SETIT 2007
4th International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 25-29, 2007 – TUNISIA

Authentification Discriminative du Locuteur Basée sur une Fusion Statistique - Connexionniste.

S. Ouamour*, M. Guerti#, H. Sayoud

*USTHB, FEI, BP 32 Bab Ezzouar, Alger. # E.N.P. 10, Avenue Hacen Badi, El Harrach, Alger

Emails: ouamour@ifrance.com mhaniag@yahoo.fr sayoud@ifrance.com

Abstract : In this paper, we describe a new approach for the task of speaker discrimination, using a fusion between two classifiers: statistical and neural classifiers. This fusion is obtained once by combining the scores of the two previous classifiers weighted by some confidence coefficients and another time by mixing them (hybrid model). The mixing is obtained by using the results of the statistical classifier at the input of the NN to improve its training.
In order to evaluate our fusion approach, we made a series of experiments on two databases: - Broadcast News 1996 corpus and - Real Telephonic recordings.
Experiments showed that the fusion improved the scores obtained by each approach alone. For instance, results of this fusion in speaker discrimination gave an EER of 7.88% on *Broadcast News* (speech segments duration of 4 seconds) and an EER of 4.29% on the telephonic database (with segments of 10 seconds).

Key words: Speaker verification, Speech Processing.

INTRODUCTION

L'authentification du locuteur par la voix représente un domaine important de la biométrie, vu que la parole reste le seul moyen utilisé à distance via canal téléphonique. Il consiste à identifier ou à vérifier l'identité proclamée par un locuteur uniquement par son empreinte vocale à partir de son signal de parole. L'authentification discriminative consiste à comparer les caractéristiques vocales extraites de deux ou plusieurs segments de parole, dans le but de décider si ces derniers ont été prononcés par un même locuteur ou par des locuteurs différents.

L'authentification discriminative par locuteur (incluant la vérification du locuteur, la segmentation par locuteur, la détection de locuteurs, …) possède plusieurs applications pratiques, parmi elles, on peut citer le contrôle d'accès aux zones sécurisées, la validation des transactions bancaires par téléphone, la segmentation par locuteur, l'indexation des documents audio, etc. Cette variété d'applications dont certaines nécessitent un système d'authentification de haute sécurité, nous a incités à développer une nouvelle approche basée sur la fusion de plusieurs classifieurs afin d'améliorer les résultats obtenus par chaque classifieur seul. Dans ce travail, nous avons opté pour la fusion des deux classifieurs suivants:

- un classifieur statistique basé sur un modèle mono gaussien et représenté par la mesure symétrique gaussienne à covariance ;
- et un classifieur connexionniste basé sur un réseau de neurones du type MLP.

Le choix de ces deux classifieurs résulte du fait que la mesure statistique gaussienne est une mesure simple à implémenter, rapide (non coûteuse en temps de calcul) et donne des résultats satisfaisants ; tandis que le classifieur connexionniste est réputé d'avoir une excellente capacité discriminative cependant il nécessite un très grand nombre d'exemples pour le training. Dans l'objectif d'accroître / renforcer d'avantage la qualité du système d'authentification, nous avons opté pour la fusion des résultats de ces deux classifieurs. La fusion est obtenue en additionnant les scores pondérés des deux

classifieurs et une seconde fois, en utilisant la sortie du classifieur statistique comme entrée au classifieur neuronal (méthode hybride).
Pour la phase d'évaluation, deux bases de données contenant des enregistrements de journaux télédiffusés [1] et des enregistrements téléphoniques ont fait l'objet de nos expériences.

Les résultats obtenus sont encourageants et montrent un bon comportement de la fusion par rapport à chaque classifieur pris seul.

1. Principales techniques d'authentification du locuteur

Plusieurs techniques ont été développées pour la tâche d'authentification, comme les GMM (*Gaussian Mixture Models*) [2], les réseaux de neurones [3], les mesures statistiques [4], les HMM (*Hidden Markov Models*) [5] etc.

