Approche immunitaire pour la classification application à la reconnaissance du locuteur

Mémoire de stage de Master Recherche - Juin 2007

Eric CHARTON

Sous la direction de Jean-François BONASTRE

Laboratoire Informatique d'Avignon
339, chemin des Meinajaries
Agroparc – B.P. 1228
F-84911 Avignon Cedex 9
eric.charton@univ-avignon.fr

RÉSUMÉ. *Nous présentons une méthode originale de classification inspirée du système immunitaire et appliquée à la reconnaissance de locuteur. La tâche que nous proposons d'explorer est l'identification du locuteur en ensemble ouvert. Notre idée et de remplacer un modèle du locuteur par un ensemble de petits modèles de non locuteurs, efficaces sur une zone réduite de l'espace de représentation. Notre approche consiste à confronter des modèles de locuteurs à bases de mélanges de gaussiennes multivariées, pour en extraire des caractères discriminants. Ces caractères sont des distributions gaussiennes multivariées extraites du GMM original. Nous utilisons ces portions de modèles pour élaborer des hyperplans séparateurs, qui constituent les lymphocytes de notre système immunitaire. En partant de ce modèle, nous proposons un ensemble de résultats préliminaires qui illustrent la faisabilité de cette approche.*

ABSTRACT. *In this report, we propose to explore an immunological approach to speaker identification in open environment. Our proposal is to view the system as a linear classifier, able to distinguish self (authorized speakers) from others. We use for our experiments a basis of gaussian mixture models. We try to build from the informations given by those GMM models, some hyperplanes that we call T-Lymphocytes, in reference to one of the mechanisms used by the immune system. Based on this analysis, we report preliminary results illustrating the feasibility of the approach.*

MOTS-CLÉS : *Classification, reconnaissance de locuteur, système immunitaire, mixtures gaussiennes.*

KEYWORDS: *Classification, speaker identification, computer immunity, gaussian mixture.*

1. Introduction

L'une des problématiques de la sécurité informatique consiste à identifier l'utilisateur et à distinguer un utilisateur légitime d'un imposteur. La nature a résolu un problème similaire depuis des millions d'années en mettant au point le système immunitaire.

Le système immunitaire est une machinerie complexe qui met en jeu de très nombreux mécanismes. Les interactions entre tous ces mécanismes sont innombrables, et demeurent d'ailleurs un champ d'investigation et d'expérimentation important des sciences du vivant.

Globalement, les mécanismes de l'immunité peuvent être séparés en deux catégories, dites *"spécifiques"* et *"non spécifiques"*, selon qu'ils ont à protéger un organisme contre un type d'intrusion pré-déterminé ou inconnu.
On aura donc des protections *"non spécifiques"* contre tout ce qui "peut venir de l'extérieur" et est considéré comme "dangereux". Le système immunitaire devra alors fonctionner selon le principe de la détection d'anomalie par rapport à l'état *"normal"* de l'organisme (un traumatisme cellulaire, par exemple), ou encore de présence de l'inconnu qui n'est pas "Soi".
Mais on aura aussi des protections spécialisées envers un type d'attaque connue : un anticorps qui réagit à une catégorie d'antigènes répertoriée comme toxique, par exemple (c'est le principe de la vaccination qui renforce le système immunitaire spécifique).

2. Etat de l'art du modèle immunitaire en informatique

Très rapidement, des chercheurs ont compris l'intérêt qu'il pourrait y avoir à modéliser une adaptation de tout ou partie du système immunitaire pour répondre à certaines tâches des systèmes d'information, et notamment celles relatives à la sécurité. On peut estimer que la recherche sur les systèmes AIS (Artificial Immune System), a débuté vers la fin des années 70 avec un article de Farmer et al [FAR 86] sur les réseaux immunitaires. Ce n'est toutefois que vers le milieu des années 90 que l'AIS devint un sujet prisé. On mentionnera les travaux de Hunt et Cooke qui initièrent un axe de recherche sur les modèles de systèmes immunitaires appliqués à la protection des réseaux [HUN 96]. Timmis et Neal ont poursuivi ce travail [NEA 00].

Plus récemment, le travail sur la "sélection clonale" en 2002 de De Castro, Von Zuben's [CAS 02], a éveillé le plus vif intérêt dans la communauté des chercheurs de l'AIS. [1]. On peut considérer que le succès de ces travaux doit beaucoup à leur proximité avec l'état de l'art de la recherche en biologie. En effet, la "sélection clonale" de De Castro *et al* est une adaptation directe des travaux théorique de Burnet et Medawar,

1. Cette communauté, de plus en plus vaste et active, se rencontre dans des conférences spécialisées : ICARIS `www.artificial-immune-systems.org`, réseau Artist `www.elec.york.ac.uk/ARTIST/`, etc

sur les implications du système immunitaire humain dans la transplantation d'organes [BUR 59], qui leur ont permis de partager le prix Nobel de Physiologie en 1960 (lire aussi [FOR 95] pour un résumé sur cette théorie et sa génèse).

Dans le domaine particulier de l'approche immunitaire appliquée à la détection d'intrusions, des travaux de mise au point de systèmes immunitaire efficaces contre les virus informatiques ont été menés dès les années 90 [KEP 94]. Dans un premier temps, la démarche était de modéliser de manière globale les grandes fonctions du système immunitaire : détecter des anomalies, évaluer le degré de probabilité qu'une anomalie soit une intrusion, archiver les signatures des intrus, enrichir une base de données de signatures. On est bien ici dans un système antiviral auto adaptatif mais peu ou pas spécialisé, et donc très éloigné de la complexité et de la richesse du véritable système immunitaire.

Par la suite, les modèles ont été affinés et seules quelques propriétés des systèmes immunitaires ont été retenues et modélisées. Dans le cadre des travaux de Forrest [FOR 94] des algorithmes de détection de modifications son proposés. Ils sont inspirés du modèle immunitaire et plus particulièrement des mécanismes de génération des lymphocytes T dans le Thymus (le Thymus est une glande située dans le médiastin antéro-supérieur, une région de la cage thoracique [SAV 06][2]). Il existe aussi des lymphocytes B nés dans la moëlle osseuse. Leur fonctionnement est utilisé dans les travaux de Hunt et Cook [HUN 96] .

2.1. *Système immunitaire à base de lymphocytes*

Les lymphocytes T sont une des nombreuses familles de cellules spécialisées présentes dans le système immunitaire. La surface d'un lymphocyte est couverte de "récepteurs" qui correspondent à des protéines étrangères (les antigènes).

Dans l'organisme, chaque lymphocyte est muni d'un sous ensemble de récepteurs qui lui est propre. Il est donc spécialisé pour une réponse immunitaire face à un petit groupe d'antigènes (la figure 1 décrit ce processus).

Les récepteurs de chaque lymphocyte sont élaborés à la suite d'un processus de sélection pseudo-aléatoire : la sélection négative. A la suite d'une phase de maturation des cellules T dans le thymus, toutes les cellules T qui correspondent à des protéines de l'organisme ("le Soi") sont détruites.

Toutes celles qui demeurent à l'issue de cette phase de sélection sont considérées comme susceptible de porter la signature d'une protéine extérieure à l'organisme (et donc potentiellement agressive). Les lymphocytes qui subsistent sont donc caractéristiques d'un "non Soi" hypothétique.

---

2. Pour une information vulgarisatrice du fonctionnement du Thymus et des lymphocytes, on pourra se reporter à `http://fr.wikipedia.org/wiki/Thymus_(anatomie)` et `http://en.wikipedia.org/wiki/Lymphocyte`

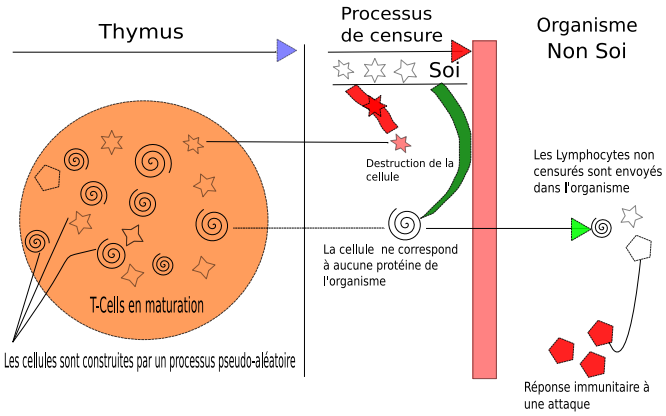

Figure 1. *Le fonctionnement du système immunitaire et de la sélection de lymphocytes dans le thymus.*

Dans les algorithmes présentés jusqu'ici [FOR 94] ce sont ces caractéristiques du système immunitaire qui sont modélisées. Elles ont pour avantage de rendre les systèmes de protection robustes : un anti-virus, par exemple, conçu selon ce modèle est capable de détecter une intrusion, sans nécessairement connaître la signature qui la caractérise [D'H 96]. L'idée de *"bases de données négatives"* en est une autre application : un ensemble de détecteurs caractérisant ce qui n'existe pas dans un jeu de données permet, par le principe du "non Soi", de prévenir ou de traiter des requêtes anormales ou des inférences aberrantes [FOR 06].

