République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Supérieure Polytechnique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Supérieure Polytechnique

Département d'Electronique

Laboratoire Signal & Communications

THESE DE DOCTORAT

PRESENTEE PAR :

Mme Siham OUAMOUR épouse SAYOUD

Thème

INDEXATION AUTOMATIQUE DES DOCUMENTS AUDIO EN VUE D'UNE CLASSIFICATION PAR LOCUTEURS - APPLICATION A L'ARCHIVAGE DES EMISSIONS TV ET RADIO -

Devant le Jury :

L. HAMAMI	**Professeur**	**ENSP**	**Présidente**
M. GUERTI	**Professeur**	**ENSP**	**Rapporteur**
A. GUESSOUM	**Professeur**	**USD-Blida**	**Examinateurs**
M. TALHA	**Maître de Conférences**	**USTHB**	
H. TEFFAHI	**Maître de Conférences**	**USTHB**	

Année 2009

بسم الله

REMERCIEMENTS

Après avoir remercié notre 'Créateur'

Un très grand Merci à Professeur Mhania GUERTI : ma directrice de thèse, enseignante et directrice de recherche à l'ENSP. Je tiens à la remercier vivement autant pour la qualité de son encadrement, que pour sa disponibilité ainsi que pour ses qualités humaines. Je la remercie pour les nombreuses lectures et pour ses remarques constructives. Je la remercie également, pour m'avoir facilité la tâche par sa gentillesse et son expérience. Sa longue expérience dans le domaine de la parole lui confère une véritable qualité de jugement scientifique. Je tiens aussi à la remercier, pour la pleine confiance qu'elle m'a accordée au cours de ces cinq années.

Mes vifs remerciements vont aussi à Mme Latifa HAMAMI, Professeur et chercheur à l'Ecole Polytechnique Supérieure d'Alger ENSP, pour l'intérêt qu'elle porte à cette thèse et pour avoir accepté très aimablement de présider mon jury. Mme HAMAMI a une longue expérience dans le domaine du traitement du signal et des images.

Je tiens à remercier vivement Mr Abderrezak GUESSOUM, Professeur et chercheur à la faculté d'Electronique de l'université Saâd Dahleb-Blida, pour avoir accepté de participer au jury de ma soutenance. Mr GUESSOUM a exprimé une disponibilité et un intérêt immédiats quant au jugement de ce travail de recherche.

Nous sommes très heureux d'avoir dans le jury, un spécialiste de la parole Mr Hocine TEFFAHI, membre du laboratoire CPTS et également Maître de conférences à la faculté d'Electronique et d'Informatique de l'USTHB. Mr Hocine TEFFAHI a été l'un des premiers membres travaillant dans le domaine de la « Communication Parlée » à l'USTHB. Il a effectué plusieurs travaux intéressants sur la parole et nous aimons toujours connaître son avis quant aux résultats obtenus. Nous donnons beaucoup d'importance aux remarques qu'il apporte.

Je remercie très chaleureusement, Mme Malika TALHA, Maître de conférences et chercheur à la faculté d'Electronique et d'Informatique de l'USTHB, d'avoir bien voulu juger, objectivement, ce travail de recherche et me faire l'honneur de participer au jury. Je la remercie particulièrement pour toutes les aides offertes ainsi que ses précieux conseils durant ma formation, alors qu'elle nous enseignait le traitement du signal.

Je n'oublie pas de remercier aussi Dr J.F. BONASTRE du LIA d'Avignon, qui nous a aidé durant la préparation de la thèse.

Je remercie enfin toute ma famille, et particulièrement mon mari, Monsieur Halim SAYOUD, Maître de conférences à l'USTHB, pour son soutien au cours du parcours aboutissant à cette thèse.

Mes remerciements vont également à ceux qui m'ont aidé de près ou de loin.

و الحمد لله رب العالمين

الملخص

هذه المذكرة تهتم بدراسة فهرسة التسجيلات الإذاعية والتلفزيونية حسب فئات المتكلمين، بهدف الحصول على أرشيف مرتب حسب مختلف المتكلمين.
مهمة الفهرسة تستدعي إختصاصين مختلفين، الأول يهتم بتجزئة التسجيل الصوتي إلى أجزاء متجانسة : إنها تجزئة الكلام، بينما يهتم الثاني بتعريف مختلف الأجزاء أو تجميعها حسب فئات المتكلمين : تسمى هذه العملية بوضع اللصائق أوبالتجميع.

للوصول إلى هدفنا، إقترحنا وأنجزنا نظامين أوتوماتيكيين :

- النظام الأول يهتم بالفهرسة مع معرفة أولية للمتكلمين حيث أن شخصيات مختلف المتكلمين معروفة مسبقا من النظام ؛
- النظام الثاني يعالج مهمة الفهرسة بدون أي معرفة مسبقة لنماذج المتكلمين.

لإنجاز النظام الأول، قمنا بإنشاء خوارزمية جديدة للفهرسة مبنية على الفهرسة المضفورة للكلام وتستعمل القياسات الإحصائية من الدرجة الثانية. بينما لإنجاز النظام الثاني، قمنا بإنشاء خاصية جديدة نسبية للمتكلم والتي سميناها الخاصية النسبية للمتكلم.
قمنا بإنشاء ثلاث مصنفات مختلفة : مصنف إحصائي و شبكة عصبية ومصنف SVM، اقترحنا عدة طرق لإدماج المصنفات.

اختبار الانظمة تم على قاعدة بيانات مكونة من اشارات صوتية حقيقية منتقاة من أخبار تلفزيونية. النتائج اظهرت الاداء الجيد للخوارزمية المنشأة، جودة الخاصية النسبية الجديدة للمتكلم و أهمية الإدماج في تحسين دقة التجزئة و الفهرسة.

كلمات المفاتيح :
الفهرسة الصوتية، تجزئة الكلام، إدماج المصنفات، المصنفات الإحصائية، الشبكات العصبية، SVM.

Résumé

Ce travail de thèse s'intéresse à l'indexation des émissions radio et télé-diffusées en classes de locuteurs, dans le but d'obtenir un archivage hiérarchique des interventions audio en fonction des différents locuteurs.
La tâche d'indexation fait appel à deux disciplines différentes, la première s'intéresse à découper le flux audio en segments homogènes : c'est la segmentation, tandis que la deuxième tâche consiste à identifier les différents segments ou bien les regrouper en classes de locuteurs : c'est l'étiquetage ou le regroupement.
Pour arriver à cette fin, nous avons proposé et implémenté deux systèmes :

- le premier s'intéresse à l'indexation avec connaissances a priori des locuteurs où les identités des différents locuteurs, sont connues à l'avance par le système ;
- le deuxième traite la tâche d'indexation sans aucune connaissance des modèles des locuteurs.

Pour réaliser le premier système, nous avons développé un nouvel algorithme d'indexation que nous avons appelé **ISI** (***I**nterlaced **S**peech **I**ndexing*). Ce dernier est basé sur une indexation entrelacée en utilisant les mesures SOSM (Mesures Statistiques du Second Ordre). Pour le second système, nous avons développé une nouvelle caractéristique relative du locuteur que nous avons appelée **RSC** (***R**elative **S**peaker **C**haracteristic*). Nous avons implémenté trois classifieurs différents : un classifieur statistique, un réseau de neurones du type MLP (*Multi-Layer Perceptron*) et un classifieur SVM (*Support Vector Machines*). Par la suite, nous avons proposé plusieurs architectures afin de fusionner ces classifieurs.
L'évaluation de nos systèmes a été faite sur une base de données de parole réelle : HUB-4 Broadcast News. Les résultats obtenus ont montré la bonne performance de l'algorithme ISI, la pertinence de la nouvelle caractéristique RSC, ainsi que l'intérêt de la fusion quant à l'amélioration de la précision de segmentation et d'indexation.

Mots clés :
Indexation Audio, Segmentation de la parole, Fusion des classifieurs, Classifieurs statistiques, Réseaux de Neurones, SVM.

Abstract

This thesis research work deals with the indexing of broadcast news documents into speakers classes, in order to get a hierarchic archiving of the audio speeches according to the different speakers.
The indexing task is based on two different operations, the first one splits the audio stream into homogeneous segments : this is the segmentation, whereas the second one consists in identifying the different segments or regrouping them into speakers classes : this is the labelling or the clustering.
For that purpose, we have proposed and implemented two automatic systems :

- the first system deals with the task of indexing with knowledge of the speakers where the different speakers are known in advance by the system ;
- the second system deals with the task of indexing without any knowledge about the speakers.