Dans ce travail de recherche, nous nous intéressons à quatre approches différentes :

- un réseau de neurones du type MLP ;
- une mesure statistique gaussienne ;
- une méthode hybride, combinant les deux méthodes précédentes ;
- une fusion résultant de la sommation pondérée des scores des deux classifieurs (statistique et MLP).

Concernant la paramétrisation, nous avons utilisé 37 coefficients MFSC (*Mel Frequency Spectral Coefficients*) [6, 7]. Cette dimension a été choisie après une étude faite sur la résolution spectrale optimale [8].

1.1. Méthode Statistique

Dans la méthode statistique on exploite le fait que les caractéristiques spectrales de tout locuteur, pour une phrase longue, suivent une loi gaussienne stationnaire du second ordre [9, 10]. Pour ce faire, les étapes suivies sont :

D'abord, on procède à l'extraction des MFSC ou Mel-énergies [5], puis pour chaque prononciation on calcul le vecteur moyenne x et la matrice de covariance X ; ainsi il existe une moyenne x et une covariance X pour chaque locuteur. Le couple (x, X) représente la référence statistique d'ordre 2 pour le locuteur ***X*** utilisé dans le dictionnaire des références.

En phase de test, une modélisation similaire de la phrase de test génèrera le couple (y, Y) représentant le modèle statistique de test pour le locuteur inconnu ***Y***.
Le test d'identification est basé, alors, sur la distance minimale (le plus proche voisin) au sens de la métrique statistique du 2e ordre: $\mu_{Gc0.5}$

$$\mu_{Gc\,0.5}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\alpha - 1 \tag{1}$$

$$\alpha = \frac{1}{p}\left(tr\left(YX^{-1}\right) + tr\left(XY^{-1}\right)\right) \tag{2}$$

p étant la dimension du vecteur des caractéristiques acoustiques et "tr" dénote la trace d'une matrice.
X représente la référence et ***Y*** représente le locuteur à reconnaître.

1.2. Méthode Connexionniste

Connaissant les hautes performances des réseaux de neurones dans la discrimination [3], nous avons opté pour l'utilisation d'un réseau de type MLP avec une ou deux couches cachées et une sortie [11] (figure 1). Ce réseau de neurones permet d'authentifier le locuteur des coefficients MFSC extraits à partir de son signal de parole.

Le réseau de neurones doit avoir à son entrée un nombre de neurones égal à la dimension du vecteur exemple [7]. Par conséquent, dans le cas où la dimension du vecteur représentant un segment est de N coefficients [6, 7], le nombre de neurones de la couche d'entrée est égal à 2N (correspondant aux deux segments de parole à comparer). L'apprentissage du NN est assuré par l'algorithme de rétropropagation du gradient.

Le réseau de neurones donnera à sa sortie NN_{sortie} une indication concernant la similarité des deux segments de parole présentés à l'entrée de ce dernier :

- si $NN_{sortie} = 0$ alors il s'agit du même locuteur ;
- si $NN_{sortie} = 1$ alors il s'agit de locuteurs différents.

Concernant l'analyse spectrale du signal, nous avons utilisé une segmentation par fenêtres de 32 ms (assurant la stationnarité), où cette analyse donne une série de vecteurs MFSC pour chaque segment [6, 7].

A partir de ces vecteurs, on extrait les caractéristiques réduites issues des matrices de covariance, pour chaque couple de segments à comparer, qui seront injectées ensuite à l'entrée du NN. Ce dernier décidera à sa sortie si ces segments appartiennent à un même locuteur ou non (fig. 1).