2.2. *L'algorithme de Forrest*

L'algorithme [FOR 94] que nous présentons dans les figures 2 et 3 nous fournit un exemple de système immunitaire très simple reprenant le principe des lymphocytes T, tel qu'il pourrait être appliqué à la reconnaissance de suites de données binaires ou de données textuelles.

Schématiquement, cet algorithme revient, pour une chaîne S, à caractériser le "non Soi" $\bar{S}$ probable en produisant un ensemble *"d'anti-chaînes"*. Ce "non Soi" est obtenu par génération aléatoire d'un ensemble de chaînes, suivie d'une sélection dans l'ensemble obtenu, de toutes les chaînes qui ne permettent pas, pour un nombre d'éléments consécutifs de $\bar{S}$ fixé par une constante r, de vérifier S (figure 3) . C'est par le choix de la valeur de r, et par la quantité d'éléments générés caractérisant $\bar{S}$, qu'est déterminé le degré de précision du système immunitaire.

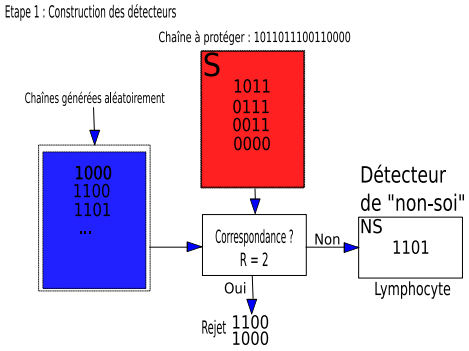

Figure 2. *Sélection de chaînes de détection du "non Soi"*

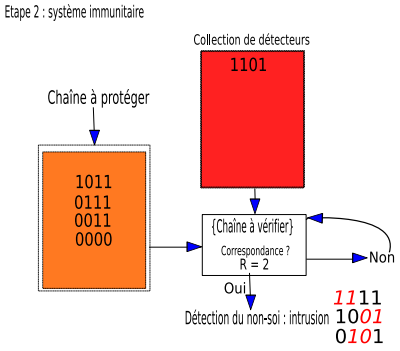

Figure 3. *Système de détection utilisant les chaînes sélectionnées*

3. Application à la reconnaissance de locuteur

L'application du principe immunitaire à la reconnaissance de locuteur suppose que nous soyons en mesure de construire des petits classifieurs simples (les lymphocytes) susceptibles de modéliser un élément qui n'est pas dans la classe recherchée. On retrouve bien ici le concept du "non Soi". Pour y parvenir, durant une phase d'apprentissage, les classifieurs apprennent à reconnaître ce que n'est pas la classe recherchée en s'appuyant sur un contre exemple.

Notre démarche est d'utiliser un système de reconnaissance de locuteur "état de l'art" [BIM 04, MAR 02, MAT 07] à base de mixture de gaussienne et d'en extrapoler un modèle de reconnaissance de type immunitaire[3].

Nous utiliserons pour cela le classifieur "Alizé" du LIA [BON 05, MAT 07]. Dans ce système de reconnaissance de locuteur à base de GMM les classes sont modélisées par des mixtures de gaussiennes multivariées. Une première proposition de méthode de reconnaissance de locuteur [REY 95], repose sur un test d'hypothèse :

– L'hypothèse $H0$ postule que l'échantillon Y appartient au locuteur L.

– L'hypothèse $H1$ postule que l'échantillon Y n'appartient pas au locuteur L.

Le test optimal de décision est donné par le ratio entre les deux hypothèses comparées à l'estimateur de vraisemblance :

$$\frac{p(Y|H0)}{p(Y|H1)} \begin{cases} \geq \theta & \text{accepter l'hypothèse H0} \\ < \theta & \text{rejeter l'hypothèse H0} \end{cases}$$

3. Les systèmes GMM que nous décrivons sont dits "Indépendants du texte", c'est-à-dire conçus pour reconnaître des locuteurs sans se préoccuper du contenu exprimé. On utilise dans les méthodes "dépendantes du texte" des modèles HMM ou GMM-HMM [RAB 89]. Ils ont pour avantage de représenter la structure temporelle du language, mais sont plus complexes, et donc coûteux à utiliser. Ils n'ont pas d'intérêt pratique dans notre cadre applicatif.

3.0.1. *Le modèle GMM-UBM*

Les modèles "état de l'art" utilisent des mixtures associées à un test de ratio basé sur la vraisemblance. Ce système dit "Système de vérification GMM UBM" est bâti sur deux classes. Un "modèle du monde" UBM [4], à savoir un GMM représentant le modèle universel de locuteurs, et un GMM par locuteur. Le test d'hypothèse [REY 00] exprime dans le domaine logarythmique est calculé comme suit :

$$\Lambda(Y) = log\, p(Y|\lambda_{USER}) - log\, p(Y|\lambda_{WORLD})$$

Ou Y est l'ensemble des vecteurs de test, λ_{USER} le modèle de l'utilisateur et λ_{WORLD} le modèle du monde. Le modèle du locuteur qui sert de base au test d'hypothèse étant adapté du modèle UBM, permet d'obtenir une méthode de scoring beaucoup plus rapide qu'avec le test d'hypothèse simple [REY 00].

Notre idée est de partir du "modèle du monde" et du "modèle de locuteur", pour construire des lymphocytes. Contrairement au paradigme classiqaue GMM-UBM, nous avons choisi de ne pas faire dériver les modèles d'utilisateurs des modèles UBM. Nous conservons en revanche un modèle UBM élaboré d'après un ensemble de données des locuteurs extérieurs à la classe des utilisateurs du système.

Nous considérons donc que la "classe recherchée" est le modèle du monde et que les contre-exemples sont issus d'un modèles de locuteur appris de manière autonome et sans adaptation. Il est aussi possible d'inverser la méthode, c'est-à-dire de considérer un modèle de locuteur comme la classe recherchée et le modèle du monde en tant que contre-exemple.

Cette différence de structuration dans le modèle exemple/contre-exemple relève du choix de l'expérimentateur et dépend uniquement de l'application recherchée.

3.1. *Le lymphocyte*

3.1.1. *Principe d'un système à base de lymphocytes*

La tâche que nous proposons d'explorer est la vérification du locuteur en ensemble ouvert. La finalité de notre application sera donc de rejeter tous les imposteurs et d'accepter tous les locuteurs connus. Par analogie, on pourrait comparer notre système à un gardien sollicité par interphone et dont la mission serait de rejeter tout individu dont la voix ne lui est pas familière. L'idée directrice sera de concevoir un ensemble de séparateurs linéaires - dits les lymphocytes - qui déterminent les frontières entre des données issues de n locuteurs enregistrés (modélisés par n modèles que nous intitulerons GMM_{USER}) et un modèle du monde (dit GMM_{UBM}) censé caractériser l'ensemble des locuteurs inconnus.

Notre idée, pour reprendre la terminologie et l'exemple des figures 2 et 3, est dans un premier temps de caractériser le "non Soi" selon le principe de l'algorithme de For-

4. Universal Backround Model ou Modèle du Monde Universel

rest. Notre méthode diffère sur la génération du "non Soi" qui est élaboré non plus à partir de méthodes pseudo aléatoires, mais d'après un modèle du monde GMM_{UBM} appris avec des échantillons non caractéristiques du "Soi". Le reste de la méthode, à savoir la sélection a postériori des caractéristiques définitives du "non Soi" après comparaison de son contenu avec le "Soi" connu, reste identique à celle proposée par Forrest.

Pour élaborer des séparateurs linéaires simples (les lymphocytes), nous comparons les densités des gaussiennes de GMM_{USER} avec celles de GMM_{UBM} (figure 4), en considérant GMM_{USER} comme un contre-exemple. Cela revient à chercher toutes les distributions gaussiennes de GMM_{USER} qui sur une dimension ou plus (figure 5) sont telles que $P(Y|UBM_{GMM})$ est le plus proche possible de 0.

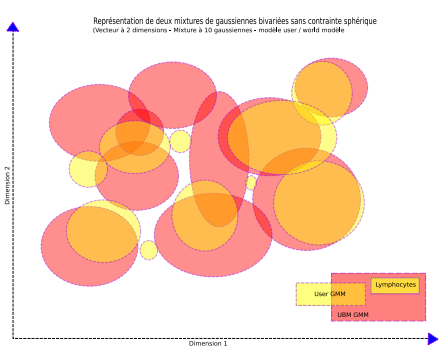

Figure 4. *Si nous observons les bases de ces deux mixtures de gaussiennes projetées dans un espace à deux dimensions, nous constatons que la plupart des distributions se superposent.*

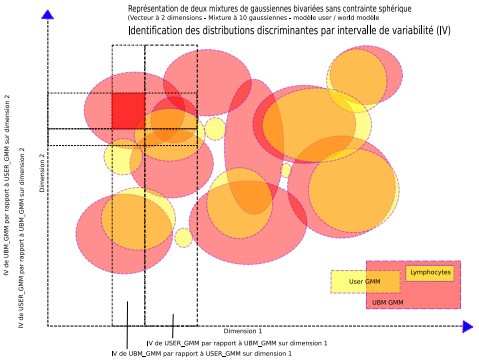

Figure 5. *En comparant les deux mixtures de gaussiennes, nous pouvons identifier des distributions qui sur une dimension ou plus, modélisent des données totalement absentes de l'autres modèle.*

Nous observons dans la figure 6 les trois distributions de GMM_{UBM} qui ne modélisent que peu de données représentées par l'ensemble des distributions du modèle user GMM_{USER}. C'est en partant de cette première information que nous élaborons des séparateurs linéaires. Notre postulat est que lorsque des échantillons de données seront projetés dans l'espace de représentation, ces derniers seront linéairement séparés et donc affectés par comptage binaire à GMM_{USER} ou GMM_{UBM} comme on l'observe dans la figure 7.

