Nouveau système hybride GMM-SVM pour la vérification du locuteur

Jamal Kharroubi, Gérard Chollet

ENST-TSI, CNRS-LTCI
46 rue Barrault – 75634 Paris Cedex 13, France
Tél.: ++33 (0)1 45 81 75 62 - Fax: ++33 (0)1 45 88 79 35
Mél: kharroub, chollet@tsi.enst.fr - http://www.tsi.enst.fr

ABSTRACT

Support Vector Machines (SVM) are a new and very promising technique in statistical learning theory, proposed by V.Vapnik in 1995 [Vap95]. In this article we address the issue of using the SVM technique for Text-independent Speaker verification experiments by proposing a new feature representation based on GMM to construct the input vector of the SVM. The results obtained are compared to the classical Log-Likelihood Ratio (LLR) technique on NIST2001 database, a part of the SWITCHBOARD database.

1. INTRODUCTION

Dans cet article, nous proposons une des premières applications des SVM pour la vérification du locuteur en mode indépendant du texte.

L'idée principale des SVM est de projeter des données sur un espace de plus grande dimension afin que les données non-linéairement séparables deviennent linéairement séparables.

Récemment , cette technique a prouvé son efficacité dans plusieurs applications du domaine de la reconnaissance de formes. La motivation derrière ce travail est les bons résultats obtenus par les SVM en traitement des images [Osu97][Fer99] et en fusion des experts pour l'authentification biométrique [Ben98]. La première application des SVM en reconnaissance de locuteur, plus particulièrement en identification du locuteur, a été réalisée par M. Schmit et H. Gish en 1996 [Sch96]. Dans cette application, les vecteurs des paramètres extraits directement du signal de parole ont été utilisés comme vecteurs d'entrée pour les SVM. Il est bien connu que ces vecteurs contiennent simultanément un certain nombre d'informations sur le canal, la parole et les locuteurs. Il est donc difficile d'extraire uniquement les informations des locuteurs directement de ces vecteurs sans passer par une modélisation. Dans ce travail, nous proposons une nouvelle représentation des données basée sur une modélisation multi-gaussiennes GMM des locuteurs.

Le reste de cet article est organisé comme suit : dans la section 2, nous présentons une courte description de la technique SVM. La section 3 est consacrée à notre proposition utilisant les SVM en vérification du locuteur ainsi que le protocole expérimental suivi des résultats et la conclusion.

2. MACHINES À VECTEURS DE SUPPORT (SVM)

Les machines à support de vecteurs est une nouvelle technique discriminante dans la théorie de l'apprentissage statistique proposée par V.Vapnick dans son livre « The natural of statistical learning theory » [Vap95] en 1995. Elle permet d'aborder des problèmes très divers comme la classification, la régression, la fusion, etc.

Dans le cadre de ce travail nous nous intéressons aux SVM comme technique de classification.

Le principe de cette technique est de projeter les données de l'espace d'entrée (appartenant à deux classes différentes non-linéairement séparables) dans un espace de plus grande dimension appelé **espace de caractéristiques**. Dans cet espace, on construit un hyperplan optimal séparant les deux classes tel que :

- les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan,
- la plus petite distance entre les points et l'hyperplan (la marge) est maximale.

2.1 Construction de l'hyperplan optimal

Dans ce paragraphe, on présente la méthode générale pour la construction d'un hyperplan optimal qui sépare deux classes. Pour cela on suppose qu'on a une base de données *D* de *m* points d'un espace de dimension p appartenant à deux classes différentes qu'on notera la classe 1 et la classe –1.

$$D = \{ (x_i, y_i) \mid x_i \in \Re^p ; y_i \in \{ 1, -1 \} ; i = 1,\ldots,m \}$$

Cas de données linéairement séparables Dans ce cas, tout hyperplan H : $(w.x)+b$ séparant les deux classes satisfait la condition suivante :

$$y_i (w . x_i + b) \quad \text{pour} \quad i = 1,\ldots, m \qquad (1)$$

Maximiser la marge M (la plus petite distance entre les données des deux classes et l'hyperplan) est équivalent à maximiser la somme des distances des classes par rapport à l'hyperplan. La marge à donc l'expression mathématique suivante :

$$M = \min_{\{x_i | y_i = 1\}} \frac{w . x + b}{\|w\|} \max_{\{x_i | y_i = -1\}} \frac{w . x + b}{\|w\|} = \frac{2}{\|w\|} \qquad (2)$$