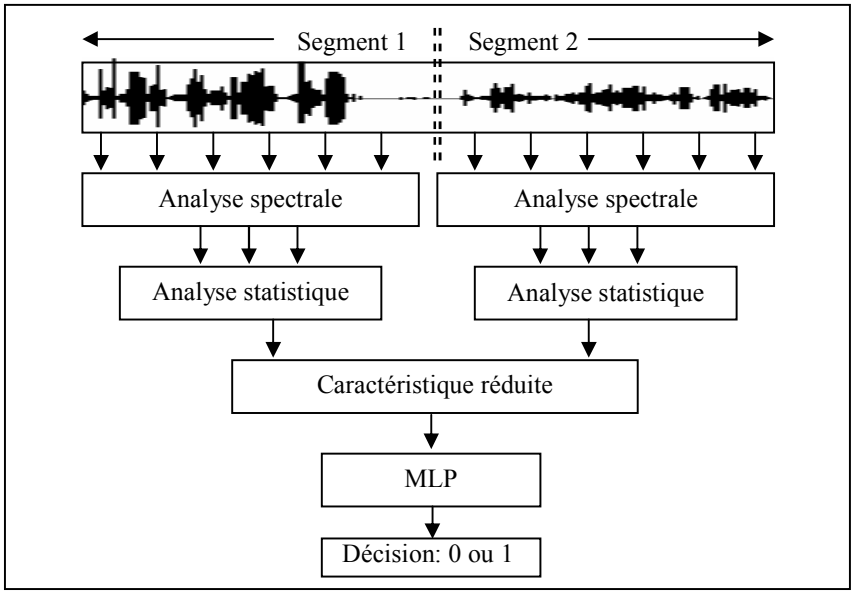

Figure 1 : Comparaison entre des segments et décision du NN.

1.3. Méthode Hybride

La méthode hybride consiste à mixer les deux classifieurs :

- statistique présenté par une mesure mono gaussienne à covariance. Celle-ci a montré un bon comportement dans les différentes tâches d'authentification [9] ;
- et neuronal (MLP). Vu qu'il présente de hautes capacités discriminatives [3].

Ce mixage est obtenu en exploitant les résultats du classifieur statistique dans l'amélioration de l'apprentissage du réseau de neurones.
Pour ce faire, nous avons ajouté une nouvelle entrée au réseau de neurones, où sont injectées les résultats de l'authentification de la mesure statistique (μ_{Gc}) pour chaque couple de segments. Le nouvel apprentissage est assuré avec cette nouvelle entrée rajoutée aux entrées présentant les caractéristiques réduites des covariances du couple de segments correspondant à cette entrée statistique

1.4. Méthode de Fusion

Nous proposons de fusionner les scores des différents classifieurs afin d'améliorer les résultats par rapport à ceux obtenus par chaque méthode seule [12].
En effet, plusieurs techniques de fusion sont proposées dans la littérature, nous pouvons citer la fusion basée sur la somme, le produit, le calcul du maximum, le calcul du minimum, le vote, le LDA et la fusion par réseau de neurones [12].

Dans notre cas, nous avons choisi une technique de fusion permettant de sommer les différents scores obtenus par chaque méthode. Ces scores sont pondérés par des coefficients, spécifiant ainsi le poids de chaque méthode dans la fusion.
Si les scores simples (obtenus par chaque classifieur seul) sont notés par S_j, alors le score de la fusion noté *Sf* est défini par l'équation 4 :

$$Sf = \sum_{j=1}^{N} C_j S_j \qquad (4)$$

où :
C_j représente le coefficient de pondération pour le classifieur "j" et N représente le nombre de classifieurs.

avec $\sum_j C_j = 1$ et $Cj \in]0,\ 1[$ (5)

Le *Cj* représente le degré de pertinence d'un classifieur j par rapport aux autres dans la fusion.

2. Résultats et Discussion

Les bases de données audio utilisées dans nos expériences comprennent des enregistrements de journaux télédiffusés extrait de la base "*Broadcast News-1996*" et téléphoniques réels. Dans les deux bases de données les exemples utilisés pour la phase d'entraînement sont différents par rapport à ceux utilisés durant la phase de test. La durée des segments traités est d'environ 4 s dans le cas de la base DB2 (*Broadcast News*) et d'environ 10 s dans le cas de DB3 (téléphonique).