Dans la figure récapitulative 8 nous présentons le test binaire résultant : les distributions ont été retirées, seuls demeurent les séparateurs linéaires orthogonaux dont

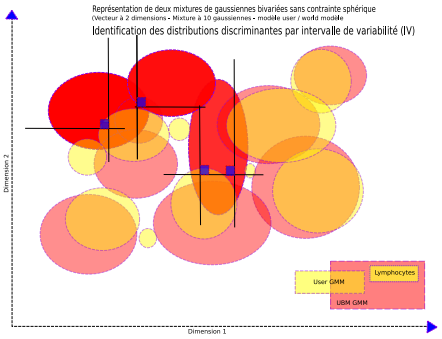

Figure 6. *Quand une distibution issue d'une mixture peut être clairement séparée de toutes les distributions du modèle qui lui est opposé nous déterminons des séparateurs linéaires.*

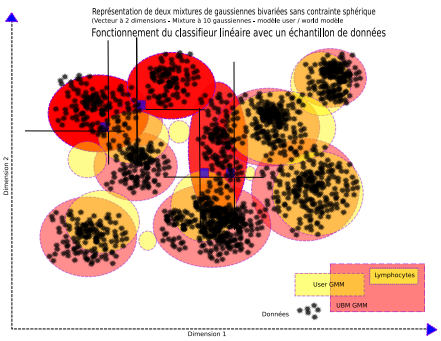

Figure 7. *En projettant des données dans l'espace de représentation, nous vérifions que seules les mixtures les plus isolées peuvent être considérées comme discriminantes.*

sont composés les lymphocytes. Dans le cas idéal présenté par cette figure [5] , on considère que l'apparition de données dans la zone délimitée par les séparateurs suffit à "activer" le lymphocyte et par voie de conséquence à affecter la population observée à la classe correspondant au lymphocyte (dans notre cadre applicatif GMM_{UBM}). On notera que dans la première version de notre algorithme, les séparateurs sont perpendiculaires aux axes. Ceci à pour effet de simplifier le comptage binaire, mais aussi de rendre plus aisée l'élaboration de séparateurs linéaires dans les espaces de grande dimensions. On espère par le choix de l'orthogonalité simplifier les équations des hyperplans. Ceci revient à perdre en précision, tout en espérant compenser cette perte par la multiplication de petits séparateurs, plus faciles à trouver et à calculer.

4. Choisir les lymphocytes

Notre modèle propose de sélectionner des distributions gaussiennes en confrontant deux modèles GMM multivariés. Nous utiliserons un modèle du monde de type GMM_{UBM} en tant que référence. Nous confronterons ce modèle à des contre-exemples, GMM_{USER} , qui sont des modèles GMM d'utilisateur appris avec peu de données. Pour simplifier l'exposé, nous présenterons la méthode comme si les gaussiennes étaient uni-dimensionnelles. Nous chercherons à localiser pour toute gaussienne $\mathcal{N}(\mu_i, \sigma_i), i = 1, ..., n$ de GMM_{USER}, un ensemble de distributions

5. Nous avons simulé ce cas théorique idéal pour valider notre modèle, en générant des échantillons aléatoires bornés, en les modélisant avec Alizé, puis en construisant des lymphocytes. Nous avons ensuite procédé à des tests. Les outils logiciels utilisés sont décrits en annexe

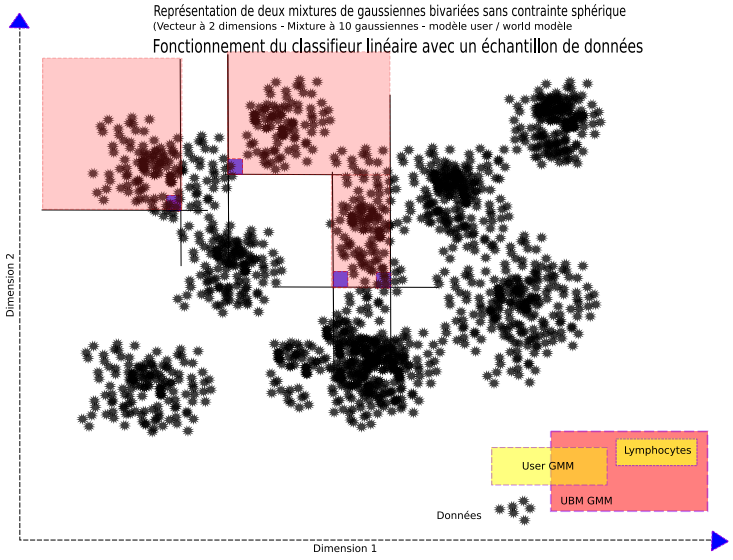

Figure 8. *Dans cette figure, nous ne conservons que les séparateurs linéaires. Nous observons qu'ils sont théoriquement suffisants par leur caractère discriminant pour identifier et classer une population.*

$\mathcal{N}(\mu_j, \sigma_j), j = 1, ..., n$, de GMM_{UBM} qui modélisent des données non représentées par GMM_{USER}. Cela revient à dire que si nous considérons A et B comme deux ensembles de distributions, nous recherchons les A_i sans intersection avec B_j avec une méthode que nous décomposerons dans la section 4.0.2. Cette recherche se fait en trois étapes :

– Nous rechercherons toutes les distributions $\mathcal{N}(\mu_j, \sigma_j)$ d'un GMM_{UBM} ne modélisant aucune des données de $\mathcal{N}(\mu_i, \sigma_i)$ et ce, dimension par dimension. Par convention, nous décrirons ces distributions comparées 2 à 2 par les termes de "similaires" ou "dissimilaires".

– Nous utiliserons une méthode d'optimisation triviale pour extraire d'une matrice contenant les résultats des mesures de similarité, les références des distributions les plus dissimilaires.

– Nous combinerons ensuite les dimensions dans lesquelles une distribution a une propriété de dissimilarité. Nous obtiendrons ainsi plusieurs distributions combinées interdimensions, qui répondent à notre critère de dissimilarité.

Pour rechercher les distributions dissimilaires, nous utilisons une mesure basée sur l'intervalle de confiance que nous appellerons l'*intervalle de variabilité*, noté IV.

Nous appellons intervale de variabilité IV au niveau $1-\alpha$ l'intervalle $[a,b]$ d'une variable aléatoire X qui est tel que :

$$P(X \leq a) = \frac{\alpha}{2} \quad et \quad P(X > b = \frac{\alpha}{2}) \tag{1}$$

L'intervalle $[a,b]$ est tel que $P(a < X \leq b = 1-\alpha)$. On dit qu'il contient une projection $1-\alpha$ de la masse de probabilité

Deux intervalles de variabilité $IV(n)$ et $IV(m)$, associés respectivement à $\mathcal{N}_A(\mu_A, \sigma_A)$ et $\mathcal{N}_B(\mu_B, \sigma_B)$ nous permettent donc dans un premier temps, de vérifier la probabilité que des échantillons d'une population, puisse simultanément appartenir à $IV(m)$ et $IV(n)$. On en déduira que pour μ, si $\mu \in [\theta_{A1}, \theta_{A2}]$ et $\mu \in [\theta_{B1}, \theta_{B2}]$ alors, $[\theta_{A1}, \theta_{A2}] \cap [\theta_{A1}, \theta_{A2}] \neq \emptyset$ et donc $\mathcal{N}_A(\mu_A, \sigma_B)$ n'est pas "dissimilaire" (ou n'a pas d'intersection vide) avec $\mathcal{N}_B(\mu_B, \sigma_B)$

Pour vérifier dans quelle mesure deux distributions peuvent êtres considérées comme dissimilaires, on procédera par itérations successives en partant d'un niveau maximal noté $\alpha - 1$, vers un niveau de confiance minimal que nous noterons $(\alpha-1)'$. Ce niveau correspond à la probabilité que l'intervalle contienne la variable aléatoire $Y = 1-\alpha$. On peut considérer que α est la probabilité d'erreur que l'on s'accorde. Pour une valeur $1-\alpha$ on construit l'intervalle de variabilité ainsi :

$$P(p_1 < p < p_2) = 1-\alpha \Longleftrightarrow P(-u_{\alpha/2} < U < u_{\alpha/2}) = 1-\alpha$$

Les coefficients multiplicateurs $u_{\alpha/2}$ sont fixés par α et sont précalculés dans les tables statistiques usuelles. En pratique nous utiliserons pour une marge d'erreur de $n\%$, la formule $1-\alpha = 1-|n\%|$, qui nous donnera par lecture de la table, la valeur du coefficient multiplicateur $u_{\alpha/2}$. Puis nous calculerons notre intervalle de confiance avec $[(\mu - (u_{\alpha/2} * \sigma)), (\mu + (u_{\alpha/2} * \sigma))]$

Les tables nous donnent les valeurs principales suivantes[6] :

- Si $1-\alpha = 15,86\%$ l'intervalle est $[(\mu - 0.2\sigma), (\mu + 0.2\sigma)]$
- Si $1-\alpha = 38,30\%$ l'intervalle est $[(\mu - 0.5\sigma), (\mu + 0.5\sigma)]$
- Si $1-\alpha = 68,26\%$ l'intervalle est $[(\mu - 1\sigma), (\mu + 1\sigma)]$
- Si $1-\alpha = 95,44\%$ l'intervalle est $[(\mu - 2\sigma), (\mu + 2\sigma)]$
- Si $1-\alpha = 99,74\%$ l'intervalle est $[(\mu - 3\sigma), (\mu + 3\sigma)]$

Pour interpréter le niveau de confiance dans l'intervalle de confiance correspondant à $1-\alpha = n\%$, on dira que n fois sur 100, l'intervalle ainsi déterminé, contient la variable aléatoire Y [BAI 02].