Par conséquent, l'hyperplan optimal défini par (w_o , b_o) est celui qui satisfait la condition (1) et qui minimise $\Phi(w)$ définie par :

$$\Phi(w) = \frac{\|w\|^2}{2} \qquad (3)$$

En utilisant les multiplicateur de Lagrange et le théorème de Kuhn-Tucker, le problème se transforme au problème dual suivant :

Maximiser :

$$L(w,b,\alpha)=F(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i.x_j) \qquad (4)$$

sous la contrainte :

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad ; \quad \alpha_i \geq 0$$

Soit la solution $\alpha^o = (\alpha_1^o,..., \alpha_m^o)$. D'après le théorème de Kühn-Tucker [21], une condition nécessaire et suffisante pour que α^o soit optimal est :

$$\alpha_i^o\{y_i[(w_o.x_i)+b_o]-1\}=0 \qquad \text{pour} \qquad i=1, \ldots,m$$

cela veut dire que : $\alpha_i^o = 0$ ou $y_i[(w_o.x_i)+b_o]=1$

Ainsi on définit ***les Vecteurs Supports VS*** par les x_i tels que $y_i[(w_o.x_i)+b_o]=1$, ce qui est équivalent :

$$VS = \{ \ x_i \ \text{tels que} \ \alpha_i > 0 \ \}$$

Ces vecteurs se placent géométriquement comme les plus proches de l'hyperplan optimal qui sépare les deux classes. En reprenant l'équation (9), la normale w_o est calculée par :

$$w_o=\sum_{VS}\alpha_i^o y_i x_i \qquad (5)$$

Le biais b_o est calculé par la formule suivante :

$$b_o = -1/2 \ [(w_o \ . \ x^*(1)) + (w_o \ . \ x^*(-1))]$$

où $x^*(1)$ est un vecteur support de la classe1, et $x^*(-1)$ un vecteur support de la classe –1.

La fonction de classification, *classe(x)*, est égale à :

$$classe\ (x)=sign\ (w_o.x+b_o)$$

$$= sign\left[\sum_{VS}\alpha_i^o y_i (x_i.x)+b_o\right] \qquad (6)$$

si *classe(x)* est inférieur à 0 alors x est de la classe –1 sinon il est de la classe 1.

Cas des données non-linéairement séparables Dans le cas où les données ne sont pas linéairement séparables, l'hyperplan optimal séparant les deux classes est celui qui sépare les données avec le minimum d'erreurs, et donc celui qui satisfait les contraintes suivantes :

- la distance entre les vecteurs bien classés et l'hyperplan doit être maximale,
- la distance entre les vecteurs mal classées et l'hyperplan doit être minimale.

Pour formaliser cela, on introduit des variables de pénalité non-négatives, ε_i pour i = 1, … ,m, appelés variables d'écart. Ces variables transforment l'inégalité (1) comme suit:

$$y_i (w. x_i + b) \geq 1- \varepsilon_i \qquad \text{pour} \qquad i = 1,\ldots, m$$

et on essaye de minimiser la fonction suivante :

$$\Phi(w,\Xi)=\frac{\|w\|^2}{2}+C\sum_{i=1}^{m}\varepsilon_i$$

où C est un paramètre de régularisation. Celui-ci permet de concéder plus ou moins d'importance aux erreurs.

Cela mène à un problème dual légèrement différent de celui du cas séparable. Dans le cas non-linéairement séparables, il faut maximiser $L(\alpha, w, b)$ par rapport à α sous les contraintes suivantes :

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \quad et \quad 0 \leq \alpha_i \leq C \quad \text{pour} \quad i = 1, \ldots , m$$

le calcul de la normale w_o, le biais b_o et la fonction de classification *classe(x)* reste exactement le même que dans le cas linéaire.