Dans le but d'évaluer les différentes techniques définies précédemment, nous avons effectué plusieurs expériences de discrimination dans les bases DB2 et DB3. Chaque expérience concerne une méthode particulière (figures 2 et 3). Ces deux figures représentent les courbes ROC (taux de Fausses Alarmes (FA) en fonction des Détections Manquées (MD)).
A partir de ces figures, nous pouvons remarquer que :

- la comparaison entre les 2 classifieurs simples : statistique et neuronal, montre que le EER obtenu par le MLP (9.25% dans le cas de DB2 et 5.02% dans le cas de DB3) est meilleur que son correspondant obtenu par la mesure statistique (11.75% dans le cas de DB2 et 5.74% dans le cas de DB3) (table 1) ;
- dans la zone centrale de la courbe ROC, Le NN-MLP est meilleur que la μ_{Gc}, par contre extrémités, c'est l'autre qui est meilleure (figure 2) ;
- la méthode hybride a amélioré le résultat de la μ_{Gc} sur DB2 et a permis l'amélioration du EER par rapport aux deux classifieurs simples sur DB3.
- la deuxième méthode de fusion basée sur la sommation pondérée des scores des classifieurs simples donne le meilleur EER dans DB2 et DB3 ;
- la comparaison des 2 techniques de fusion : hybride et sommation pondérée montre que la sommation pondérée est plus efficace que la méthode hybride sur les deux bases traitées (table 1).

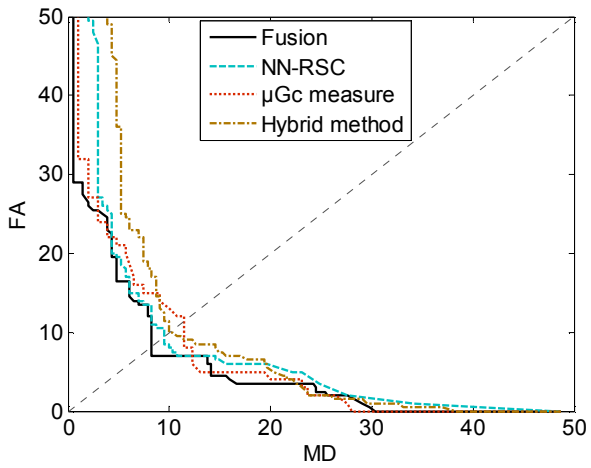

Figure 2 : Courbe ROC de discrimination de locuteur sur *Broadcast news*.

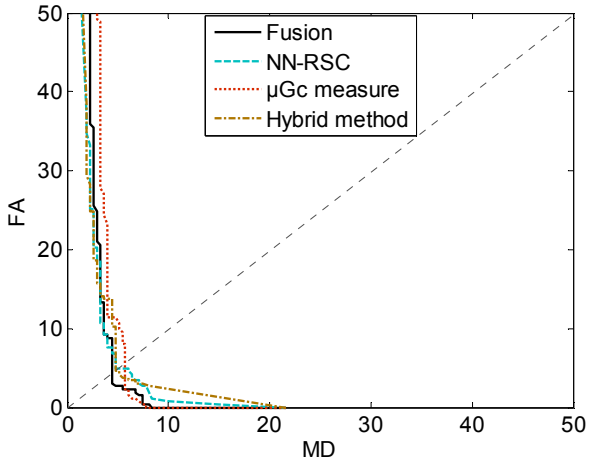

Figure 3 : Courbe ROC de discrimination de locuteur sur parole téléphonique.

Table 1 : EERs obtenus pour les différentes méthodes.