6. Voir par exemple `http://www.bibmath.net/formulaire/tablenormale.php3`

4.0.2. *Etape 1 recherche des dissimilarités*

Considérons un GMM_A composé de n distributions sur une dimension, et un GMM_B composé de m distributions sur une dimension. Pour un intervalle de confiance donné, nous cherchons à localiser toutes les distributions de GMM_A pour lesquelles

$$\forall i \in \{1, ..., n\}, \forall j \in \{1, ..., m\}, IV_{A_i} \cap IV_{B_j} = \emptyset \tag{2}$$

Nous calculons le taux de recouvrement de $\mathcal{N}_{A_i}$ et de $\mathcal{N}_{B_j}$ pour un intervalle de confiance $u_{\alpha/2}$ comme suit [7] :

– Nous choisissons un niveau de confiance pour comparer $\mathcal{N}_{A_i}$ et $\mathcal{N}_{B_j}$, qui nous donne après lecture des tables un coefficient de multiplication $u_{\alpha/2}$

– Nous calculons pour $\mu_{A_i} < \mu_{B_i}$ [8], $(\mu_{A_i} + (\sigma_{A_i} * u_{\alpha/2})) - (\mu_{B_j} - (\sigma_{B_j} * u_{\alpha/2}))$.

– Nous décidons que si la valeur obtenue est négative, le recouvrement est établi, et nous fixons la valeur de similarité à 1 (figure 9). Si la valeur est positive, le recouvrement n'est pas établi, donc $\mathcal{N}_{A_i}$ et $\mathcal{N}_{B_j}$ sont dissimilaires et nous fixons pour ce couple la valeur de similarité à 0 (figure 10).

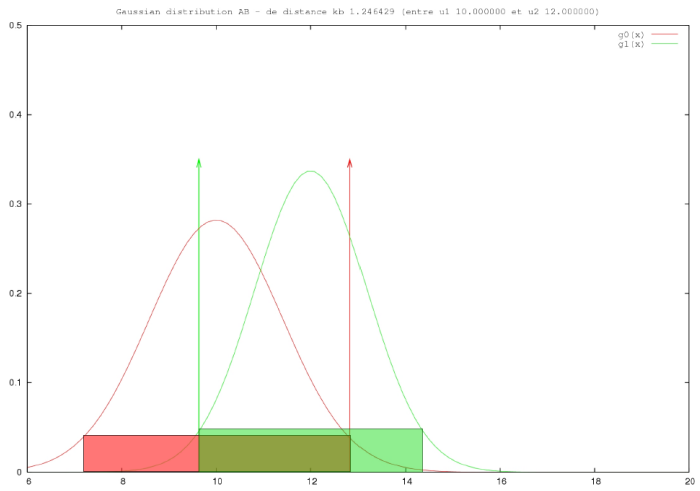

Figure 9. *Dans cette figure les deux intervalles de confiance se recouvrent :* $(\mu_A + (\sigma_A * u_{\alpha/2})) > (\mu_B - (\sigma_B * u_{\alpha/2}))$. *Les deux intervalles se recouvrent. Notre mesure de similarité est fixée à 1.*

Lorsque nous introduisons un échantillon Y, cette suite d'étapes revient, pour un niveau $1 - \alpha$ donné, à attribuer à toute distribution i du GMM_A, confrontée à toutes

7. nous proposons un logiciel utilitaire de calcul de distance et de recouvrement d'intervalles de confiance entre deux distributions gaussiennes. Ce dernier est présenté dans l'annexe B.3
8. La formule pour $\mu_A > \mu_B$ est $(\mu_{A_i} - (\sigma_{A_i} * u_{\alpha/2})) - (\mu_{B_j} + (\sigma_{B_j} * u_{\alpha/2}))$

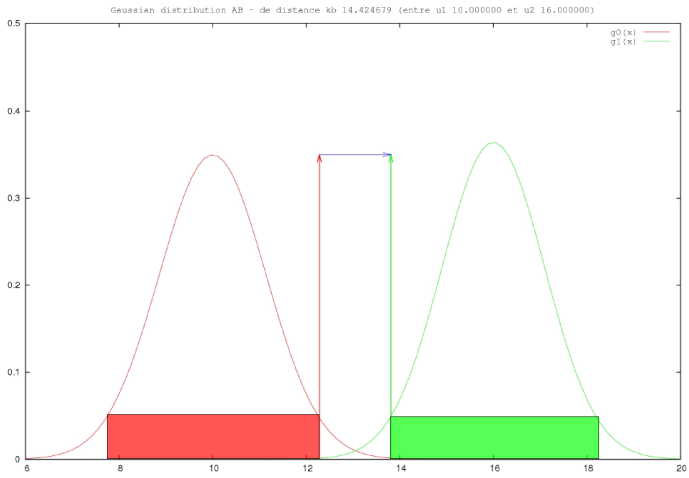

Figure 10. *Dans cette figure les deux intervalles de confiance ne se recouvrent pas :* $(\mu_A + (\sigma_A * u_{\alpha/2})) < (\mu_B - (\sigma_B * u_{\alpha/2}))$. *Les deux intervalles n'ont pas d'espace commun. Notre mesure de similarité est fixée à 0.*

les distributions j du GMM_B, l'indice de dissimilarité $D_{A_iB_j}$. Cet indice est de type binaire et prend les valeurs suivantes :

– si $P(Y \in IV_{A_i}|GMM_A) > 0$ et $P(Y \in IV_{B_j}|GMM_B) > 0$ alors $D_{A_iB_j} = 1$
– si $P(Y \in IV_{A_i}|GMM_A) = 0$ et $P(Y \in IV_{B_j}|GMM_B) > 0$ alors $D_{A_iB_j} = 1$
– si $P(Y \in IV_{A_i}|GMM_A) > 0$ et $P(Y \in IV_{B_j}|GMM_B) = 0$ alors $D_{A_iB_j} = 0$

Nous construisons avec ces indices, une matrice $D_{AB}(n, m)$ qui nous indique que pour tout A_i si cette dernière est dissimilaire de toute B_j, dans un intervalle de confiance donné. Le paramètre de cet intervalle est conservé dans une matrice C_{AB}. Nous cherchons ensuite à construire la matrice de référence des distributions dissimilaires du GMM_A pour chaque $\mathcal{N}_{A_i}$ comparée à toute distribution du GMM_B.

Nous évaluons si une distribution i de GMM_A est dissimilaire de toutes les distributions j de GMM_B (et donc considérée comme isolée et retenue pour élaborer un lymphocyte) en définissant un indice de dissimilarité D_i. Cet indice est obtenu en additionnant pour chaque indice i de la matrice, la totalité des indices $D_{A_iB_j}$. Si aucune B_j n'était similaire à A_i, alors la somme sera égale à 0, et on pourra considérer que A_i est isolée de tous les $B_{j,j=1,\ldots,n_B}$

$$\forall i, D_i = \sum_{j=1}^{n_B} D_{AB}(i, j) \tag{3}$$

Où, pour un intervalle de variabilité donné, si $D_i = 0$ nous conservons la distributions pour un lymphocite par ce que dissimilaire, et si $D_i \neq 0$ nous ne la retenons pas.

Pour rendre plus aisée la compréhension, nous donnons un exemple appliqué dans le tableau 1. Nous considérons un GMM_A à une dimension et 3 distributions ($i = 3$), comparé à un GMM_B à une dimension et 4 distributions ($j = 4$) et la matrice de dissimilarité résultante.

GMM	A1	A2	A3	Dissimilarité de B sur A
B1	0	1	1	>1
B2	0	1	1	>1
B3	0	0	0	>1
B4	0	1	0	>1
Dissimilarité de A sur B	0	>1	>1	

Tableau 1. *On déduit de l'observation de cette matrice que l'intervalle de confiance de la distribution 1 du GMM_A n'inclut aucune des valeurs comprises dans les intervalles de confiance des distributions 1 à 4 du GMM_B. La somme de la colonne est donc égale à 0 et la distribution A1 est retenue pour élaborer les séparateurs. Toutes les autres distributions, en revanche, ont une probabilité plus ou moins élevée de modéliser les mêmes données dans les mêmes intervalles de confiance (généralement par ce que leurs moyennes sont très proches), et ne sont donc pas retenues.*

4.1. *Itérations des valeurs de l'intervalle de variabilité*

L'intervalle de variabilité est calculé par l'application d'un coefficient $u_{\alpha/2}$ multiplicateur à σ. On obtient $u_{\alpha/2}$ après lecture d'une table de la loi normale centrée réduite dans laquelle ces valeurs de coefficients sont précalculées.