For the realisation of the first system, we have developed a new indexing algorithm which we called **ISI** (**I**nterlaced **S**peech **I**ndexing), based on an interlaced indexing and using the SOSM measures (Second Order Statistical Measures). For the realisation of the second system, we have developed a new Relative Characteristic for the Speaker which we called **RSC** (**R**elative **S**peaker **C**haracteristic). We have implemented three different classifiers: a statistic classifier, a neural network MLP (Multi Layer Perceptron) and a SVM (Support Vectror Machines) classifier. Thereafter, we have proposed several architectures to fuse these classifiers.
The evaluation of our systems was done on real speech database: HUB-4 Broadcast News. Results show the good performance of the ISI algorithm, the pertinence of the new RSC characteristic, and the interest of the fusion for the enhancement of the indexing accuracy.

Keywords :
Audio Indexing, Speech segmentation, Fusion of classifiers, Statistical classifiers, Neural Networks, SVM.

LISTE DES ABRÉVIATIONS

ABC :	**A**merican **B**roadcasting **C**ompany
BBC :	**B**ritish **B**roadcasting **C**orporation
BIC :	**B**ayesian **I**nformation **C**riterion
BP :	**B**ande **P**assante
CMS :	**C**epstral **M**ean **S**ubtraction
CNN :	**C**able **N**ews **N**etwork
Cf :	**C**on**f**usion
CW :	**C**oefficient de **W**olf
BD :	**B**ase de **D**onnées
DB1 :	**D**ata **B**ase 1
Det :	**D**éterminant
DISTBIC :	Speaker based Segmentation for Audio Data Indexing.
DRSC :	**D**iagonal of the **R**elative **S**peaker **C**haracteristic
DSD :	**D**ivergence **S**hape **D**istance
DVS :	**D**écomposée en **V**aleurs **S**ingulières
EER :	**E**qual **E**rror **R**ate
Exp :	**Exp**érience
FA :	**F**alse **Al**arms
Fe :	**F**réquence d'**é**chantillonnage
FFT :	**F**ast **F**ourier **T**ransform
GLR :	**G**eneralized **L**ikelihood **R**atio
GMM :	**G**aussian **M**ixture **M**odels
GSM :	**G**lobal **S**ystem for **M**obile **C**ommunications
HUB-4 :	base de données parlée Américaine enregistrée à partir des chaînes de télévision et de radio.
IAL :	**I**dentification **A**utomatique du **L**ocuteur
ISI :	**I**nterlaced **S**peech **I**ndexing
KL :	**K**ullback-**L**eibler
LFCC :	**L**inear **F**requency **C**epstral **C**oefficient
LPC :	**L**inear **P**redictive **C**oding
LSP :	**L**ine **S**pectral **P**airs
MD :	**M**issed **D**etections
MFCC :	**M**el **F**requency **C**epstral **C**oefficient
MFSC :	**M**el **F**requency **S**pectral **C**oefficient
MLP :	**M**ulti **L**ayer **P**erceptron
MP3 :	**M**otion **P**icture Experts Group Audio Layer **3**
NHK :	**N**ippon **H**ôsô **K**yôkai
NPR / PRI :	**N**ational **P**ublic **R**adio / **P**ublic **R**adio **I**nternational

PLP : **P**erceptual **L**inear **P**redictive
RAL : **R**econnaissance **A**utomatique du **L**ocuteur
RAP : **R**econnaissance **A**utomatique de la **P**arole
RBF : **R**adial **B**asis **F**unction
RN : **R**éseaux de **N**euronres
RNA : **R**éseaux de **N**euronres **A**rtificiels
ROC : **R**eceiver **O**perating **C**haracteristic
RSB : **R**apport **S**ignal sur **B**ruit
RSC : **R**elative **S**peaker **C**haracteristic
SAD : **S**ilence **A**ctivity **D**etection
SEP : **S**ous-**E**space **P**ropre
SOSM : **S**econd **O**rder **S**tatistical **M**easures
SVM : **S**upport **V**ector **M**achines
TB1 : **T**elephonic Data **B**ase 1
TPZ : **T**aux de **P**assages par **Z**éro
VAL : **V**érification **A**utomatique du **L**ocuteur
VoIP : **V**oice **o**ver **IP**
VQ : **V**ector **Q**uantization

LISTE DES FIGURES

Figure 1.1 : Principe de base de la tâche d'Identification Automatique du Locuteur 8

Figure 1.2 : Principe de base de la tâche de Vérification Automatique du Locuteur 9

Figure 1.3 : Principe de base de la tâche d'Indexation par Locuteurs d'un flux audio 9

Figure 1.4 : Principe de base de la tâche de suivi de locuteurs 11

Figure 1.5 : Indexation avec connaissances à l'avance des identités des locuteurs 11

Figure 1.6 : Indexation avec connaissances à l'avance des identités des locuteurs 12

Figure 1.7 : Représentation des points de rupture (points de changements de locuteurs) 13

Figure 1.8 : Exemple de Segmentation du signal de parole 14

Figure 1.9 : Principe du regroupement par locuteurs 14

Figure 1.10 : Segments homogènes 16

Figure 1.11 : Principe des fenêtres glissantes 18

Figure 1.12 : Exemple de regroupement par agglomération 20

Figure 1.13 : Exemple de regroupement séquentiel 21

Figure 2.1 : Problème d'indexation 23

Figure 2.2 : Représentation graphique du regroupement ascendant et descendant 34

Figure 3.1 : Quelques silhouettes de neurones 47

Figure 3.2 : Communication neuronale à l'aide des synapses 48

Figure 3.3 : Exemple d'unité neuronale 49

Figure 3.4 : Quelques types de fonctions de transfert 49

Figure 3.5 : Architecture d'un réseau monocouche à deux sorties 50

Figure 3.6 : Architecture d'un réseau multicouche à (n+1) niveaux 51

Figure 3.7 : Exemples d'un problème de discrimination à deux classes, avec une séparatrice 57

Figure 3.8 : Ensemble de points linéairement séparables 58

Figure 3.9 : Hyperplan optimal (en rouge) avec la marge maximale 58

Figure 3.10 : Exemple simple de transformation 61

Figure 3.11 : Principe du système de discrimination utilisant un RN basé sur la DRSC 65

Figure 3.12 : Fusion sérielle entre les classifieurs statistique et neuronal 66

Figure 3.13 : Fusion parallèle (au niveau des scores) de deux classifieurs 67

Figure 3.14 : Fusion sérielle-parallèle de deux classifieurs 68

Figure 3.15 : Segmentation entrelacée de la parole (*Seg* signifie segment) 69

Figure 3.16 : Banc de Filtres du type Hamming (dans ce cas : 12 filtres) 70

Figure 3.17 : Principe d'extraction des coefficients MFSC 71

Figure 3.18 : Organigramme d'apprentissage dans le cas de l'indexation avec connaissances a priori des locuteurs 72

Figure 3.19 : Organigramme d'étiquetage des segments 73

Figure 3.20 : Calcul de la distance minimale 74

Figure 3.21 : Étiquetage des segments 74

Figure 3.22 : Etapes de l'algorithme ISI avec une itération 76

Figure 3.23 : Fenêtre glissante d'analyse 77

Figure 4.1 : Erreur d'indexation pour différentes mesures statistiques 81

Figure 4.2 : Erreur d'indexation avec et sans l'ISI correction 82

Figure 4.3 : Erreur d'indexation pour différentes longueurs de segment 82

Figure 4.4 : Erreurs de discrimination de locuteurs dans DB1 86

Figure 4.5 : Erreurs de discrimination de locuteurs dans TB1 86

Figure 4.6 : Courbes ROC de discrimination de locuteurs 89

Figure 4.7 : Courbes ROC de segmentation 91

Figure 4.8 : Résultat de la segmentation (ensemble de segments homogènes) 93

Figure 4.9 : Regroupement par la distance $\mu_{G\beta}$ avec 10 groupes du document de Broadcast-News 94

Figure 4.10 : Indexation de référence du document de 30 minutes de Broadcast-News 94

LISTE DES TABLEAUX

Tableau 4.1 : Taux d'erreurs d'indexation avec différentes mesures (après l'algorithme ISI) 80

Tableau 4.2 : Taux d'erreurs utilisant $\mu_{Gc\beta}$ avec et sans correction ISI 81

Tableau 4.3 : Taux EER obtenus dans DB1, avec les différentes caractéristiques 85

Tableau 4.4 : Performances obtenues dans TB1, avec les différentes caractéristiques 85

Tableau 4.5 : Erreurs de discrimination obtenues dans les 4 BD : DB1, DB2, TB1 et TB2 87

Tableau 4.6 : Erreurs EER données par les différentes méthodes 92

Tableau 4.7 : Répartition des seg trouvés à l'issue de la segmentation des 3 premières mn de parole 93

Tableau 4.8 : Affichage des seg avec les nouveaux numéros et leurs durées pour les 3 mn du début du fichier de parole 93