2.2 Théorie des SVM

Rappelons que le principe des SVM est de projeter les données de l'espace d'entrée dans un espace de plus grande dimension et de construire un hyperplan optimal séparant les deux classes dans cet espace. L'objectif est de trouver une surface de séparation linéaire dans l'espace de caractéristiques qui corresponde à une surface non-linéaire dans l'espace d'entrée. Considérons l'ensemble de données *D* défini dans la sous-section 2.1, la projection Ψ de $\Re^p$ dans $\Re^q$ tel que q ≥ p et *D*' l'image de *D* dans $\Re^q$.

$$D' = \{ \ (\Psi(x_i), y_i) \mid x_i \in \Re^p ; \ y_i \in \{1, -1\} ; i = 1,\ldots,m \ \}$$

En appliquant la technique de construction de l'hyperplan dans cet espace on aura la fonction de classification suivante :

$$classe\ (x) = sign\left[\sum_{VS}\alpha_i^o y_i[\Psi(x_i).\Psi(x)]+b_o\right]$$

$$= sign\left[\sum_{VS}\alpha_i^o y_i K(x_i,x)+b_o\right] \qquad (7)$$

On note que pour calculer cette fonction de classification on n'utilise que le produit scalaire dans l'espace de caractéristiques qui peut également être exprimé dans l'espace d'entrée par ce qu'on appelle le noyau K(x,y) [Vap95]. Parmi les noyaux les plus utilisés on trouve :

- le noyau linéaire : $K(u, v) = u \ . \ v$,
- le noyau polynomial : $K(u, v) = [(u \ .v)+1]^d$,
- le noyau RBF : $K(u, v) = \exp[-\gamma |u – v|^2]$.

Vous trouverez une description plus détaillée sur les SVM dans les références [Vap95][Bur98].

3. SVM POUR LA VÉRIFICATION DU LOCUTEUR

La plupart des systèmes de vérification du locuteur actuels sont basés sur le principe de test d'hypothèses binaire suivant :

$$Log\left[\frac{P(X/\lambda)}{P(X/\overline{\lambda})}\right]\underset{H_1}{\overset{H_0}{\gtrless}}\beta \qquad (8)$$

où X est le segment de test, $\overline{\lambda}$ représente le modèle indépendant du locuteur, appelé souvent modèle du monde et λ est le modèle de l'identité proclamée. H_0 correspond à l'hypothèse : « X provient de l'identité proclamée λ », H_1 correspond à l'hypothèse : « X ne provient pas de l'identité proclamée λ » et β représente le seuil de décision [Gra00]. Ce seuil peut être dépendant ou indépendant du locuteur.

Dans cette section, nous allons décrire notre approche qui consiste à utiliser la technique SVM en vérification du locuteur en mode indépendant du texte. L'idée principale de notre proposition est d'utiliser les SVM dans la phase de décision en construisant un modèle qui permet de discriminer la classe des *accès clients* de la classe des *accès imposteurs*. Ce modèle est obtenu grâce à la nouvelle représentation des données que nous proposons et qui permet de construire les vecteurs d'entrée des SVM en se basant sur les modèles GMM des clients et du monde obtenus en phase de modélisation.

3.1 Construction des vecteurs d'entrée pour le modèle SVM

Dans notre application, les SVM sont utilisés en phase de décision. Pour cela on a besoin d'un segment de test X, du modèle de l'identité proclamée λ et du modèle du monde $\overline{\lambda}$. Supposons que le monde et les clients soient modélisés par des GMM de n gaussiennes alors le vecteurs d'entrée est également de dimension n. La construction de ces vecteurs est réalisé de la manière suivante : D'abord, on initialise les composantes du vecteur d'entrée par zéro. Ensuite pour chaque trame t du segment de test X, on détermine l'indice de la gaussienne g_i qui maximise la probabilité $P(t / g_j)$ avec $g_j \in \lambda, \overline{\lambda}$. Un score S_t est ainsi calculé en utilisant l'équation suivante :

$$S_t = Log\left[\frac{P(t/g_i^{\lambda})}{P(t/g_i^{\overline{\lambda}})}\right] \qquad (9)$$

Enfin, le score S_t est ajouté à la $i^{ème}$ composante du vecteur V_{λ}^{X}, vecteur d'entrée pour le module de décision basé sur les SVM. La Figure 1 représente un schéma modulaire de la construction des vecteurs d'entrée pour notre système SVM. Après avoir traité tout le segment de test, le vecteur obtenu est alors normalisé par le nombre de trames du segment.