Par exemple, pour un niveau de confiance de $99,99\%$, la table nous propose $u_{\alpha/2} = 3,2$. La plus petite valeur de coefficient que nous puissons obtenir avec les tables est de $0,1$ correspondant à un niveau de confiance de $7,9\%$. On utilise les valeurs $u_{\alpha/2}$ de ces tables de la plus grande, jusqu'à la plus petite, jusqu'à trouver la valeur $argmin(u_{\alpha/2})$ pour $IV_{A_i} \in IV_{B_j} = \emptyset$. On archivera alors la valeur $argmin(u_{\alpha/2})$ obtenue en lui associant le numéro de distribution avec laquelle elle a été obtenue. Ceci permet d'associer à chaque distribution i dissimilaire, le niveau de confiance précis ayant permis de la définir comme dissimilaire.

Dans notre cadre applicatif la finalité est de mesurer raisonnablement, dans un espace de représentation basé sur des GMM, le degré de séparation de deux lois normales. En conséquence, il n'y a aucun intérêt à choisir une valeur $u_{\alpha/2}$ élevée, qui aurait pour conséquence de produire un IV incluant des valeurs d'échantillons dont la probabilité d'apparaître pour la distribution concernée est faible ou très faible.

En conséquence, nous avons borné $argmax(u_{\alpha/2})$ à 3 (soit un niveau de confiance de $99,74\%$ ce qui donne une probabilité d'erreur correspondant de $0,0026$ que des données puissent être identiques dans deux échantillons de locuteurs, n'appartenant pas à la même classe, quand deux intervalles de variabilité de deux distributions, n'incluent pas de valeurs commune).

Mais l'espace de représentation, hors du cas de variables aléatoires très différenciées (ce que ne sont pas les données issues de la parole) fournit très rarement ce cas presque idéal.

En pratique, dans un GMM conçu d'après des échantillons de parole, les distributions se recouvrent largement et nous devons, pour sélectionner un nombre le plus important possible de caractères discriminants, réduire progressivement l'amplitude de l'intervalle de variabilité. Nous procédons par itérations avec décrémentation de 1 pour $u_{\alpha/2} \in \{3; 2; 1\}$ puis par décrémentation à chaque itération de $0,1$ pour tout $u_{\alpha/2} < 1$ (voir illustration 11). En moyenne, nous avons pu observer que notre algorithme permet selon ce mode opératoire d'isoler des distributions jugées dissimilaires dans les proportions suivantes :

– Un modèle entrainé sur les données NIST de nos expériences peut contenir 30 à 35% de distributions isolées. Ces 35% de distributions sont réparties comme ci-dessous.

– 0,78% des distributions pour $u_{\alpha/2} = 3$

– 3,68% des distributions pour $u_{\alpha/2} = 2$

– 13,28% des distributions pour $u_{\alpha/2} = 1$

– 82% des distributions pour un $1 > u_{\alpha/2} > 0,09$ dont 26% pour $u_{\alpha/2} = 0,1$, 14.67% pour $u_{\alpha/2} = 0,2$, 6% pour $u_{\alpha/2} = 0,5$, 3,26% pour $u_{\alpha/2} = 0,9$

4.2. *Combinaisons interdimensions dans le cas multivarié*

Nous pouvons élaborer, comme nous le verrons plus loin, un détecteur par discriminant linéaire dans l'espace monodimmensionnel matérialisé par les distributions du GMM vues de façon unidimensionnelles. Le séparateur linéaire est alors un point dans un espace à une dimension. Mais nous pouvons aussi construire des regroupements sur plusieurs dimensions des lymphocytes pour augmenter leur caractère discriminant. Nous aurons alors un sous-ensemble de composantes d'une distribution multivariée, retenues pour leur dissimilarité dans leur dimensions.

Dans le cas multivarié, pour chaque dimension $k, k = 1, ..., o$ du GMM_{UBM} nous avons un vecteur de dissimilarité binaire qui nous indique si la distribution j du GMM_{UBM} est dissimilaire par rapport à toutes les distributions i du GMM_{USER}. Nous regroupons ces informations dans une matrice de k lignes et w colonnes (ou w est égal au nombre de distribution j de GMM_UBM) contenant 0 ou 1 pour chaque distribution, dans chaque dimension k, soit $A'(k, w)$. Nous conservons par ailleurs

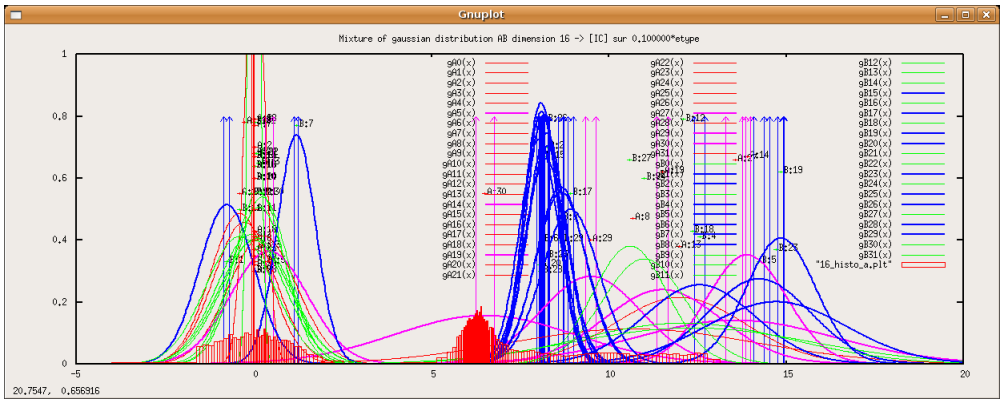

Figure 11. *Un exemple de distributions issues de deux modèles GMM comparés dans une dimension. En trait fin, les distributions non retenues, en trait épais, celles jugées isolées de leurs voisines pour un intervalle de confiance donné. Les vecteurs verticaux délimitent l'intervalle de variabilité retenu pour comparer les distributions avec leurs voisines. L'histogramme, en rouge, correspond aux données ayant servi à construire le GMM_A affiché en rouge et violet*

dans une matrice $IV(k, w)$, les paramètres $u_{\alpha/2}$ des intervalles de variabilité que nous avions utilisés pour attribuer ou non à une distribution un caractère de dissimilarité.

L'algorithme d'exploration de la matrice est trivial [9] et conduit à la création d'un tableau de lymphocytes multidimmensionnels contenant :

– Un numéro d'index L correspondant au numéro de distribution de GMM_{UBM}

– l numéros de dimensions d'une distribution retenue comme dissimilaires conservées depuis GMM_{UBM}

– Pour chaque l, la moyenne μ_l, l'écart type σ_l, l'estimateur d'intervalle de confiance retenu IV_l

– Pour chaque L_l retenue de GMM_{UBM}, la valeur μ_i de sa plus proche voisine i dans GMM_{USER}

Nous possédons ainsi des regroupements de distributions dont la taille peut varier de 1 à k' dans la limite des k dimensions de GMM_{UBM}. Avec ces informations, si l'on fait abstraction du paramètre de densité de la distribution gaussienne, il devient possible de travailler dans espace euclidien de dimension k, dans lequel la moyenne de chaque distribution retenue devient une coordonnée du foyer d'un hyperellipsoïde. Cet hyperellipsoïde est définie dans chaque dimension par la demi-amplitude correspondant à la distance entre les bornes de l'intervalle de confiance retenu pour sélectionner la distribution.

9. voir code en C++ dans le programme ImunAliz

L'intérêt de ce regroupement est qu'il nous permet d'élaborer un hyperplan séparateur dans le sous espace k' de l'espace à k dimensions, qui nous permettra de procéder à une affection des valeurs y_i composant un échantillon de locuteur Y, à la classe UBM ou à l'une des classes $USER_n$.