Sommaire

INTRODUCTION GENERALE 1

CHAPITRE 1 : GENERALITES SUR LA RECONNAISSANCE DU LOCUTEUR

1.1. Introduction 5
1.2. Qu'est-ce-qu'une Reconnaissance ?
1.3. Notions sur la Variabilité du Signal Vocal
1.4. Reconnaissance Automatique du Locuteur (RAL) 6
1.4.1. Niveau de Dépendance au Texte 7
1.4.2. Différentes Tâches en RAL
1.4.2.1. Identification Automatique du Locuteur (IAL)
1.4.2.2. Vérification Automatique du Locuteur (VAL) 8
1.4.2.3. Détection de Locuteurs 9
1.4.2.4. Indexation par Locuteurs et ses Variantes
1.5. Indexation par Locuteurs 11
1.5.1. Quelques Applications Pratiques de l'Indexation Automatique par Locuteurs 12
1.5.2. Tâches et Phases de l'Indexation par Locuteurs 13
1.5.3. Hypothèses pour l'Indexation par Locuteurs 14
1.6. Segmentation en Locuteurs 16
1.6.1. Techniques de Segmentation par Locuteurs
1.6.1.1. Segmentation par Détection de Silences 17
1.6.1.2. Segmentation par Détection de Changement de Caractéristiques Acoustiques
1.6.1.3. Segmentation par Identification de la Nature des Segments 18
1.7. Regroupement en Locuteurs 19
1.7.1. Techniques du Regroupement
1.7.1.1. Regroupement Hiérarchique
1.7.1.2. Regroupement Séquentiel 21
1.8. Conclusion 22

Chapitre 2 : PRINCIPALES TECHNIQUES UTILISEES EN INDEXATION DES DOCUMENTS AUDIO

2.1. Introduction 23
2.2. Domaines d'Application de l'Indexation de Documents Audio
2.3. Paramétrisation Acoustique 24

2.3.1. Vecteurs Acoustiques 25
2.3.2. Suppression des Informations sur le Canal de Transmission
2.3.3. Coefficients Différentiels et Energie
2.4. Segmentation en Locuteurs 26
2.4.1. Segmentation Basée sur la Métrique 27
2.4.1.1. Critère d'Information Bayésien (BIC) 28
2.4.1.2. Rapport de Vraisemblance Généralisé (*Generalized Likelihood Ratio*, GLR) 29
2.4.1.3. Distance de Gish 30
2.4.1.4. Distance de Kullback-Leibler (KL ou KL2) 31
2.4.1.5. Distance de Divergence de Forme (*Divergence Shape Distance,* DSD)
2.4.1.6. Autres Distances 32
2.4.2. Segmentation non Basée sur des Métriques
2.4.2.1. Segmentation Basée sur les Silences
2.4.2.2. Segmentation Basée sur le Modèle
2.5. Regroupement en Locuteurs 33
2.5.1. Techniques de Regroupement Hiérarchique 34
2.5.1.1. Techniques de Regroupement Ascendant
2.5.1.2. Techniques de Regroupement Descendant 35
2.5.1.3. Combinaison des Méthodes de Regroupement 36
2.5.2. Regroupement Séquentiel
2.5.2.1. Utilisation de Sous-Espace Propre du Locuteur (SEP)
2.5.2.2. Utilisation du BIC (*Bayesian Information Criterion*) 38
2.5.3. Utilisation de l'Information de Support dans l'Indexation
2.5.3.1. Aide de l'Indexation en Utilisant la Transcription de Parole
2.5.3.2. Indexation en Locuteurs Utilisant l'Information Multi-Canaux 39
2.6. Conclusion

Chapitre 3 : METHODES PROPOSEES POUR L'INDEXATION

3.1. Introduction 41
3.2. Classifieur Mono-Gaussien ou "*Second Order Statistical Measures*" (SOSM)
3.2.1. Propriétés du Modèle Gaussien
3.2.2. Notion de Mesure de Similarité 42
3.2.3. Différentes Mesures Statistiques du 2ème Ordre
3.2.3.1. Mesure de Vraisemblance Gaussienne 43
3.2.3.1.1. Propriété de la μ_G 44
3.2.3.1.2. Inconvénient de la μ_G

3.2.3.2. Mesure Arithmétique- Géométrique Sphérique 45
3.2.3.3. Mesure de Déviation Absolue
3.2.4. Procédures de Symétrisation 46
3.3. Classifieur à Base de Perceptron Multi Couches (ou *Multi-Layer Perceptron,* MLP)
3.3.1. Aspect Biologique des Réseaux de Neurones (RN) 47
3.3.2. Fonctionnement et Modélisation du RN 48
3.3.3. Construction des RN 50
3.3.3.1. Architecture des Réseaux Monocouches
3.3.3.2. Mode d'apprentissage 51
3.3.3.2.1. Apprentissage Supervisé
3.3.3.2.2. Apprentissage non Supervisé 52
3.3.4. Perceptron
3.3.5. Perceptron Multicouche MLP 53
3.3.6. Algorithme de Rétropropagation du Gradient (RPG) 54
3.4. Classifieur à Base de Machine à Vecteurs de Support (SVM) 55
3.4.1. Historique des SVM
3.4.2. Principe Général des SVM 56
3.4.2.1. Discrimination Linéaire et Hyperplan Séparateur
3.4.2.2. Marge Maximale 58
3.4.2.3. Recherche de l'Hyperplan Optimal
3.4.2.4. Technique du "*Kernel trick*" (Cas non séparable) 60
3.4.2.5. Choix de la Fonction Noyau 62
3.5. Caractéristique Relative du Locuteur (RSC) 63
3.5.1. Dimension des Entrées des Classifieurs MLP et SVM
3.5.2. Notion de DRSC (*Diagonal of* RSC)
3.6. Fusion des Classifieurs 65
3.6.1. Fusion Sérielle 66
3.6.2. Fusion Parallèle 67
3.6.3. Fusion Sérielle-Parallèle
3.7. Algorithmes d'Indexation 68
3.7.1. Indexation avec Connaissances des Locuteurs
3.7.1.1. Segmentation et Etiquetage du Signal Audio
3.7.1.2. Indexation Entrelacée de la Parole : ISI 74
3.7.1.3. Regroupement en Locuteurs 76
3.7.2. Indexation Sans Connaissances a Priori du Locuteur

3.7.2.1. Segmentation du Signal Audio
3.7.2.2. Regroupement en Locuteurs 78
3.8. Conclusion

Chapitre 4 : EXPERIENCES ET RESULTATS DES DEUX TYPES D'INDEXATION

4.1. Introduction 79
4.2. Indexation Avec Connaissances a Priori des Locuteurs (Exp 1)
4.2.1. Corpus d'Indexation 80
4.2.2. Protocole Expérimental
4.2.3. Résultats Obtenus de l'Indexation Avec Connaissances des Locuteurs 83
4.3. Indexation Sans Aucune Connaissance des Locuteurs
4.3.1. Exp 2 : Test des Différentes Caractéristiques Réduites en Caractérisation du Locuteur
4.3.2. Exp 3 : Discrimination des Locuteurs 87
4.3.3. Exp 4 : Segmentation Sans Connaissances a Priori des Locuteurs 90
4.3.4. Exp 5 : Regroupement Séquentiel des Groupes de Locuteurs 92
4.4. Discussion sur les Expériences d'Indexation Sans Connaissance des Locuteurs 95
4.5. Conclusion 96

CONCLUSIONS GENERALES ET PERSPECTIVES 97

REFERENCES BIBLIOGRAPHIQUES 100

Introduction Générale

La science ne cesse d'évoluer, les moyens de calcul aussi et l'homme ne cesse de demander plus de confort et d'innovation. Ainsi sous cette continuelle pression, de nouvelles spécialités et domaines de recherche naissent et étoffent l'ancienne structure scientifique existante : d'où l'émergence de la spécialité : "Indexation par locuteurs des documents audio" venant compléter les différentes disciplines du domaine de la Reconnaissance du locuteur par la voix.

La voix reste un des moyens les plus utilisés par les êtres humains pour communiquer leurs idées et transmettre des informations au monde extérieur. En effet, la quantité d'informations disponibles sous forme de parole (téléphone, radio, télévision, meetings, conférences, internet, etc.) est énorme et le travail croît rapidement avec le temps, vus les moyens actuels de stockage, disponibles.

