BECARS : un logiciel libre pour la vérification du locuteur

Raphaël Blouet[†]*, Chafic Mokbel*[‡] *et Gérard Chollet*[†]

[†] ENST, dépt. TSI
46 rue Barrault, 75634 Paris
France
{blouet,chollet}@tsi.enst.fr

[‡]University of Balamand
El-Koura, BP 100 Tripoli
Libanon
chafic.mokbel@balamand.edu.lb

ABSTRACT

This article presents BECARS (Balamand-ENST-CEDRE Automatic Recognition of Speakers) : a free software for training Gaussian Mixture Models (GMM). BECARS permits the use of many classical adaptation techniques (such as MAP) and proposes original one (namely MAP_TREE and MAP_TREE_SPEC). In this paper, each of them are precisely described and evaluated on the data of the NIST'2003 Speaker Verification Evaluation campaign [Przybocki, 2003].
We introduce this work with a recall of generalities on Automatic Speaker Verification (ASV). We then present main characteristics of Gaussian Mixture Models (GMM) which are the most common tool for speaker modelization in ASV system. Following is the describtion of each adaptation technique available in BECARS. We finally provide performances evaluation of each of them before concluding the paper.

1. INTRODUCTION

Soit une suite de vecteurs acoustiques $\mathbf{Y}$ extraits d'un signal de parole Y et une identité proclamée ou supposée X, on formalise le problème de la vérification du locuteur en considérant les deux hypothèses H_X et $H_{\bar{X}}$ suivantes :

H_X : l'énoncé a bien été prononcé par X,
$H_{\bar{X}}$: l'énoncé n'a pas été prononcé par X.

Dans le cadre d'une modélisation statistique des hypothèses H_X et $H_{\bar{X}}$, le score de décision optimal $S_X(\mathbf{Y})$ permettant de choisir l'une de ces deux hypothèses correspond à l'estimation du rapport de vraisemblance entre H_X et $H_{\bar{X}}$. La règle de décision s'écrit alors :

$$S_X(\mathbf{Y}) = \log \frac{p_X(\mathbf{Y})}{p_{\bar{X}}(\mathbf{Y})} \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta \qquad (1)$$

où $p_X(\mathbf{Y})$ et $p_{\bar{X}}(\mathbf{Y})$ représentent respectivement la vraisemblance des données observées suivant les hypothèses H_X et $H_{\bar{X}}$ et θ le seuil de décision.
La connaissance des probabilités *a priori* des hypothèses H_X et $H_{\bar{X}}$ permet théoriquement de fixer le seuil optimal. En pratique, celui-ci est cependant déterminé de manière empirique sur un ensemble d'évaluation et de manière à minimiser le coût de fonctionnement associé à l'application visée.
En Vérification Automatique du Locuteur (VAL), la modélisation par modèle de mélange de gaussiennes [Reynolds, 1992] est celle de l'état-de-l'art pour calculer la vraisemblance des hypothèses H_X et $H_{\bar{X}}$. Sa description fait l'objet de la section suivante.

2. MODÉLISATION PAR GMM

La densité de probabilité d'un vecteur y de dimension d suivant une loi de mélange de M gaussiennes s'écrit :

$$p(y|\lambda) = \sum_{m=1}^{M} \alpha_m . \mathcal{N}_m(y, \mu_m, \Sigma_m)$$

où α_m et $\mathcal{N}_m(y, \mu_m, \Sigma_m)$ sont le poids et la gaussienne de vecteur moyenne μ_m et de matrice de covariance Σ_m associés à la $m^{\text{ème}}$ composante du mélange. $\lambda = \{\alpha_m, \mu_m, \Sigma_m\}_{m=1}^{M}$.