5. Système et algorithme de détection

Les systèmes de détection à base de discriminants linéaires seront construits à partir des tableaux de lymphocytes obtenus. Il est possible d'élaborer d'après le vecteur de dissimilarité monodimensionnel obtenu précédemment, un système immunitaire par discriminant linéaire rudimentaire. On procèdera à un comptage des valeurs y d'un ensemble de features Y, telles que ses coordonnées soient $\mu_{L_l} + (u_{\alpha/2_L} * \sigma_L) < y < \mu_L - (u_{\alpha/2_L} * \sigma_L)$ pour tout $\mathcal{N}_L$

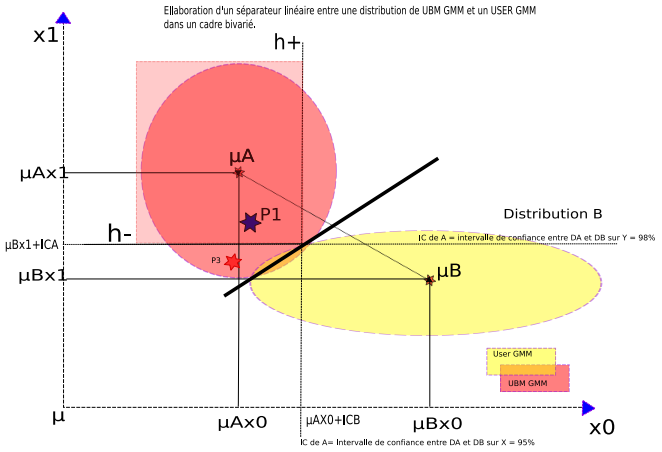

Figure 12. *Dans cette illustration, nous observons un couple de séparateurs $h+$ et $h-$ obtenus par intervalle de variabilité, ainsi qu'un séparateur linéaire optimal. Pour un coût de calcul et une complexité moindre les séparateurs $h+$ et $h-$ sont capables de discriminer des masses de données d'autant plus importantes qu'elles seront centrées sur le point μ_A. Le séparateur linéaire optimal permettrait seulement de compter des données moins fréquentes et donc moins significatives, telles que celles représentées le point $P3$.*

Dans le cas multivarié, les hyperplans séparateurs orthogonaux h_- et h_+ sont élaborés d'après chaque lymphocyte L et ont pour coordonnées $h_{k'}-$ et $h_{k'}+$ sur k' dimensions s'ils sont entourés d'autres dimensions (si la distribution est isolée dans une direction, un seul des deux séparateur $h-$ ou $h+$ est utilisé, comme dans la figure

12) :

$$h_{L-}\begin{pmatrix} \mu_1 - (u_{\alpha/2} * \sigma_1) \\ \mu_{...} - (u_{\alpha/2} * \sigma_{...}) \\ \mu'_k - (u_{\alpha/2} * \sigma'_k) \end{pmatrix} h_{L+}\begin{pmatrix} \mu_1 + (u_{\alpha/2} * \sigma_1) \\ \mu_{...} + (u_{\alpha/2} * \sigma_{...}) \\ \mu'_k + (u_{\alpha/2} * \sigma'_k) \end{pmatrix}$$

On précisera ici que les paramètres $u_{\alpha/2}$ sont symétrisés dans cette version de l'algorithme, c'est à dire que c'est la distribution GMM_{USER} la plus proche d'une distribution GMM_{UBM} qui sert de référence pour calculer l'IV. Il est possible d'envisager à terme une disymétrisation de $u_{\alpha/2}$, en positionnant différement les séparateurs $h-_k$ et $h+_k$, en fonction de la distance qui les sépare des deux distributions les plus proches, de parametre μ supérieur et inférieur.

La conception des hyperplans revient, pour une population Y de dimension k, à comparer les dimensions k' des donnée y à tous les séparateurs tels que $h- < y < h+$, pour chaque k'.

Ce qui revient à situer le point W dans toutes les dimensions de l'espace, par rapport à l'hyperplan (voir figure 12).

6. Evaluation et résultats

L'objectif de notre expérience est de vérifier que des locuteurs ne sont pas des imposteurs en utilisant les lymphocytes construits d'après le modèle du monde. La finalité de cette expérimentation est de concevoir un système capable de rejeter le plus d'imposteurs possibles, en minimisant les faux rejets (qualification d'un locuteur enregistré en tant qu'imposteur). Cette approche conforme au modèle immunitaire, doit permettre d'enrichir un système de détection classique pour lui permettre de réduire au maximum le taux de fausses acceptations. Notre évaluation part de la campagne NIST standard 2006 (37400 tests de *NIST 2006 1conv_1conv*). Nous apprenons un modèle UBM-GMM complet, composé d'un modèle du monde et d'un modèle par locuteur. Les modèles de locuteurs ne sont pas adaptés du modèle du monde. Le modèle du monde est composé d'une mixture à 2048 gaussiennes. Les modèles d'utilisateurs sont composés de mixtures à 512 gaussiennes. Les résultats de cette expérience de base sont représentés dans la figure 13 par la courbe foncée.

Dans un second temps, nous réaprenons un modèle du monde à 512 gaussiennes, puis nous générons un jeux de lymphocytes par modèle d'utilisateur confronté au modèle du monde. Nous procédons ensuite à une sélection des lymphocytes les plus efficaces en comparant chacun d'eux avec la totalité des fichiers de features ayant servis à concevoir les autres modèles d'utilisateurs. A l'issue de ce processus de sélection, la moitié des lymphocytes a été écartée en raison de son caractère peu discriminant. Cette sélection est indispensable en l'état de nos travaux, pour éviter que les lymphocytes introduisent de nouveaux faux rejets dans l'expérience. Les lymphocytes restants nous ont permis de conserver 20192 tests sur les 37400 d'origines. Nous utilisons les lymphocytes pour procéder aux 20192 tests, en remplaçant l'outil ComputeTest de LIA_SPK_DET par notre application immunitaire. La finalité de ces nouveaux tests

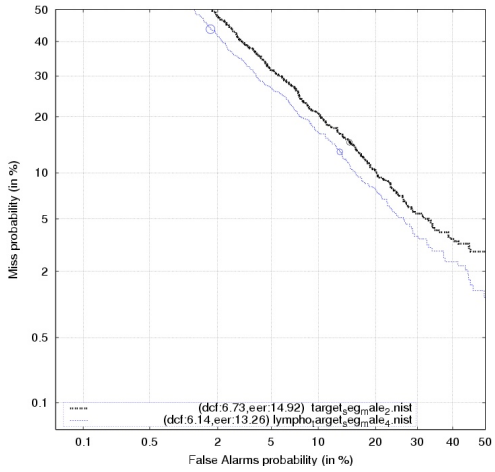

Figure 13. *Nous observons dans cette figure la courbe Det du système GMM UBM Alizé appliqué sur un jeux d'expérimentation NIST en gras, et en clair, la courbe Det obtenue après intégration des informations fournies par les lymphocytes. Nous obtenons une amélioration sensible des performances du classifieur qui détecte mieux les imposteurs.*

est de déterminer si une confrontation entre un lymphocyte et un fichier de features doit donner une décision "imposteur".

Notre système attribue la qualité d'imposteur selon deux critères :

– La quantité de lymphocytes activés en mode test, qui doit être supérieure à la quantité q_l de lymphocytes activés en mode entrainement. Nous règlons la précision du système par ajout d'une valeur v à la quantité q_l.

– Le nombre de features attribuées à un groupe de lymphocytes, qui doit être supérieur au nombre de feature q_f attribuées à un lymphocytes en mode train. Nous règlons la précision du système par multiplication de q_f par un coefficient c.

Nous retrouvons bien dans ces deux critères les fondements théoriques de notre système qui veulent qu'un lymphocyte issu du modèle du monde après confrontation avec un modèle d'utilisateur, reconnaisse plus de données pour un imposteur, que pour l'utilisateur qu'il modélise.

Nous générons ensuite un index indiquant pour chaque test si le lymphocyte a attribué à ce test la qualité d'imposteur. Dans cette expérience, 3600 tests sur 20192 réalisés se sont vus attribuer la qualité d'imposteurs par les lymphocytes. Nous utilisons cet index pour améliorer notre base d'évaluation NIST.

La méthode employée consiste à minorer les scores de vraissemblances de la baseline de test, afin de les rammener de manière certaine sous le seuil de décision du

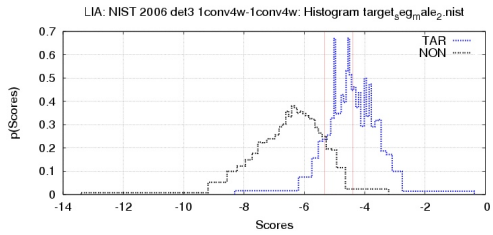

Figure 14. *Comparaison des histogrammes de scores des imposteurs et des utilisateurs avant intégration des résultats produits par les lymphocytes*

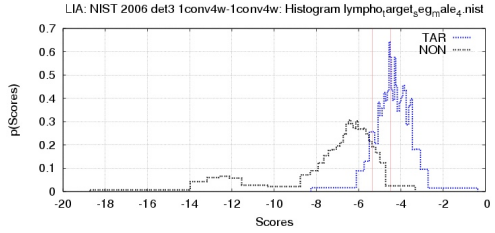

Figure 15. *Comparaison des histogrammes de scores des imposteurs et des utilisateurs après intégration des résultats produits par les lymphocytes.*

classifieur. L'évolution de la précision du système ainsi enrichi est représentée par la courbe claire de la figure 13. On observe que la Dcf passe de 6.73 dans le système de base à 6.14 dans le modèle avec introduction du système immunitaire et que le taux EER passe de 14,92 dans le système de base à 13,26 dans le système enrichi par le modèle immunitaire. Les histogrammes représentés par les figures 14 et 15 démontrent comment l'enrichissement des scores de vraissemblances obtenus avec ComputeTest, d'après l'index des résultats produits par les lymphocytes, améliore le caractère discriminant du modèle GMM.

7. Conclusions et perspectives

Nous avons conçu un modèle de reconnaissance de locuteur inspiré de l'approche immunitaire. Il repose sur l'extrapolation d'un ensemble de petits séparateurs linéaires obtenus après comparaison d'un modèle GMM UBM et d'un modèle GMM de locuteur.

Ces petits séparateurs sont constuits après extraction des distributions dissimilaires pour les deux modèles GMM comparés. La dissimilarité est mesurée avec un intervalle de variabilité. Les séparateurs sont discriminants chacun pour une région donnée et peuvent êtres combinés sur plusieurs dimensions.