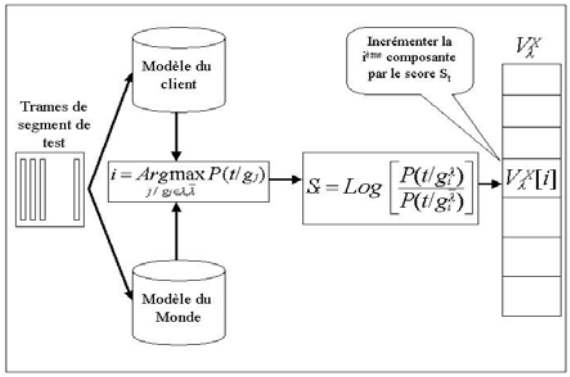

Figure 1 : Construction des vecteurs d'entrée des SVM pour notre système hybride GMM-SVM

3.2 Protocole expérimental

Base de données Nous utilisons dans nos expériences la base de données NIST'2001 qui est une partie de la base SWITCHBOARD. Cette base est constituée de 1003 locuteurs dont 546 femmes et 557 hommes. Chaque locuteur a enregistré chacun une session de 2 minutes de parole et quelques dizaines de segments de test d'une durée de plus de 3 secondes. Deux types de combinés téléphoniques ont été utilisés pour les enregistrements de cette base, *électret* et *carbone*. Pour le développement de notre système (construction des modèles du monde et les modèles SVM) nous avons utilisé un sous ensemble de la base de données NIST1999.

Paramétrisation Pour la phase de paramétrisation, le module standard du Consortium ELISA [Mar01] est utilisé. Le signal de parole est représenté toutes le 10ms par des trames de 32 composantes calculées sur une fenêtre de 20ms, 16 coefficients cepstraux et 16 delta des coefficients calculés sur 5 trames. Une normalisation basée sur le retrait de la moyenne cepstrale (Cepstral Mean Subtraction) permet de minimiser les perturbations dues aux différents canaux de transmission de la voix. Enfin un algorithme de suppression des trames, basé sur une modélisation bi-gaussienne de l'énergie, est utilisé pour supprimer toutes les trames qui appartiennent à la gaussienne de plus faible moyenne [Mar01].

Modélisation La phase de modélisation est réalisée grâce au module de modélisation de la plate-forme ELISA. Les modèles du monde et des clients sont des GMMs de 128 gaussiennes avec une matrice de covariance diagonale. Quatre modèles indépendants du locuteur, dépendants du sexe et du combiné téléphonique ont été construits utilisant environ 50 locuteurs pour chaque modèle. Les modèles des clients sont obtenus par adaptation du modèle du monde de même sexe et de même combinés téléphoniques en utilisant la technique MAP (Maximum a Posteriori). Deux modèles SVM dépendent du sexe ont été construits sur l'ensemble de développement utilisant 100 locuteurs dont 50 femmes et 50 hommes. On a réalisé 272 accès clients et 2720 accès imposteurs pour le modèle des accès femmes et 247 accès clients contre 2470 accès imposteurs pour construire le modèle des accès hommes.

Les accès clients et imposteurs sont respectivement étiquetés +1 et –1.

Décision Pour la phase de décision, la technique SVM est utilisée comme expliqué dans la section précédente. Deux modèles SVM dépendant du sexe ont été construits sur l'ensemble de développement. Sur l'ensemble d'évaluation, on a réalisé 94517 tests dont 48655 accès femmes et 45862 accès hommes.

Nous avons utilisé le logiciel SvmFu, disponible gratuitement sur le site http://svm.first.gmd.de

Pour valider nos systèmes, nous avons comparé avec un système de référence utilisant la technique classique du Log du Rapport de Vraisemblance LLR. La technique LLR est considérée comme étant la plus performante et la plus utilisée dans le domaine de la vérification du locuteur**.**

4. RESULTATS

La Figure 2 montre les résultats de notre meilleur système GMM-SVM en utilisant le noyau RBF avec ($\gamma = 50$) comparé au système de référence LLR sous forme de courbes DET [Mar97] et sur deux conditions « Primary : on ne prend en compte que les décisions prises sur des segments de test d'une durée de 15s à 45s et ou le combiné utilisé est électret » « All : prend en compte les décision prises sur tous les tests »