L'algorithme *EM* [Dempster, 1977] permet l'estimation de ces paramètres. Cet algorithme itératif garantit la croissance de la vraisemblance des données d'apprentissage avec les itérations. Chacune d'elle est formée de deux étapes : une étape **E** (*Estimation*) où la fonction vraisemblance des données complètes étant donnés les paramètres du modèles à l'itération précédente est estimée, et une étape **M** (*Maximization*) où une nouvelle estimation des paramètres du modèle est obtenue en maximisant la fonction de vraisemblance précédente. La qualité des paramètres estimés et donc de la modélisation dépend de la quantité et de la représentativité des données d'apprentissage.

Le modèle associé à l'hypothèse $H_{\bar{X}}$ est fréquemment appelé modèle du monde car il est censé représenter tous les locuteurs autres que le locuteur du test courant. Les paramètres de ce modèle doivent être estimés dans cette perspective et l'on utilise généralement une grande quantité de parole produite par de nombreux locuteurs. Ainsi, on dispose d'une quantité suffisante de données pour apprendre ce modèle. Ceci n'est généralement pas le cas pour estimer les paramètres du modèle associé à H_X (appelé modèle client). Il s'agit alors d'un apprentissage non supervisé à partir d'un ensemble réduit de données et dont on ne peut contrôler la qualité. La meilleure solution consiste alors à ajuster les paramètres du modèle du monde pour décrire au mieux les données du client. Le logiciel BECARS[1] propose un ensemble d'exécutables qui permettent l'apprentissage des densités $p_X(\mathbf{Y})$ et $p_{\bar{X}}(\mathbf{Y})$.

Dans la section suivante nous présentons les différentes stratégies mises en place dans BECARS pour estimer les paramètres associés au hypothèse H_X et $H_{\bar{X}}$.

[1] BECARS peut être librement téléchargé à partir de http ://www.tsi.enst.fr/~blouet/Becars/index.html

3. ESTIMATION DES PARAMÈTRES

3.1. *Estimation des paramètres associés à* $H_{\bar{X}}$

Une fois l'ensemble d'apprentissage défini, BECARS initialise l'algorithme *EM* grâce à l'algorithme *LBG* [Linde, 1980]. À chaque itération de l'algorithme *EM*, les données sont complétées d'une manière probabiliste. Plus précisément, la probabilité (ou la vraisemblance) qu'un vecteur observé provienne d'une composante du mélange est calculée lors de l'étape *Estimation*. Puis, les paramètres du modèle sont réestimés en utilisant les formules de réestimation [Dempster, 1977] :

$$\begin{aligned}
\alpha_m^{n+1} &= \frac{1}{T}\sum_{t=1}^{T} p(y_t|c_t = m, \lambda^n) \\
\mu_m^{n+1} &= \frac{\sum_{t=1}^{T} p(y_t|c_t = m, \lambda^n)\cdot y_t}{\sum_{t=1}^{T} p(y_t|c_t = m, \lambda^n)} \\
\Sigma_m^{n+1} &= \frac{\sum_{t=1}^{T} p(y_t|c_t = m, \lambda^n)\cdot (y_t \cdot y_t^T)}{\sum_{t=1}^{T} p(y_t|c_t = m, \lambda^n)} \\
&\quad -(\mu_m^{n+1})\cdot(\mu_m^{n+1})^T
\end{aligned} \tag{2}$$

où m est l'indice de la composante du mélange, n le numéro de l'itération et c_t la composante qui se réalise à l'instant t.

3.2. *Estimation des paramètres associés à* H_X

Utilisées pour la première fois en Vérification du locuteur dans [Reynolds, 1997], les techniques d'adaptation ont été largement étudiées dans la dernière décennie. Dans [Mokbel, 2001] un cadre les unifiant a été proposé et diverses techniques dérivant de ce cadre sont implémentées dans BECARS. Dans ce cadre théorique, l'adaptation est vue comme un apprentissage contrôlé des paramètres du modèle. Le contrôle doit dépendre des données utilisées pour cette estimation ainsi que de leur qualité. Ainsi, l'adaptation est vue comme une fonction de transformation des paramètres du modèle ayant un degré de liberté variable.