Le choix de l'orthogonalité pour les séparateurs est l'une des originalités de ce système. Ce choix est motivé par la rapidité de calcul qu'elle procure, et le grand nombre de séparateurs qu'elle permet de produire. Dans ce cadre théorique, on espère compenser la diminution de précision que provoque immanquablement la contrainte de perpendicularité des séparateurs aux axes du plan, par l'augmentation d'information fournie par la multiplicité des petits séparateurs produits rapidement et simplement dans des sous-espaces d'un espace de grande dimension. A notre connaissance, s'il existe des systèmes de reconnaissance de locuteur associés à des méthodes de sélection des meilleures distributions pour élaborer un score ("Top Distrib", "Bayesian Factor Analysis"), aucun ne repose sur une sélection interdimensionnelle, conduisant à l'élaboration d'hyperplans. Les machines à vecteurs de support, dans le même contexte, posent le problème de la complexité du modèle et de la transformation de l'espace de représentation en espace de redescription. Notre méthode est donc un hybride entre les systèmes à noyaux et ceux de types gaussiens avec test de vraisemblance.

Nos expérimentations sur un ensemble d'expériences de NIST ont démontrées qu'un tel classifieur pouvait améliorer sensiblement les performances d'un classifieur UBM-GMM "état de l'art" en enrichissant ses résultats. Elles ont aussi démontrées que la difficulté du modèle immunitaire résidait dans sa capacité à réduire au maximum l'introduction de "fausses acceptations" dans le système qu'il enrichit. La sélection des meilleurs lymphocytes par un processus d'aprentissage est la solution que nous avons retenue pour remédier à cette difficulté.

7.1. *Perspectives*

Il reste un travail très important à mener pour fiabiliser ce système et lui permettre d'atteindre un niveau de performance le plus proche possible de l'état de l'art en reconnaissance de locuteur. Nous proposons plusieurs pistes de recherches pour y parvenir : prise en compte, lors de la conception des hyperplans et des calculs de scores, des probabilités de chaque classe (pour le moment, nous considérons chaque distribution du GMM comme équiprobable), dysimétrisation des séparateurs. L'approche immunitaire prévoit aussi des processus de sélection naturelle (phase d'apprentissage sur des données étiquetées). Cet aspect est peu exploré dans ce travail alors qu'il permet probablement de maximiser la détection des imposteurs tout en minimisant l'introduction de faux rejets. Dans ce dernier point réside probablement un fort potentiel d'amélioration.

8. Bibliographie

[BAI 02] BAILLARGEON G., « *Probabilité et Statistiques avec applications en technologie et ingéniérie* », chapitre 7, Editions SMG, 2002.

[BIM 04] BIMBOT F., BONASTRE J.-F., FREDOUILLE C., GRAVIER G., MAGRIN-CHAGNOLLEAU I., MEIGNER S., MERLIN T., ORTEGA-GARCIA J., PETROVSKA-DELACRETAZ D., REYNOLDS D. A., « A Tutorial on Text-Independent Speaker Verification », vol. 4, 2004, p. 430-451.

[BON 05] BONASTRE J. F., WILLS F., MEIGNER S., « ALIZE, a free toolkit for speaker recognition », Minneapolis, Minnesota, United States, 2005, IEEE International Conference, p. 737 - 740.

[BUR 59] BURNET F. M., « The Clonal Selection Theory of Acquired Immunity », , 1959, Cambridge University Press.

[CAS 02] CASTRO L. N. D., TIMIS J., *Artificial Immune Systems : A New Computational Intelligence Approach*, Springer-Verlag, 2002.

[D'H 96] D'HAESELEER P., « An immunogical Approach to Change Detection : Thorical Results », *9th IEEE Computer Security Foundations Workshop*, 1996.

[FAR 86] FARMER J., PACKARD A., PERELSON A., « The immune system, adaptation and machine learning », *"Physica D"*, vol. 22, 1986, p. 187-204.

[FOR 94] FORREST S., PERELSON A. S., ALEN L., CHERUKURI R., « Self NonSelf Discrimination in a Computer », *IEEE Symposium on Research in Securaty and Privacy*, 1994.

[FOR 95] FORSDYKE D. R., « The Origins of the Clonal Selection Theory of Immunity », *FASEB Journal*, , 1995.

[FOR 06] FORREST S., « Negative Databases », Rapport de recherche, 2006, The university of New Mexico.

[HUN 96] HUNT J. E., COOKE D. E., « Learning Using an Artificial Immune System », *Journal of Network and Computer Applications*, vol. 19, 1996.

[KEP 94] KEPHART J. O., « A Biologically Inspired Immune System for Computers », MIT Press, 1994.

[MAR 02] MARIANI J., DE MORI R., BIGI B., BIMBOT F., ET AL, *Reconnaissance de la parole*, Hermès Lavoisier, traitement automatique du langage parlé 2 édition, 2002.

[MAT 07] MATROUF D., FAUVE B., « Recent Advances in Text Independent Speaker Verification », *IEEE Transactions on Audio, Speech and Language Processing*, 2007.

[NEA 00] NEAL M., TIMMIS J., HUNT J., « An artificial immune system for data analysis », , 2000, PubMed.

[RAB 89] RABINER L. R., « A tutorial on hidden markov models and selected applications on speech recognition », vol. 77, IEEE International Conference, 1989, p. 257-286.

[REY 95] REYNOLDS D. A., ROSE R. C., « Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models », *tassp*, vol. 3, n° 1, 1995, p. 72-83.

[REY 00] REYNOLDS D. A., QUATIERI T. F., DUNN R. B., « Speaker Verification Using Adapted Gaussian Mixture Models », *dsp*, vol. 10, 2000, p. 19-41.

[SAV 06] SAVINON W., « The Thymus Is a Common Target Organ in Infectious Diseases », *Plos Pathogens*, vol. 2, 2006.

A. Annexe Logiciel ImunAliz

Le logiciel ImunAliz permet de réaliser toutes les phases d'élaboration et d'utilisation d'un système immunitaire basé sur des mixtures de gaussiennes (GMM). Il exploite les modèles GMM produits avec la librairie Alizé [10].
Le principe d'ImunAliz est de générer des séparateurs linéaires (les lymphocytes), en comparant deux modèles de locuteurs conçus avec le toolkit Lia_Spk_Det ou avec toute application compatible avec Alizé. Pour construire les lymphocytes, ImunAliz peut comparer deux mixtures multivariées dont le nombre de distributions peut prendre n'importe quelle valeur. Le nombre de dimensions doit en revanche être équivalent dans les deux modèles comparés.
Les séparateurs linéaires produits peuvent prendre plusieurs formes (mono dimmensionnelle ou combinée sur plusieurs dimensions pour créer des hyperplans). Ils peuvent ensuite être utilisés pour tester des données et décider si ces dernières appartiennent à un modèle. ImuneAliz propose également un ensemble de fonctions graphiques (génération de "plots" de mixtures, outils d'analyse graphique d'un GMM, visualisation de dissemblance entre les distributions d'un modèle, par intervalle de variabilité) et utilitaires (calcul de distances de Kullback, de Mahalanobis). ImunAliz permet donc en plus de sa fonction expérimentale de classifieur immunitaire, d'analyser en profondeur des mixtures générées par Alizé.
Il est accompagné d'un ensemble d'utilitaires qui simplifie son déploiement, ou aident à la réalisation d'expériences.

A.1. *Options de la ligne de commande*

Toutes les options d'ImunAliz sont activées d'après des options saisies en ligne de commande.

– imune –help Affichage de l'aide en ligne (liste des commandes disponibles)

– imune –config [configfilename] Définir un fichier de configuration (format Alizé)

– imune –version Affiche le numéro de version en cours d'ImunAliz

A.1.1. *Fonctions de test de feature d'après des lymphocytes*

Les fonctions de test de features d'après un modèle de lymphocytes sont les suivantes :

– imune –test Tester des features avec un groupe de lymphocytes dimension par dimension

10. ALIZE est un projet de développement issu des travaux du consortium ELISA et financé par le Ministère de la Recherche dans le cadre du programme Technolangue. Plus d'informations sur `http://www.lia.univ-avignon.fr/heberges/ALIZE/`

– imune –testmulti Tester des features avec un groupe de lymphocytes, dans un espace multidimensionnel, avec les hyperplans séparateurs

– imune –lymphoname filename Déclarer le nom du fichier de lymphocytes à utiliser

– imune –inputFeatureFilename featurefilename Indiquer le nom d'un fichier de features

– imune –ndx filename Indiquer un fichier ndx (format Alizé contenant les noms de fichiers de features à tester)

– imune –ndxtrain filename Entraîner un ensemble de lymphocytes d'après les données ayant servies à la construction de leur modèle GMM

– imune –select filename Sélectionner un sous-groupe de lymphocyte d'après toutes les features ayant servies à élaborer l'ensemble des lymphocytes

A.1.2. *Construction d'un système immunitaire*

La construction d'un système immunitaire est réalisée avec les commandes suivantes :

– imune Suivi des commandes

– –vaccine Construire un système immunitaire d'après deux modèles comparés

– –ma modelename Définir le modèle A (fichier GMM Alizé au format Raw ou XML)

– –mb modelename Définir le modèle B (fichier GMM Alizé au format Raw ou XML)

– –plotlog Sortir sur disque toutes les étapes de recherche de distributions discriminantes sous format Gnuplot (illustration)

A.1.3. *Utilitaires et analyse de modèle*

ImunAliz propose un ensemble de fonctions utilitaires et d'analyses de modèles :

– imune Suivi des commandes

– –readmodel Afficher le contenu d'un modèle en console

– –InputModelFilename [modelname] Nom du modèle à afficher

– –mesure [kb/mh] mesurer la distance entre deux distributions (Kullback et Mahalanobis) (à associer à –ma et –mb indiquant les noms des fichiers gmm à mesurer)

– –ma [modelename] Définir le modèle A (fichier GMM Alizé au format Raw ou XML)

– –mb [modelename] Définir le modèle B (fichier GMM Alizé au format Raw ou XML)

A.1.4. *Plateforme graphique*

ImunAliz propose un ensemble de fonctions graphiques au format Gnuplot.