Beaucoup d'organismes numérisent et archivent les émissions radio et télé-diffusées, permettant ainsi aux auditeurs de les consulter (les voir ou les écouter) ultérieurement, à partir d'archives audiovisuelles, internet ou tout simplement à partir de Bases de Données (BD) spécialisées. La manipulation des documents numériques facilite le transfert d'informations, l'archivage des collections de grand volume et la consultation de ces documents. Cependant, lire, écouter ou regarder l'ensemble de ces documents multimédia, durant la recherche d'une information bien précise, apparaît une tâche très difficile (voire impossible) vus le temps qu'elle nécessite et le nombre de personnes à solliciter pour la recherche de cette information : imaginez quelqu'un rechercher les discours d'un ministre particulier, durant une époque bien définie et appartenant à un pays particulier, au sein de toutes les archives audiovisuelles de la télévision nationale.

De ce fait, l'accès direct à l'information recherchée, sans parcourir (écouter) la totalité des documents, suppose que les collections soient archivées, classées et décrites. L'exploitation efficace des collections nécessite de décrire le contenu et la structure des documents, en les annotant. Vu le volume de documents disponibles, l'annotation manuelle n'est pas envisageable, seuls des procédés automatiques (ou avec une faible intervention humaine) peuvent remplir cette tâche.

Pour résoudre ce problème, les technologies de la parole peuvent offrir une grande contribution par des techniques spécialisées. Concernant les documents audio, plusieurs clés d'indexation, et par conséquent, clés de recherche sont envisageables. La clé d'indexation (de recherche) peut être un mot, un sujet, un thème. Dans ce cas, tous les mots ou expressions relatifs à ce sujet sont recherchés (une autre clé particulière d'indexation est la voix du locuteur). Ainsi, il peut être intéressant de trouver les moments où un locuteur donné parle.

Aussi, il peut être intéressant de connaître la séquence ou l'enchaînement des locuteurs dans un document audio : en d'autres termes, connaître le moment où un locuteur prend la parole et le moment où il cède celle-ci à un autre. La connaissance de cette séquence repose sur **l'indexation de ce document audio.** L'indexation des documents permet, une recherche et une extraction beaucoup plus simple de l'information désirée.

La plupart du temps, quand une personne parle, sa parole est adressée à une ou plusieurs autres personnes, avec lesquelles elle communique. Ce qui fait que le signal de parole contient des paroles provenant de différentes personnes. Pour extraire plus tard une information de cet enregistrement, il est important de répondre à des questions comme : "Qu'est ce qui a été dit ?" pour transmettre le message, mais aussi "Qui est-ce qui l'a dit?" Vu que l'information varie selon la personne qui l'a prononcée, ainsi, l'information correspond à l'indication de la classe du locuteur. Quand il s'agit de classes correspondantes aux différents locuteurs présents dans l'enregistrement, ces techniques sont appelées indexation en locuteurs.

Les algorithmes d'indexation localisent chaque changement de locuteur dans le document audio et attribuent chaque segment à un locuteur présent dans le document. La sortie du système est un ensemble de segments homogènes contenant chacun un locuteur à la fois, laissant ainsi, aux systèmes d'identification, la tâche de déterminer les identités successives des personnes.

A l'heure actuelle, il y a trois domaines d'application de l'indexation qui ont attiré l'attention des chercheurs :

- la parole téléphonique : les systèmes d'indexation évalués lors des évaluations du NIST (*National Institute for Standards and Technology 2006*) utilisent un seul canal téléphonique pour les signaux de parole ;
- les informations télédiffusées (radio et TV) ;
- les meetings (cours, débats et conférences).

L'indexation par locuteurs d'un document audio englobe deux phases, en l'occurrence : la **segmentation** et le **regroupement**. La segmentation (appelée aussi la détection de changements de locuteurs) consiste en la recherche dans le document des moments où il y a eu un changement de locuteur. Tandis que le regroupement rassemble les différents segments homogènes en classes globales. Chaque classe contient l'intervention globale d'un locuteur dans le document.

Parmi les applications de l'indexation, on cite le suivi du locuteur. Dans ce cas, l'identité d'un (ou de plusieurs locuteurs) appelé "cible" est connue a priori par le système et le but sera alors de localiser son intervention dans le document audio.

Par ailleurs, les algorithmes d'indexation par locuteurs sont souvent utilisés dans les différents systèmes de technologie de la parole. Nous citons, à titre d'exemples :

- l'indexation par locuteurs et la transcription : en indexant le document audio selon l'intervention des locuteurs et en ajoutant une information supplémentaire aux systèmes de transcription, il devient plus simple de localiser l'information et de la traiter ;
- la segmentation en locuteurs aidant les systèmes de Reconnaissance Automatique de la Parole (RAP) : les algorithmes de segmentation sont utilisés pour découper le flux audio en petits segments, qui seront traités ensuite par les systèmes RAP ;
- modules de prétraitement pour les algorithmes basés sur le locuteur : l'indexation par locuteurs peut être utilisée avant le suivi de locuteur, l'identification du locuteur, la vérification du locuteur et d'autres algorithmes basés sur le locuteur, pour découper le flux audio en locuteurs individuels.

Notre thèse s'intéresse à l'indexation par locuteurs de documents audio enregistrés à partir des informations télédiffusées (ex : BBC, CNN). Cette tâche se complique à cause de la spontanéité de la parole qui constitue la première source de difficultés. De plus, les paroles des différents locuteurs peuvent se recouvrir et dans ce cas, il est difficile d'indexer (mélange de paroles).

Durant notre travail expérimental, la tâche d'indexation est exécutée une fois avec connaissances a priori des locuteurs et une autre fois, sans connaissances a priori des locuteurs (ni de leur nombre). Dans le premier cas, les identités des locuteurs engagés dans la conversation sont connues au préalable par le système. Par contre, dans le deuxième cas, le système ne dispose d'aucun modèle des locuteurs participants à la conversation ou de leur nombre.

Ainsi, dans le cadre de cette thèse, nous avons étudié l'indexation avec et sans connaissances a priori des locuteurs, de documents audio enregistrés à partir des informations télédiffusées. Nous avons aussi développé des algorithmes pour les deux types d'indexation.

Dans le cas de l'indexation avec connaissances a priori du locuteur, nous avons développé une nouvelle technique de segmentation que nous avons appelée segmentation entrelacée, utilisant un classifieur statistique mono-gaussien basé sur les métriques statistiques du 2ème ordre (SOSM). Tandis que pour l'indexation sans connaissances a priori du locuteur, nous avons implémenté trois types de classifieurs : le classifieur statistique mono-gaussien, le Multi-Layer Perceptron (MLP) et les Machines à Vecteurs de Support (SVM). De plus, des techniques de fusion entre ces classifieurs sont proposées avec différentes architectures pour améliorer les résultats de l'indexation. Concernant l'évaluation expérimentale, nous avons mené les différents tests sur la base de donnés HUB-4 Broadcast News.

Nous avons organisé notre document en quatre chapitres :

- dans le chapitre 1, nous définissons certaines généralités sur la reconnaissance du locuteur et ses différentes disciplines, ensuite, nous présentons l'indexation par locuteurs et ses applications. Nous détaillons après, les deux tâches principales de

l'indexation, notamment la segmentation et le regroupement, tout en définissant les différents algorithmes existants dans la littérature, en rapport avec ces deux tâches ;

- le chapitre 2 expose un état de l'art détaillé sur les principales caractéristiques et les différents algorithmes relatifs à la segmentation et au regroupement par locuteurs ;
- au chapitre 3, nous définissons les différents classifieurs implémentés durant notre étude. Nous expliquons ensuite les différents algorithmes que nous avons développés pour accomplir la tâche d'indexation avec et sans connaissances a priori des locuteurs. Nous présentons aussi dans ce chapitre les différentes architectures que nous avons développées pour fusionner les différents classifieurs utilisés ;
- le dernier chapitre expose les résultats expérimentaux obtenus durant cette étude, avec des interprétations et des conclusions.

Une conclusion générale clôture ce travail de thèse, et quelques perspectives concernant les deux tâches étudiées (segmentation et regroupement) sont proposées.

Finalement, des références bibliographiques ainsi que des annexes utiles sont mises à la disposition du lecteur pour plus de détails.

Chapitre 1 :

GENERALITES SUR LA RECONNAISSANCE DU LOCUTEUR

1.1. Introduction

Cet important chapitre définit les différents domaines de la RAL telles que l'Identification et la Vérification Automatique du Locuteur et des tâches plus récentes comme le suivi de locuteur ou l'indexation par locuteur de flux audio. Ensuite, il présente les applications de l'indexation par locuteurs ainsi que ses étapes fondamentales notamment la segmentation en locuteurs et le regroupement en locuteurs.

1.2. Qu'est-ce-qu'une Reconnaissance ?

Le terme "reconnaissance" est défini comme étant l'identification de quelque chose, sachant qu'on doit connaître au préalable le modèle de référence de cette chose.

La Reconnaissance des formes consiste, alors, à identifier une forme donnée, après avoir déjà conçu le modèle de référence de ces formes. Cette phase s'appelle la phase d'apprentissage.