La variation du degré de liberté de la fonction d'adaptation est associée à une classification variable des composantes du mélange allant d'une seule classe à un nombre de classes égal au nombre de composantes du mélange. Dans ce cas, chaque classe est un singleton formé d'une unique composante du mélange. Une structure arborescente binaire permet de définir cette classification variable des composantes du mélange. Cet arbre est construit sur les distributions gaussiennes du modèle du monde en partant des feuilles vers la racine et en regroupant à chaque fois deux des noeuds libres. À chaque étape de la construction, on regroupe les deux noeuds les plus proches au sens d'une distance. La distance $d(\mathcal{N}_1, \mathcal{N}_2)$ entre les deux gaussiennes $\mathcal{N}_1$ et $\mathcal{N}_2$ utilisée dans BECARS correspond à la perte en vraisemblance sur les données d'apprentissage si on les remplace par la distribution gaussienne $\mathcal{N}_3$ qui les représente. Cette distance correspond à :

$$d(\mathcal{N}_1, \mathcal{N}_2) = \log \frac{|\underline{\underline{\Sigma_3}}|^{\frac{\alpha_1+\alpha_2}{2}}}{|\underline{\underline{\Sigma_1}}|^{\frac{\alpha_1}{2}}|\underline{\underline{\Sigma_2}}|^{\frac{\alpha_2}{2}}}$$

Une fois l'arbre construit, on applique un algorithme *EM* modifié pour effectuer l'adaptation.

À l'étape **E** les données acoustiques sont complétées d'une manière probabiliste. Les poids d'occupation des gaussiennes sont ensuite remontés dans l'arbre des feuilles jusqu'à la racine en s'ajoutant en passant des noeuds fils au noeud parent. Ensuite, l'arbre est reparcouru de la racine vers les feuilles en s'arrêtant à un noeud si l'un des noeuds successeurs possède un poids inférieur à un seuil fixé *a priori*. Ceci permet d'obtenir une classification et donc de définir le degré de liberté. Cette procédure, décrite sur la figure 1 permet d'assurer que chaque classe possède suffisamment de données et d'adapter correctement les paramètres des distributions gaussiennes membres.

À l'étape **M** les paramètres des distributions gaussiennes d'une classe sont réestimés en supposant une transformation par distribution gaussienne mais que ces transformation partagent les mêmes paramètres par classe. Ces paramètres des transformations sont estimés en utilisant le critère Maximum A Posteriori (MAP). Au cas où l'on suppose des distributions gaussiennes à matrices de covariance diagonales, on utilise des régressions linéaires sur chaque dimension avec comme paramètres (a_i, b_i) tels que :

$$\begin{aligned}
\mu_i &= a_i m_i + b_i \\
\sigma_i^2 &= a_i^2 s_i^2
\end{aligned} \tag{3}$$

où m_i est la moyenne *a priori* pour la $i^{\text{ème}}$ distribution qui peut être prise comme la moyenne correspondante du modèle du monde, et s_i^2 la variance correspondante. Pour une classe q, les fonctions d'estimation dans le cadre le plus générale des paramètres de transformation par régression sont données dans [Mokbel, 2001]. Dans BECARS quatre cas particuliers sont implémentés :

- MLLR_MAP
- MAP
- MAP_TREE
- MAP_TREE_SPEC

Leur description fait l'objet des quatre sous-sections suivantes.