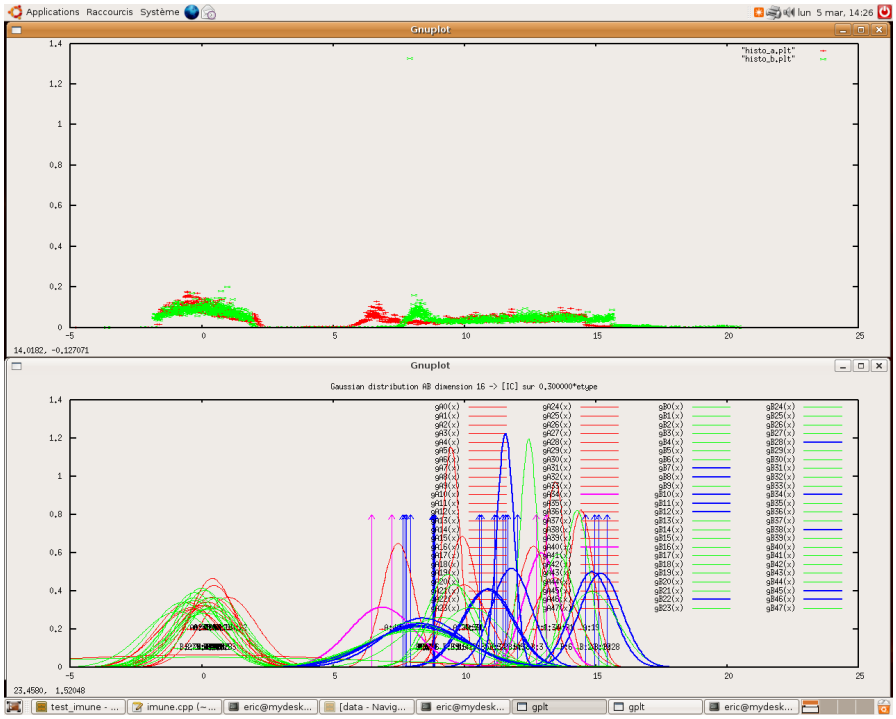

Figure 16. *Sur cette figure, on observe en haut le résultat d'un plot des points composants deux fichiers features. Ces fichiers sont à rapprocher des deux distributions comparées, en bas, qui modélisent ces données*

– –plotpath [chemin] définir le chemin de sortie des fichiers plot pour gnuplot

– –histoplot [dim_number] sortir le fichier plot de deux fichiers de features pour la dimension indiquée

– –fa [featurefilename] Nom du fichier de features a pour histoplot

– –fb [featurefilename] Nom du fichier de features b pour histoplot

– –mixplot Sortir le fichier Gnuplot d'un GMM dimension par dimension

– –InputModelFilename [modelname] Nom du modèle à afficher avec mixplot

– –bimixplot Sortir le fichier Gnuplot de deux dimensions d'un GMM (distributions bivariées)

– –mb [modeleafilename] nom du modele A pour bimixplot

– –mb [modelebfilename] nom du modele B pour bimixplot

– –da [num] dimension 1 pour bimixplot bivarié

– –db [num] dimension 2 pour bimixplot bivarié

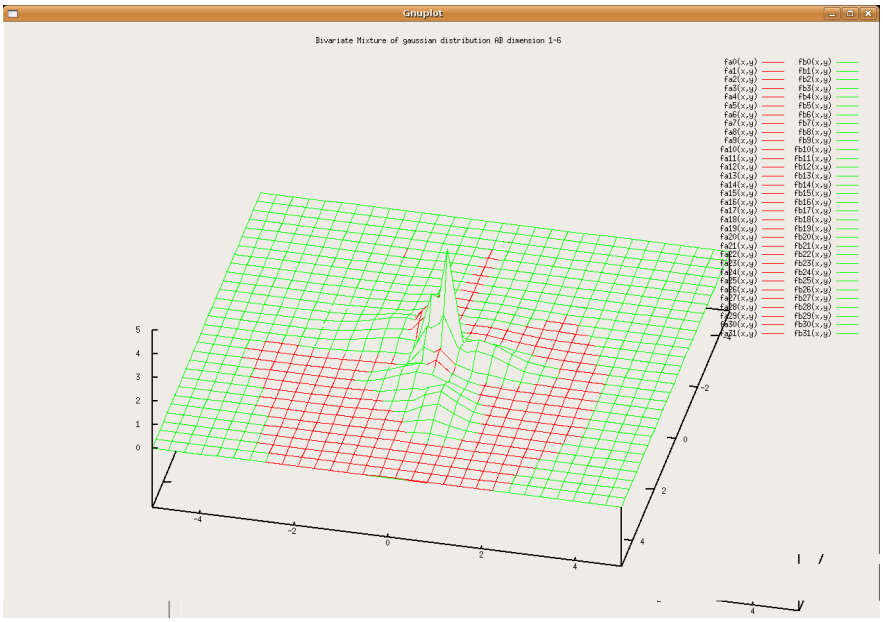

Figure 17. *Exemple de mixture affichée sous une forme bivariée.*

ImuneAliz est totalement compatible avec les options des fichiers de configuration de Alizé et du toolkit Lia_Spk_Det.

A.2. *Fonctions graphiques d'analyse*

En tant que plateforme expérimentale, le logiciel ImunAlize propose un ensemble d'options graphiques dont la finalité est de permettre l'observation des résultats qu'il produit ou des données qu'il manipule. Ces options permettent de générer des fichiers répondant au format Gnuplot (note). On pourra ainsi afficher des informations sur l'écran, ou produire des fichiers images au format EPS, intégrable dans LaTeXou plus généralement dans toute forme de présentation.

A.2.1. *Types d'images proposés*

– Les mixtures de distributions comparées

– Les "plots" d'histogrammes

– Les distributions discriminantes sélectionnés par distance binaire pour chaque itération

– Les distributions bivariée

Pour afficher un histogramme saisissez la commande gnuplot :

```
plot "n_histo_a.plt"
```

Ou n est le numéro de dimension pour lequel l'histogramme a été généré. Le fichier de points généré ne contient que des coordonnées. Pour modifier l'aspect de l'histogramme, vous pouvez utiliser toutes les commandes d'aspect de gnuplot soit :

```
plot "0_histo_a.plt" with [style]
```

Ou [style] pourra être *boxes, impulses, steps* etc ... Les histogrammes peuvent êtres affichés concomitament par une commande :

```
plot "0_histo_a.plt, "0_histo_b.plt"
```

B. Annexe Logiciels utilitaires

B.1. *test_dissemblance : Outil de simulation de comparaisons entre deux distributions*

Cet outil présent dans le répertoire testdissemblance permet de comparer deux distributions gaussiennes, de mesurer leur distance de Kullback et d'évaluer leur recouvrement par calcul de l'intervalle de confiance. Il a été utilisé pour produire les illustrations du chapitre relatif aux mesures de dissimilarités entre deux distributions. Il permet de produire un fichier au format Gnuplot, affiché sur écran ou enregistré au format .eps. Il sert également de calculateur de distance de Kullback entre deux distributions gaussiennes.

B.2. *twin.pl : calcul des scores des lymphocytes par rapport à des features*

Ce script en perl compare le fichier de scores généré par ImunAliz d'après une liste d'expérience Nist, avec les résultats obtenus pendant l'entrainement (l'entrainement des lymphocytes permet d'obtenir un ensemble de valeurs de référence après comparaison du lymphocyte avec le fichier de feature qui a servi a élaborer son modèle. Ces valeurs de références constituent en quelque sorte la signature du lymphocyte). Il n'utilise que les Lymphocytes conservés après le processus de sélection. Il détermine d'après comparaison entre les données de référence et les données obtenues lors du test, si l'ensemble feature / lymphocyte en cours correspond à un imposteur. Ce logiciel génère un fichier de résultat .res, propre à ImunAliz, et réutilisé ensuite par ly2nist.pl.

B.3. *ly2nist.pl : Outil d'enrichissement d'un fichier de résultat*

Ce script en perl permet de modifier automatiquement le fichier .nist contenant les résultats d'une expérience, en minorant les scores de vraissemblances des tests auxquels ImuneAliz a attribué la qualité d'imposteur. Il utilise pour cela le fichier .res produit par twin.pl.