Classifieur / Méthode	EER % sur :	
	Broadcast News	Téléphonie
Mesure Statistique	11.75	5.74
NN-RSC	9.25	5.02
Méthode Hybride	9.95	4.65
Fusion: NN-Statistique	7.88	4.29

3. Conclusion

L'authentification discriminative de locuteurs consiste à comparer les caractéristiques vocales extraites de deux ou plusieurs segments de parole, afin de décider si ces segments ont été prononcés par un même locuteur ou par des locuteurs différents. Dans le but d'aborder ce problème, nous avons testé plusieurs techniques notamment, un MLP, une mesure statistique et des techniques de fusion : hybride et sommation pondérée. Sur les deux bases de données testées : *Broadcast-News* et téléphonique, les différents résultats ont été clairement présentés par les courbes ROC.
Les résultats nous ont permis de comparer ces 4 méthodes selon leurs EER ; Concernant les classifieurs simples, nous avons constaté que le EER du MLP est meilleur que celui de la mesure statistique dans les deux bases : DB2 et DB3, ce qui confirme les bonnes performances des réseaux de neurone [3].

La mixture entre ces deux classifieurs simples, pour obtenir le modèle hybride a amélioré d'avantage les résultats de l'authentification discriminative puisque ce dernier a réduit le EER jusqu'à 4.65% dans la base téléphonique.

Par ailleurs, la fusion par sommation pondérée des scores a donné les meilleurs taux de discrimination : un EER de 7.88% sur DB2 et un EER de 4.29% sur DB3 ce qui montre l'excellente discrimination de cette dernière. En conclusion, cette étude confirme, encore, le bon comportement des réseaux de neurone en discrimination et montre que les techniques de fusion, utilisées ici, sont efficaces en discrimination de locuteurs, surtout pour ce qui est de la fusion par sommation pondérée des scores.

REMERCIEMENTS

Nous remercions les équipes de recherche du laboratoire LIA d'Avignon pour leurs aides quant à l'achèvement de ce travail.

BIBLIOGRAPHIE

[1] P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young The Development of the 96 HTK broadcast news transcription system. In *DARPA Speech Recognition Workshop*, pages 97-99, 1997.

[2] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, volume 17 number. 1-2, pages 91-108,1995**.**

[3] Y. Bennani, and P. Gallinari. Neural Networks for discrimination and modelization of speakers. *Speech Communication,* volume 17, number 1-2, pages 159-175, 1995.

[4] H. Sayoud, and S. Ouamour. Discrimination Parole/ Non Parole en suivi de locuteur, *RJC'01*, Mons Belgium, pages 130-132, 8-11 September 2001.

[5] S. Meignier. Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection. PhD thesis, LIA Avignon, France, 2002.

[6] H.S. LEE, and A.C. TSOI. Application of multi-layer perceptron in estimating speech / noise characteristics for speech recognition in noisy environment. *Speech Communication*, volume 17, number 1-2, pages 59-76 1995.

[7] H. Sayoud, and S. Ouamour. 'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation. *Acta Acustica*, volume 89, number 4, pages 702-710, 2003.

[8] H. Sayoud, and S. Ouamour. Reconnaissance Auto du Locuteur en Milieu Bruité. *JEP'00*, Aussois, France, pages 345-348, June 2000.

[9] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan. Second-Order Statistical Measures for text-independent Broadcaster Identification. *Speech Communication*, volume 17, number 1-2, pages 177-192, August 1995.

[10] F. Bonastre, and L. Besacier. Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur, Actes du 4ème Congrès Français d'Acoust., Marseille 14-18 Apr. 97, pages 357-360, 1997.

[11] H. Sayoud. Reconnaissance Automatique du locuteur – Approche Connexionniste-, PhD thesis, USTHB University, Algiers, 2003.

[12] J. Kitler. Multiple classifier systems in decision-level fusion of multimodal biometric experts. *1st BioSecure residential workshop*, Paris, France, 1- 26 August 2005.