MLLR_MAP : Dans le cas MLLR_MAP, les coefficients de régression sont réestimés par :

$$b_q = \frac{\sum_{j=1}^{J}[r0_j \cdot n_j \cdot (\bar{y}_j - a_q \cdot m_j)]}{\sum_{j=1}^{J}[r0_j \cdot n_j]} \tag{4}$$

$$\begin{aligned}
0 &= |a_q^2|\cdot\left[\sum_{j=1}^{J} n_j\right] + |a_q|\cdot\left[\sum_{j=1}^{J} r0_j \cdot n_j \cdot m_j \cdot \bar{y}_j \right. \\
&\quad \left. - \frac{\left(\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y}_j\right)\cdot\left(\sum_{k=1}^{J} r0_k \cdot n_k \cdot m_k\right)}{\sum_{j=1}^{J} r0_j \cdot n_j}\right] \\
&\quad - \left[\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y^2}_j - \frac{\left(\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y}_j\right)^2}{\sum_{j=1}^{J} r0_j \cdot n_j}\right]
\end{aligned} \tag{5}$$

où J est le nombre de distributions gaussiennes dans la classe q, n_j est le poids de la $j^{\text{ème}}$ gaussienne après l'étape **E** de l'algorithme *EM*, $r0_j$ la précision *a priori*, m_j la moyenne *a priori*, $\bar{y}_j$ et $\bar{y^2}_j$ la moyenne et le moment d'ordre 2 observés suite à l'étape **E**.

Le coefficient a_q est la solution de l'équation de second degré de l'équation 5 qui a toujours une solution [Mokbel, 2001]. L'estimation de la moyenne est alors directement obtenue par :

$$\mu_i = 0.8 \cdot (a_i m_i + b_i) + 0.2 \cdot m_i$$

MAP : Dans le cadre de l'adaptation MAP, la seule classification utilisée est celle où une distribution gaussienne par classe est utilisée. En d'autres termes on suppose comme nul le seuil pour que les données d'une classe soient considérées suffisantes. La moyenne est ainsi réestimée par :

$$\mu_i = \frac{\tau \cdot m_i + n_i \cdot \bar{y}_i}{\tau + n_i} \quad (6)$$

Cette formule d'adaptation est identique à celle proposée dans [Reynolds, 1997]. Dans les évaluations présentées à la section suivante, nous avons fixé $\tau = 14$.

MAP_TREE : MAP_TREE est une adaptation MAP où un seuil réel est utilisé pour effectuer la classification. Dans ce cas, le cadre unifié est appliqué avec un coefficient de régression constant de 1.0. Seul le biais est calculé par MAP pour l'ensemble de la classe et ensuite utilisé pour adapter les moyennes de chaque distribution gaussienne. La réestimation de ce biais se fait selon :

$$b_q = \frac{\sum_{j=1}^{J} [r0_j \cdot n_j \cdot (\bar{y}_j - m_j)]}{\sum_{j=1}^{J} [r0_j \cdot (n_j + \tau)]} \quad (7)$$

MAP_TREE_SPEC : MAP_TREE_SPEC est une version lissée de MAP_TREE. L'obtention des classes est obtenue suivant la procédure décrite au début de cette section. Ensuite, pour chaque distribution gaussienne d'une classe, on suppose fixe sa contribution dans la classe, à savoir son poids β. On suppose que l'estimation de la moyenne de la classe sur les données observées est robuste et que cette moyenne est la somme pondérée de la moyenne de la distribution gaussienne considérée et d'une autre distribution gaussienne. Finalement, la moyenne de la distribution gaussienne de la classe est estimée par maximum a posteriori afin de vérifier les hypothèses précédente et connaissant la distribution a priori de cette moyenne. Ceci permet d'obtenir une équation de réestimation de la forme :

$$\mu_j = (\bar{y}_q - m_q) \cdot \frac{\beta \cdot \tau \cdot r0_j}{(1-\beta)^2 \cdot \tau \cdot r0_j + \beta^2 \cdot \tau \cdot r0_{q-j}} + m_1 \quad (8)$$

où les mêmes notations sont utilisées et où $r0_{q-j}$ représente la précision de la distribution gaussienne complémentaire à la $j^{\text{ème}}$ distribution de la classe.