La reconnaissance automatique d'un individu consiste à utiliser des caractéristiques physiques dans le but de faire une discrimination entre les différents individus. Pour ce faire, plusieurs caractéristiques sont proposées dans la littérature : la photographie du visage, les empreintes digitales, les traits génétiques ou encore le signal de parole.

La caractéristique la plus pratique reste le signal de parole, vues la simplicité de son extraction et la possibilité d'offrir une authentification à distance.

L'authentification par la voix est appelée "Reconnaissance Automatique du Locuteur" (RAL). Cependant, ici nous rencontrons un problème majeur, qui est défini par la difficulté de trouver des caractéristiques pertinentes en discrimination : ce problème implique la nécessité de trouver des caractéristiques possédant une grande variabilité inter-locuteur et une faible variabilité intra-locuteur.

1.3. Notions sur la Variabilité du Signal Vocal

Il existe deux types de variabilités (pour une caractéristique acoustique donnée), la variabilité :

- **intra-locuteur** : elle identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de

manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.

- **inter-locuteur** : la voix de chaque individu possède des qualités qui lui sont propres. L'âge, le sexe, le tempérament du locuteur (et bien d'autres facteurs encore) lui confèrent une identité vocale originale qui est la combinaison de multiples paramètres dont la hauteur (pitch), l'intensité et le timbre de sa voix, la qualité de son articulation, ou encore son accent (national et régional) ;

Le Coefficient de Wolf "CW" est le rapport de la variabilité inter-locuteur sur la variabilité intra-locuteur [1].

Une grande variabilité inter-locuteur : indique qu'on peut séparer, facilement, les locuteurs par leurs caractéristiques.

Une petite variabilité intra-locuteur : indique que chaque locuteur peut être représenté par une référence qui représente très bien ce locuteur.

Par conséquent, un grand coefficient de Wolf indique, alors, que le paramètre choisi est pertinent en identification du locuteur et devrait donner un bon score de reconnaissance.

Dans ce qui suit, nous définirons les différentes étapes associées à la RAL.

1.4. Reconnaissance Automatique du Locuteur (RAL)

La caractérisation automatique du locuteur est un vaste domaine dans lequel la "machine" a pour tâche d'extraire du signal de parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc. Elle s'applique à différents thèmes de recherche traitant des informations véhiculées par la voix tels que la classification d'individus, ou l'étude psychique ou physiologique d'une personne.

La RAL est un sous-problème de la caractérisation automatique du locuteur. Son objectif est de reconnaître l'identité d'une personne à l'aide de sa voix. La variabilité de la parole entre locuteurs (variabilité inter-locuteur) est l'essence même de la RAL. Sans cette variabilité, il serait impossible de reconnaître une voix parmi plusieurs voix possibles.

La RAL, contrairement à la Reconnaissance Automatique de la Parole (RAP) s'intéresse tout particulièrement aux informations extra-linguistiques véhiculées par un signal vocal (signal de parole). Pourtant, la RAL a très souvent bénéficié des avancées de la RAP. Ainsi, de nombreuses techniques ont été appliquées en RAP avant d'être adaptées au domaine de la RAL. Les applications de la RAL sont principalement liées aux problèmes d'authentification ou de confidentialité.

1.4.1. Niveau de Dépendance au Texte

Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte.

En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système.

En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

Concernant le mode dépendant du texte, une terminologie plus fine peut être donnée à un système suivant l'application visée, systèmes à :

- messages fixés : la personne est contrainte de prononcer un message, qu'elle aurait au préalable (mots de passe personnalisés : [2] et [3]) ou qui sera imposé par le système ;
- messages prompts : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle [4] ou auditive [5]. Ces systèmes ont pour première motivation de se protéger des attaques de personnes malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne ;
- unités segmentales : la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connus du système.

1.4.2. Différentes Tâches en RAL

L'Identification et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL, où plusieurs chercheurs sont en train de travailler : plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio (travaux de Johnson et de Delacourt) [6] et [7] ou le Suivi de Locuteurs ou de nouvelles variantes telles que la détection du locuteur dans une conversation [8] et [9].

1.4.2.1. Identification Automatique du Locuteur (IAL)

L'Identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vue schématique (figure 1.1), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL.

Deux modes sont proposés en IAL, l'identification en ensemble :

- fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système ;

- ouvert pour lequel le locuteur peut ne pas être connu.

En mode "ensemble ouvert", le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée. De par son principe - déterminer une identité parmi les identités potentielles - les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

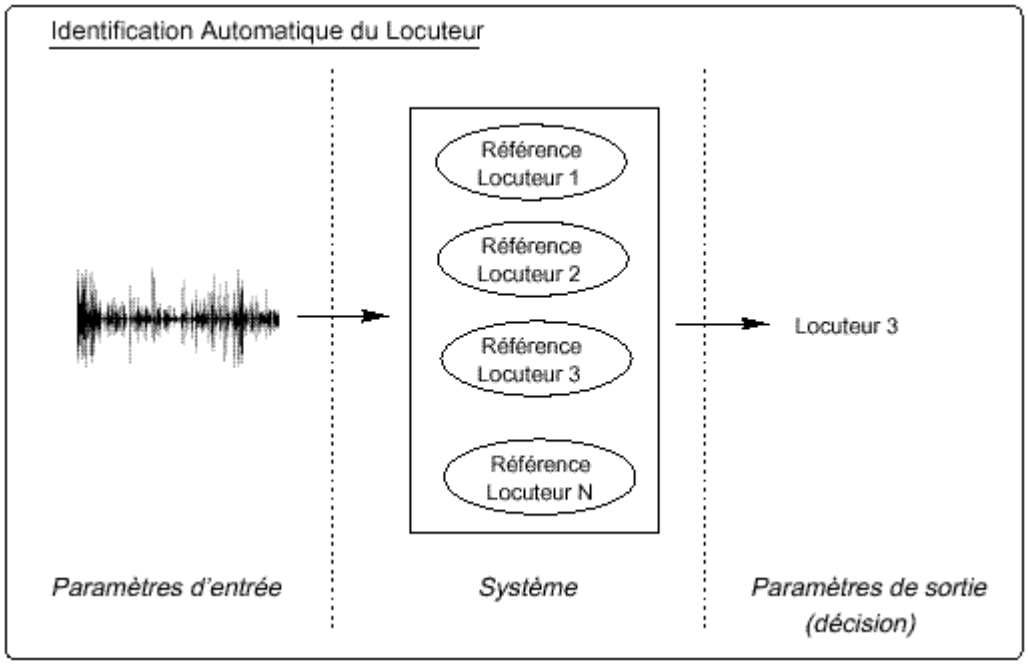

Figure 1.1 : Principe de base de la tâche d'Identification Automatique du Locuteur [10].

1.4.2.2. Vérification Automatique du Locuteur (VAL)

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu (figure 1.2). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est inférieure au seuil, l'individu est accepté, dans le cas contraire, l'individu est considéré comme un imposteur, est rejeté.

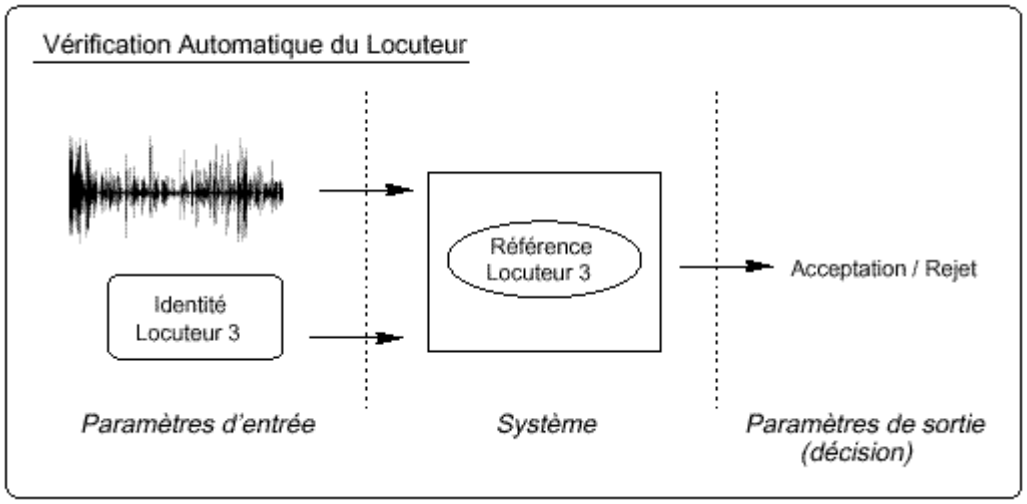

Figure 1.2 : Principe de base de la tâche de Vérification Automatique du Locuteur [10].