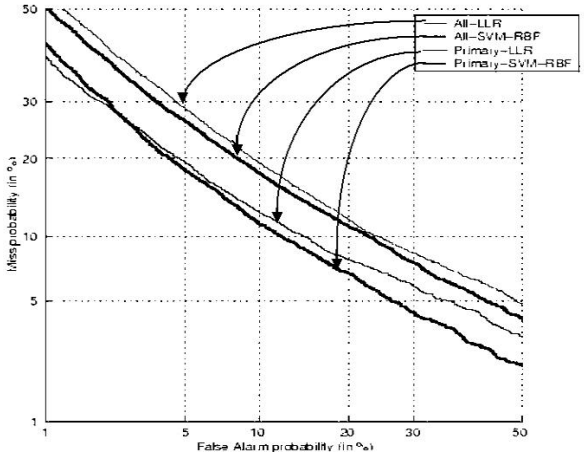

Figure2 : Courbes DET du meilleur système GMM-SVM utilisant le noyau RBF avec $\gamma = 50$ comparé au système de référence LLR sur les deux conditions « Primary » et « All » sans aucune normalisation.

Tableau 1 : Résultas obtenus sur les données de l'évaluation NIST2001

	Primary	All
Sys de réf (LLR)	11.5% [11.4 , 11.6]	16% [15.9 , 16.1]
SVM (RBF $\gamma = 50$)	**10.75% [10.6, 10.9]**	**15% [14.9,15.2]**
SVM (RBF $\gamma = 40$)	11.25% [11.1 , 11.3]	16% [15.9 , 16.1]
SVM (Linéaire)	11.5% [11.4 , 11.6]	16.25%[16.1,16.4]

Les résultats montrent que notre système hybride est plus performant que le système de référence. Les Taux d'Egale Erreur (TEE) obtenus par ces deux systèmes avec des intervalles de confiance de 95% sont présentés dans le tableau1 ainsi que les résultats de deux autres systèmes hybrides GMM-SVM avec différents noyaux (linéaire et RBF avec $\gamma = 40$).

5. CONCLUSION

Dans cet article, nous avons présenté un nouveau système hybride GMM-SVM pour la vérification du locuteur. Dans ce système nous avons proposé une nouvelle représentation de données réunissant la capacité de modélisation des GMM et l'efficacité de décision des SVM. Ce travail représente la suite d'un premier système que nous avons présenté dans [kha01]. Les résultats montrent que les performances de notre système hybride GMM-SVM au noyau RBF avec $\gamma = 50$ sont meilleurs que celles du système de référence basé sur la technique classique LLR sans aucun recouvrement entre les intervalles de confiance. Notons que les résultats présentés dans cet article sont obtenus sans aucune normalisation comme Znorm Hnorm et Tnorm [Gra00].

BIBLIOGRAPHIE

[Vap95] Vapnik V. (1995), " The Natural of Statistical Learning Theory" , Spring-Verlag

[Sch 96] Schmidt M. et H. Gish H. (1996) " Speaker identification via Support Vector Classifier, Proc ICASSP'96, pp. 105-108

[Mar97] Martin A. and al (1997), " The DET curves in assessment of detection task performance " EUROSPEECH, Vol 4, pp 1895-1898.

[Osu97] Osuna E., Freund R. and Girosi F. (1997) " Training Support Vector Machines: Application to Face Detection " Proceedings CVPR.

[Ben98] Ben-yacoub S. (1998) " Multi-Modal Data for person Authentication using SVM " IDIAP-PR-07-98.

[Bur98] Burges C. (1998), " A Tutorial on Support Vector Machines for Pattern recognition " Data Mining and Knowledge discovery, 2 (2).

[Fer99] Fernandez R., Viennet E. (1999), " face identification with support Vector Machine ", Proceedings ESCANN.

[Gra00] Gravier G., Kharroubi J., Chollet G. (2000), " On the use of Prior knowledge in normalization scheme for speaker verification ", DSPJ, Vol. 10 N 1-3, pp 213-225.

[Kha01] Kharroubi J., Petrovska D., Chollet G. (2001), " Combining GMM's with Support vector Machines for Speaker verification ", Eurospeech, pp. 1761-1764

[Mar01] Magrin-Chagnolleau I., Gravier G., Blouet R. (2001), " Overview of the Elisa Consortium Research Activities ", Odyssey.