4. ÉVALUATION DES PERFORMANCES

Les différentes techniques d'adaptation proposées dans BECARS ont été évaluées sur les locuteurs féminins de l'évaluation NIST 2003 [Przybocki, 2003]. Pour chaque locuteur de cette base, la quantité de données d'apprentissage correspond à environ 2 minutes de parole spontanée, enregistrée de manière transparente lors de conversation téléphonique sur le réseau cellulaire nord-américain. Les accès de test ont une durée comprise entre 15 et 35 secondes. Enfin, on dispose dans cette base d'environ 200 locuteurs et environ 23000 tests sont effectuées. L'ensemble des donnés permettant l'estimation des paramètres du modèle associé à l'hypothèse $H_{\bar{X}}$ sont issues des données cellulaires de l'évaluation NIST 2001. On dispose ainsi d'une centaine de locuteurs pour environ 1h30 de parole.

4.1. Caractéristiques communes des systèmes de VAL

L'analyse acoustique, qui permet l'extraction de la suite de vecteurs $\mathbf{Y} = \{y_1, \cdots, y_T\}$ à partir du signal de parole Y a consisté, pour chacune des évaluations présentées ici, en :

1. Filtrage des données dans la bande de fréquence téléphonique *i.e* 300-3400 Hz,
2. Extraction de 16 coefficients cepstraux de banc de filtre et ajout des coefficents dynamiques,
3. Égalisation aveugle du canal par Feature Warping [Pelecanos, 2001].

Le nombre de composantes M utilisé pour toutes les configurations du système est fixé à 256.

Aucune des techniques de normalisation classique (du type *h-norm* ou *t-norm* [Auckenthaler, 2000]) n'a été appliquée pour obtenir les résultats que nus présentons.

4.2. Résultats

La tableau 4.2 présente les pourcentages d'erreur à l'EER (Equal Error Rate) obtenus par les différentes techniques d'adaptation implémentées dans BECARS. Sur la figure 2 sont tracées les courbes DET [Martin, 1997] associées à ces résultats.

Suivant les évaluations que nous avons réalisées, MLLR_MAP, MAP et MAP_TREE permettent d'obtenir des performances quasiment similaires et très proches de celles des meilleurs systèmes. La méthode MAP_TREE_SPEC conduit quant à elle à une légère dégradation des performances à l'EER. Ceci peut s'expliquer par le lissage trop fort qu'elle effectue sur l'estimation des paramètres. On peut espérer réduire l'écart de performance entre MAP_TREE_SPEC et les autres méthodes proposées, en utilisant un coefficient τ spécifique à MAP_TREE_SPEC pour diminuer la contribution des paramètres *a priori*. En effet, dans toutes les expériences réalisées nous avons utilisé une unique valeur en fixant $\tau = 14$.

	EER (%)
MLLR_MAP	9.5%
MAP	9.9%
MAP_TREE	9.6%
MAP_TREE_SPEC	11.1%

TAB. 1: Performances suivant l'EER et l'HTER des différentes techniques d'adaptation implémentée dans BECARS.

