiers enregistrements VB surpasse la méthode classique EM/ML tandis que pour les deux derniers, les deux techniques sont équivalentes. Nous pensons qu'il y a principalement deux raisons pour expliquer les meilleures performances de l'apprentissage VB par rapport à l'apprentissage EM /ML. D'une part, les paramètres finaux tirent avantage de l'effet de régularisation venant des distributions a priori. D'autre part, le système VB converge pour chaque locuteur vers un GMM qui peut avoir un nombre de composantes moindre que le modèle original (15 composantes dans notre cas) selon nombre d'observations par locuteur. Il en résulte généralement une plus grande pureté par locuteur comme le montre le tableau (1).

Dans un second jeu d'expériences, nous avons initialisé le modèle avec un nombre élevé de groupes (30) et avons laissé l'apprentissage VB réduire automatiquement le nombre de groupes vers un nombre final de groupes inférieur. En réalité, dans les enregistrements HUB-4, il y a énormément de segments non-parole qui influencent radicalement le regroupement. Pour les trois premiers enregistrements, le système surpasse le système de base tandis que pour le quatrième, les performances sont similaires. Analysons les résultats enregistrement par enregistrement :

Enregistrement 1 : le nombre final de groupes est plus élevé que le nombre réel de locuteurs ; de fait les groupes excédentaires organisent les différents événements non-parole. Le regroupement final montre une amélioration de asp et acp.

Enregistrement 2 : le nombre final de groupes est proche du nombre réel de locuteurs. On observe que cet enregistrement est très majoritairement composé d'événements parole. Les valeurs de asp et acp sont très élevées.

Enregistrement 3 : à nouveau, le nombre de groupes est proche du nombre de locuteurs et le score final est très élevé (asp=0.9).

Enregistrement 4 : dans cette expérience, plusieurs locuteurs sont regroupés dans un même groupe. Cela est sans doute du à un facteur de réduction Φ_0 trop élevé et au fait que dans cet enregistrement , de nombreux locuteurs ont des temps de parole très brefs. Cependant les performances en termes de acp et d'asp sont proches du système de base.

6. CONCLUSION ET FUTURS TRAVAUX

Dans cette contribution, nous avons appliqué avec succès l'apprentissage bayesien variationnel au regroupement non-supervisé de locuteurs. Des expériences sur la base de données d'évaluation NIST 1996 HUB-4 montrent que l'apprentissage VB peut surpasser le système de référence. Cette conclusion mérite cependant quelques considérations. Tout d' abord le système tirerait assurément avantage d'une discrimination préliminaire entre parole et non-parole parce que les événements non-parole perturbent souvent le regroupement. Ensuite dans la méthode que nous proposons, nous laissons le système se réduire lui-même au nombre correct de groupes. Cette technique est très sensible aux maxima locaux. Une solution possible consiste à surgrouper les données (c'est à dire à utiliser une faible valeur de Φ_0 qui conservera plus de groupes) et ensuite d'essayer de fusionner les groupes en utilisant la borne VB (expression 10) comme mesure.

RÉFÉRENCES

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

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

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

[4] H. Attias , "A Variational Bayesian framework for graphical models", *Advances in Neural Information Processing Systems 12, MIT Press,Cambridge, 2000*.

[5] S. Watanabe et al. "Application of the Variational Bayesian approach to speech recognition" *MIT Press, NIPS 2002*.

[6] O.-W. Kwon , T.-W. Lee ,K. Chan , "Application of variational Bayesian PCA for speech feature extraction," pp. I-825–I-828, *Proc. ICASSP 2002*.

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

[8] D.J.C. MacKay "Local Minima, symmetry breaking and model pruning in variational free energy minimization", http ://www.inference.phy.cam.ac.uk/mackay/BayesVar.html.

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

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

[11] A. Cohen et V. Lapidus " Unsupervised text independent speaker classification", *Proc. of the 18th Convention of IEEEI 1995*.

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

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