1.4.2.3. Détection de Locuteurs

La détection de locuteurs dans un flux audio est une variante de la VAL. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux audio monolocuteur, la tâche de détection se résume à la tâche de vérification.

1.4.2.4. Indexation par Locuteurs et ses Variantes

La tâche d'Indexation Automatique par Locuteurs consiste à cibler les interventions des locuteurs dans un flux audio (figure 1.3). En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité.

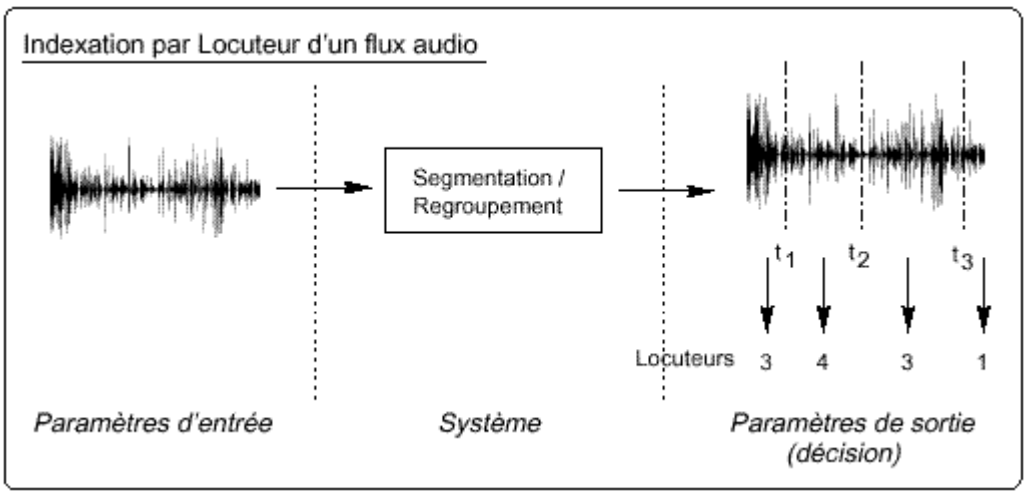

Figure 1.3 : Principe de base de la tâche d'Indexation par Locuteurs d'un flux audio [10].

Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation "aveugle" en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet finalement d'identifier les différents locuteurs présents dans le document.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'Indexation par Locuteurs d'un flux audio (figure 1.4). Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La simplification réside dans le fait que le système de suivi de locuteurs connaît nécessairement les locuteurs présents dans le document à indexer ou, du moins, ceux dont il doit détecter les interventions. Il possède une référence caractéristique pour chacun des locuteurs [11].

Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

- une segmentation "aveugle" en locuteurs, identique à celle employée pour l'Indexation par Locuteurs d'un flux audio, est appliquée sur le signal de test. Les segments - résultant de la segmentation - sont soumis à un système de VAL classique afin de déterminer les segments appartenant effectivement au locuteur cible [12] ;
- le signal de test est découpé en une suite de blocs de trames, de taille fixe (ce découpage en blocs est entièrement indépendant des événements acoustiques observés sur le signal de parole.), sur lesquels sont appliqués un système de VAL. Un processus de décision, à base de seuils, permet en phase finale d'accepter ou de rejeter les blocs appartenant au locuteur cible [13] et [14] ;
- la troisième approche est similaire à la précédente excepté pour le processus de décision. Dans ce cas, la décision repose sur un Modèle de Markov Caché, HMM ergodique composé d'états correspondant au locuteur cible, à un modèle générique de parole et de non parole (silence, bruit, etc.) [15] et [16].

Les systèmes d'Indexation Automatique par Locuteurs d'un flux audio sont principalement utilisés pour le traitement de bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débats, etc.). D'autres applications sont envisageables comme la recherche de messages par locuteur sur un répondeur téléphonique ou sur une boîte vocale.

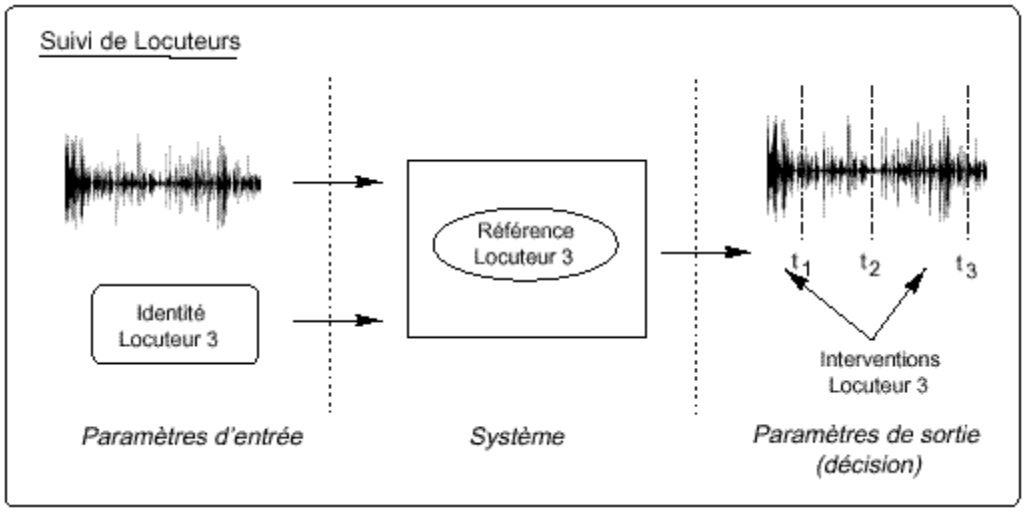

Figure 1.4 : Principe de base de la tâche de suivi de locuteurs [10].

1.5. Indexation par Locuteurs

L'indexation par locuteurs est une méthode fiable qui peut être utilisée pour l'archivage automatique des documents audio. En effet, l'étiquetage de ces documents par les identités des locuteurs intervenant dans le discours permet ultérieurement une accessibilité simple, facile et rapide à ces documents.

L'indexation par locuteurs d'un document audio consiste à reconnaître la séquence ou l'enchaînement des locuteurs engagés dans la conversation. En d'autres termes, il s'agit de savoir ***"Qui parle ? et Quand ?"*** afin de saisir la cohérence du dialogue (figure 1.5). Le résultat du processus de l'indexation sera de la forme suivante : le locuteur *A* parle de l'instant t_1 à l'instant t_2, puis le locuteur *B* intervient de t_3 à t_4, un autre locuteur *C* intervient entre t_4 et t_5, puis *A* reprend la parole de t_5 à t_6, ensuite le locuteur *C* parle de t_6 à t_7, etc.

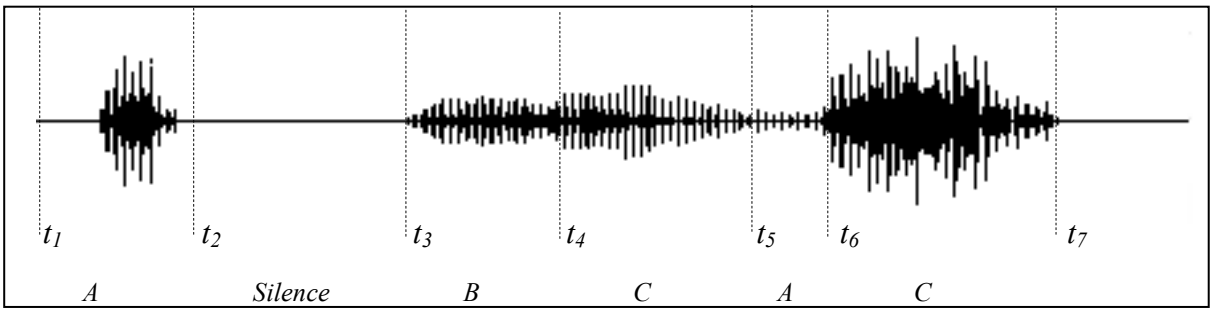

Figure 1.5 : Indexation avec connaissances à l'avance des identités des locuteurs.

Il est important de préciser que l'identité des locuteurs n'est pas pour autant connue. Dans ce cas, le système ne dispose d'aucun modèle des locuteurs (figure 1.6).