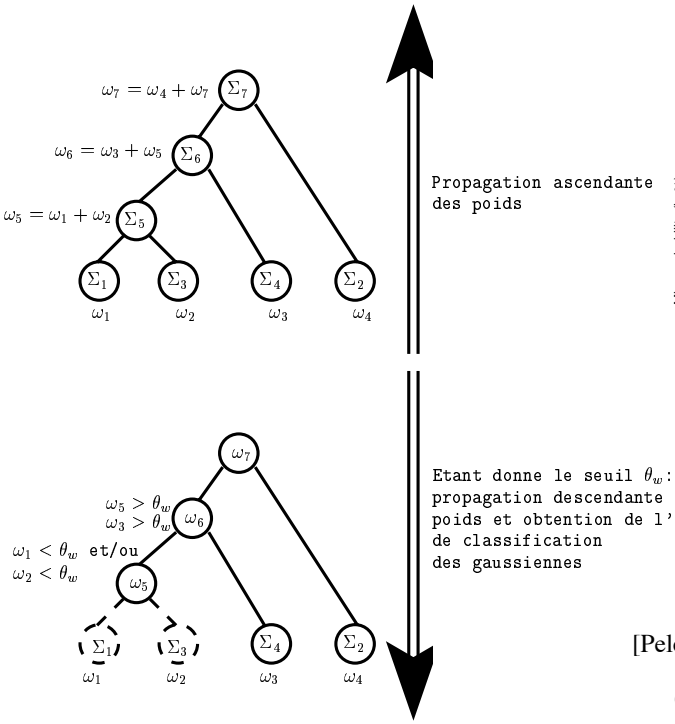

FIG. 1: Description des étapes d'obtention de l'arbre de classification des gaussiennes

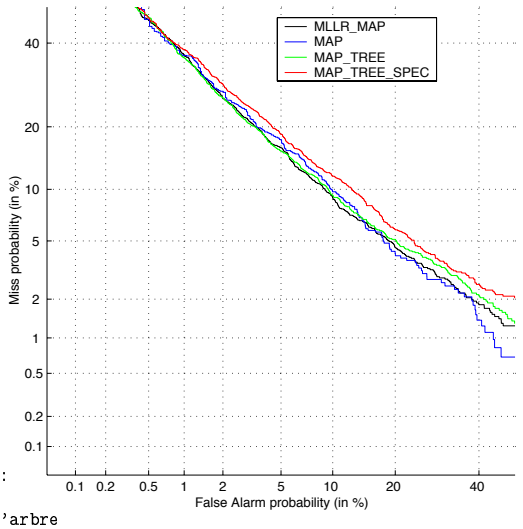

FIG. 2: Courbes DET

CONCLUSION ET PERSPECTIVES

Nous avons présenté et évalué les principales techniques d'adaptation implémentées dans BECARS, un système libre pour la vérification du locuteur. Les performances obtenues sont très proches de celles des meilleurs systèmes. Nous souhaitons maintenant évaluer le comportement de chacune des techniques d'adaptation de BECARS sous d'autres contraintes d'apprentisage et de test par exemple en diminuant les quantités de données disponible pour l'apprentissage et/ou le test.

RÉFÉRENCES

[Auckenthaler, 2000] R. Auckenthaler, M. Carey et H. Lloyd-Thomas, *Score Normalization for Text-Independant Speaker Verification Systems*, Digital Signal Processing Vol 10., Nos 1-3, Janvier 2000

[Dempster, 1977] A.P. Dempster, N.M. Laird, D.B. Rubin,*Maximum likelihood from incomplete data using the EM algorithm*, Journal of the Royal Statistical Society, 39(B), 1777.

[Linde, 1980] Y. Linde, A. Buzo et R. Gray,*An Algorithm for Vector Quantizer Design*, IEEE Transactions on Communications,1980.

[Martin, 1997] The DET Curve in Assessment of Detection Task Performance,
A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki,
EuroSpeech 1997, Proceedings Volume 4, pp. 1895-1898.

[Mokbel, 2001] C. Mokbel, *Online adaptation of hmms to real life conditions : A unified framework*, IEEE Transaction on Speech and Audio Processing, 2001.

[Pelecanos, 2001] J. Pelecanos and S. Sridharan, *Feature Warping for Robust Speaker Verification*, Workshop Odyssey, 2001.

[Przybocki, 2003] M. Przybocki et A. Martin, *The NIST Year 2003 Speaker Recognition Evaluation Plan*, 2003.

[Reynolds, 1992] , *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Georgia Institute of Technology, 1992.

[Reynolds, 1997] D.A. Reynolds, *Comparison of background normalization methods for text independent speaker verification*, Eurospeech'97, 1997.

Remerciement : Le développement du logiciel BECARS a été partiellement financé par le projet de coopération franco-libanaise CEDRE et a bénéficié du soutien scientifique du consortium ELISA.