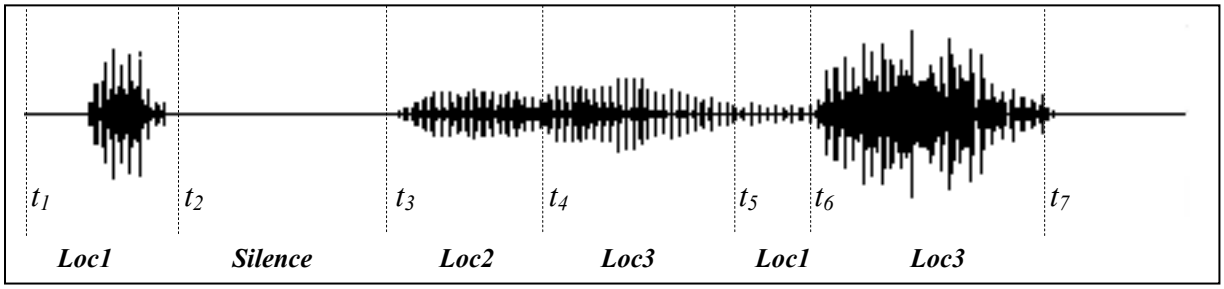

Figure 1.6 : Indexation sans connaissances des identités des locuteurs.

1.5.1. Quelques Applications Pratiques de l'Indexation Automatique par Locuteurs

Bien que la finalité de l'indexation soit la recherche de l'information, l'indexation peut aussi être abordée dans d'autres domaines de recherche comme la RAP, la RAL, etc. En effet, l'indexation par locuteurs s'applique dans des domaines aussi variés tels que :

- l'indexation de bases de données audio. C'est en effet son premier rôle. Couplée à un processus d'identification du locuteur, elle peut permettre par exemple la recherche de tous les discours prononcés par telle personnalité politique, cela peut servir par exemple à calculer son temps de paroles au cours d'une compagne électorale. Cela peut aussi permettre la recherche des paroles du journaliste présentateur et l'extraction à partir de ces paroles des thèmes abordés dans le journal télévisé ;
- la recherche des messages par locuteur sur un répondeur téléphonique ou sur une boîte vocale. Avec l'explosion de la téléphonie, de plus en plus de services sont proposés. Parmi ces services, figureront peut être bientôt, la classification par locuteurs des messages déposés sur un répondeur téléphonique ou sur une boîte vocale. Cette classification reposera alors sur une indexation par locuteurs de ces messages ;
- dans un système de poursuite de locuteur, un modèle du ou des locuteurs "cible(s)" est disponible afin de nous permettre de reconnaître le locuteur intervenant pendant la séquence de parole où il intervient, ceci peut se faire en rajoutant une étape de vérification qui va permettre de comparer les caractéristiques vocales du locuteur intervenant dans le segment du document audio avec celles du modèle de référence du locuteur "cible" [13] et [17].
- la transcription automatique de documents audio, en particulier les nouvelles radio- ou télé-diffusées [7], [18] et [19]. Les systèmes de transcription automatique utilisent des modèles de parole pré-entraînés sur de larges bases de données, de sorte qu'ils ne contiennent aucune spécificité du locuteur. Or, il a été prouvé que lorsque ces modèles sont adaptés aux locuteurs présents dans le document audio, alors le taux de reconnaissance s'en trouvait amélioré. Ainsi, une étape préliminaire dans ces systèmes de transcription automatique consiste à indexer par locuteurs, permet alors l'extraction des données correspondantes à chaque locuteur. Ces données servent ensuite à adapter

les modèles de parole aux locuteurs et le processus de transcription peut alors travailler sur les segments de chaque locuteur présent dans le document audio [7] ;

- le repérage automatique de messages issus de personnes suspectes dans des conversations téléphoniques (intérêt sécuritaire). Dans cette application, l'indexation est utilisée pour un filtrage de données par locuteurs [20]. Adaptée à un système de vérification par locuteurs, le système d'indexation peut détecter tous les messages ou les interventions d'une personne bien déterminée (suspecte) dans une collection de documents audio enregistrées (ex : conversations téléphoniques enregistrées).

1.5.2. Tâches et Phases de l'Indexation par Locuteurs

Du fait que l'indexation automatique en locuteurs consiste à rechercher les points de changement de locuteur (appelés aussi points de rupture) [21] dans un document audio multilocuteur et reconnaître ensuite le nombre des différents intervenants et éventuellement leurs identités (dans le cas d'une indexation par connaissances a priori de locuteurs), l'indexation se compose de deux tâches principales :

- la segmentation en locuteurs des documents sonores consiste à découper le flux audio en segments homogènes les plus longs possibles et ne contenant que les paroles d'un seul locuteur. A l'issue de la segmentation, les segments ne sont pas encore étiquetés ; c'est-à-dire que le locuteur qui a prononcé les paroles contenues dans un segment n'est pas encore identifié. En effet, le processus de segmentation est utilisé pour chercher les points de rupture (ou points de changement de locuteurs) dans un discours multilocuteur (figure 1.7). Ces points représentent les moments où un des locuteurs cède la parole à un autre locuteur. De ce fait, ces points délimitent les instants de début et de fin de chaque intervention des locuteurs dans le document audio (figure 1.8) ;

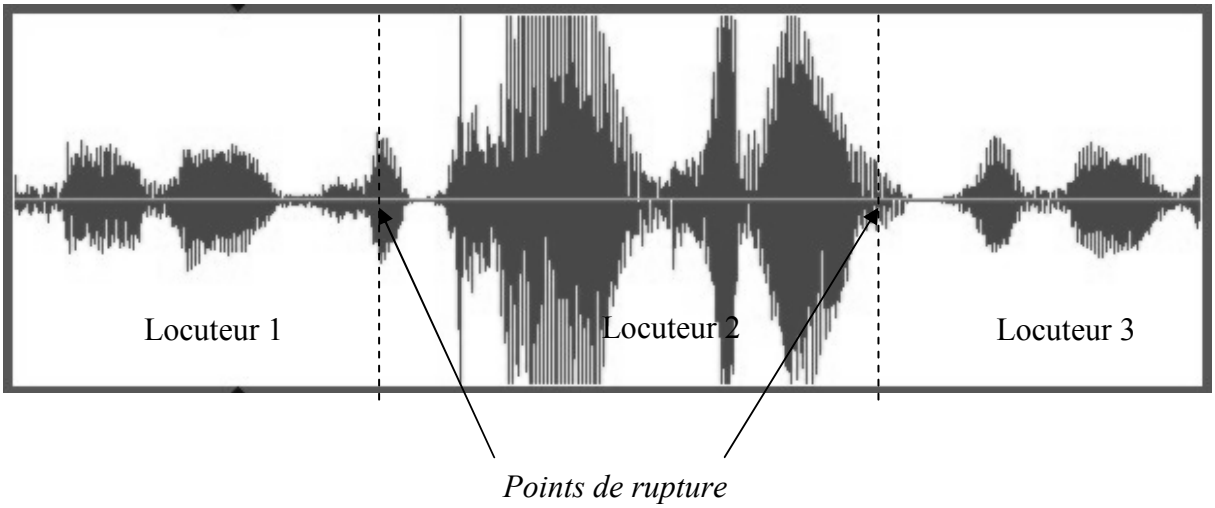

Figure 1.7 : Représentation des points de rupture (points de changements de locuteurs).

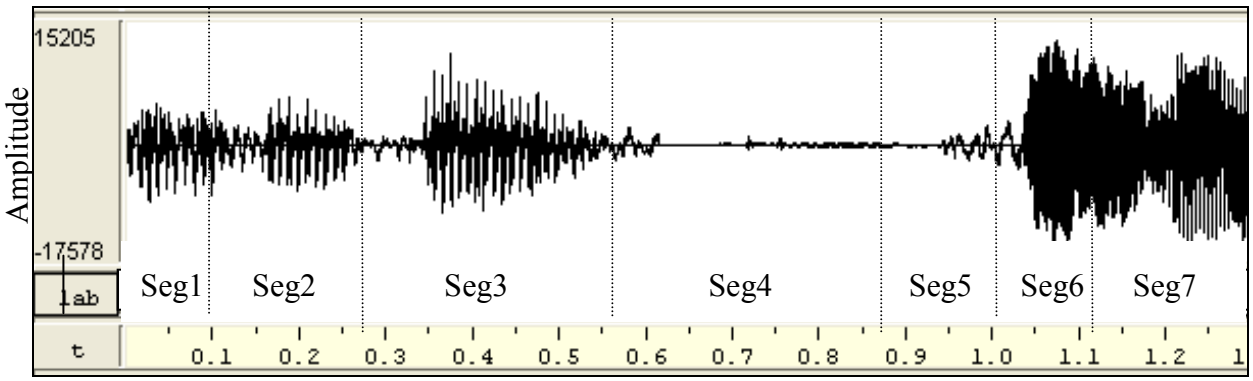

Figure 1.8 : Exemple de Segmentation du signal de parole (voir les segments en bas de la figure).

- dans cette phase, tous les segments homogènes appartenant à un même locuteur seront regroupés ensemble, cette opération est effectuée pour tous les locuteurs présents dans le document. A la fin de cette opération, nous obtiendrons des groupes (ou classes) dont le nombre est égal au nombre de locuteurs participant à la conversation. Chaque groupe contiendra l'intervention totale d'un locuteur, dans le document audio traité avec un index de ce locuteur. Ce qui donne à cette tâche un grand intérêt pratique (par exemple : dans le cas où on veut archiver toutes les paroles d'un locuteur "cible" à partir de ces interventions différentes dans des émissions radio ou TV) (figure 1.9).

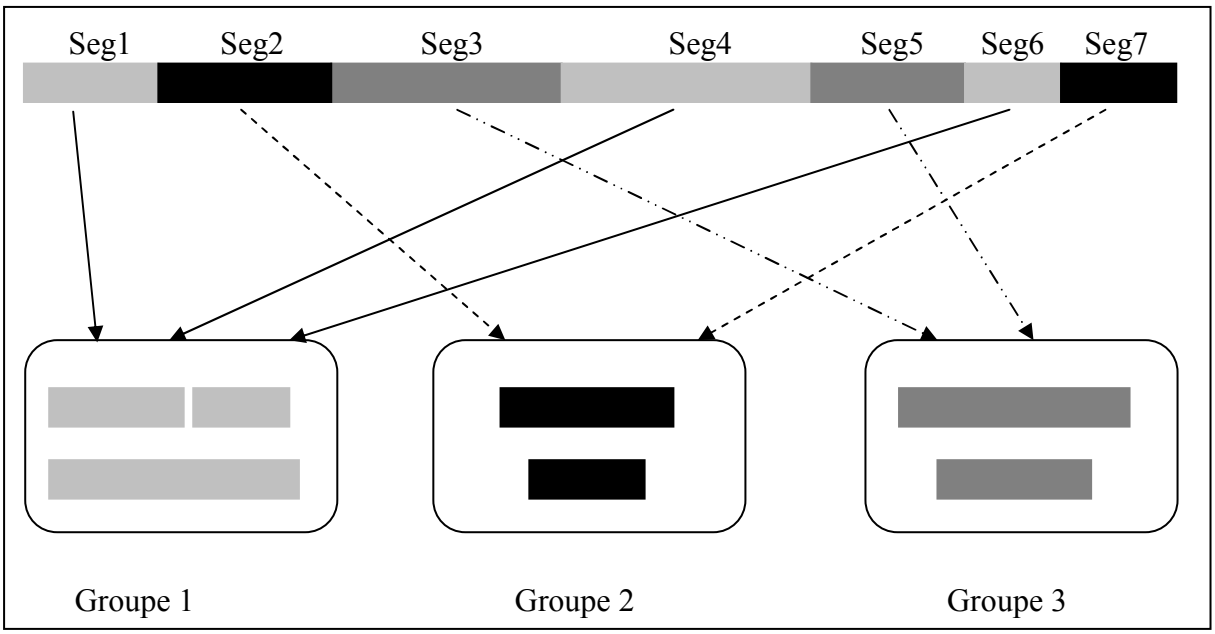

Figure 1.9 : Principe du regroupement par locuteurs.

1.5.3. Hypothèses pour l'Indexation par Locuteurs

Au cours de cette thèse, nous avons traité l'indexation automatique des émissions de radio et de télévision, ces émissions comprennent des discours multilocuteur. Pour cela, nous nous sommes fixés les hypothèses suivantes :

- **les locuteurs :** nous proposons de traiter les documents audio, une fois en utilisant les modèles des locuteurs (modèle de référence) et une autre fois sans aucune connaissance au préalable des identités des locuteurs :

  - avec modèle de référence : dans ce cas, le système d'indexation possède en mémoire des caractéristiques vocales (un modèle de référence) de chaque locuteur présent dans le document audio. Ce type d'indexation peut être adapté, par exemple, dans le cas de discussions où tous les locuteurs sont connus au préalable (nécessite une phase d'apprentissage) ;

  - aucune connaissance des locuteurs (pas de modèle, pas de phase d'apprentissage). En pratique, il n'y a pas toujours, à notre disposition, des données d'apprentissage pour construire un modèle sophistiqué du locuteur. Par exemple, dans un journal radio ou télédiffusé, il est rare de posséder des données d'apprentissage pour une personne interviewée lors d'un reportage.

- **la langue :** quant à la langue, il est possible qu'un document audio mélange plusieurs langues ;

- **le nombre de locuteurs est inconnu :** dans le cas où nous n'avons aucune connaissance des locuteurs, le nombre de ces derniers devient inconnu aussi. En considérant cette hypothèse, on élargit le domaine d'application et notre système d'indexation sera capable de traiter n'importe quel document audio.

  Nous pourrons éventuellement supposer que le nombre de locuteurs est de deux dans une conversation téléphonique (si ces locuteurs ne sont pas en pluri-conférence). Mais, il reste difficile de faire au préalable une hypothèse sur le nombre d'intervenants dans un document audio multilocuteur ;

- **avec et sans détection des silences :** dans certains discours qui contiennent des silences de durée importante (quelques secondes), il est intéressant de détecter la partie parole de la partie non-parole. Mais dans le cas de la parole spontanée, les silences entre les locuteurs sont très petits et parfois les locuteurs se coupent la parole (recouvrement de parole) ;

- **aucun contrôle sur l'environnement :** les enregistrements sont réalisés en environnement réel. Les documents audio traités peuvent contenir en plus de la parole, du bruit, de la musique, etc ;

- **les personnes ne parlent pas simultanément :** cette hypothèse peut sembler peu réaliste. En effet, il arrive souvent qu'au cours d'une conversation, une personne commence à parler alors que la précédente n'a pas achevé sa phrase. Il y a donc recouvrement des paroles des deux locuteurs. Cependant, ce type d'évènement est difficile à indexer même à la main. Faut-il considérer qu'il n'y a qu'une personne et dans ce cas laquelle, ou faut-il considérer qu'il y a deux personnes ? Enfin, cette

hypothèse reste réaliste dans le cadre de journaux radio ou télédiffusés car la parole est bien souvent préparée et non spontanée. Il y a donc peu de recouvrement de parole. Les méthodes utilisées jusqu'ici arrivent difficilement à détecter si plusieurs locuteurs interviennent en même temps ;

- **pas de contrainte de temps réel :** l'indexation est une tâche qui est réalisée une fois pour toute sur une base de données, c'est un processus "*off-line*". Aussi, cette opération peut prendre un temps non négligeable, sans contrainte de temps préalable.

1.6. Segmentation en Locuteurs

Dans le domaine de l'indexation automatique par locuteurs, la segmentation consiste à détecter (chercher) les points de changement (de rupture) de locuteurs [22], afin d'obtenir des segments homogènes, où chaque segment contient la parole d'un seul locuteur à la fois, d'autre part, le résultat de segmentation indique que deux segments adjacents appartiennent toujours à deux locuteurs différents (figure 1.10).

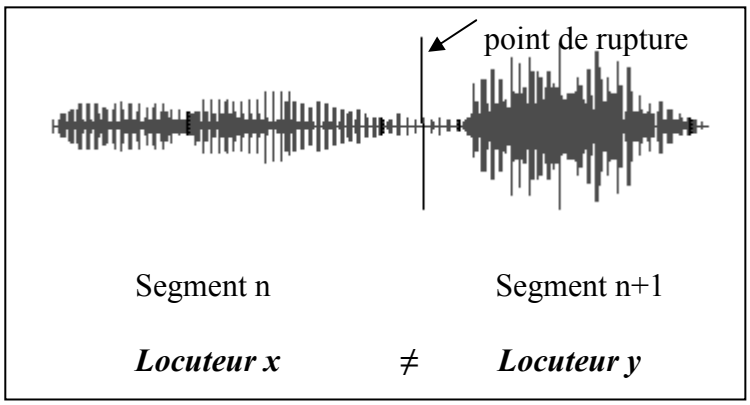

Figure 1.10 : Segments homogènes.

Le problème de la segmentation revient à la recherche de ces points de rupture dans le document audio. Plusieurs techniques ont été proposées pour la détection de ces points, la segmentation par :

- détection de silences ;
- détection de changement de caractéristiques acoustiques [23] et [24] ;
- identification de la nature des segments.

1.6.1. Techniques de Segmentation par Locuteurs

Nous citons trois types de techniques qui sont les plus utilisées en segmentation.

1.6.1.1. Segmentation par Détection de Silences

Cette technique de segmentation repose sur la supposition que les interventions provenant de locuteurs différents sont séparées par des silences significatifs. La détection de ces silences permet la détection de ces points de rupture.

Un silence est significatif s'il a une durée suffisante pour être détecté. Les silences sont généralement caractérisés par un très faible niveau d'énergie. Il existe plusieurs méthodes pour les détecter [25], par l'utilisation :

- de la puissance moyenne : étant donné que le niveau d'énergie du silence est faible, il en est de même de sa puissance moyenne. Cependant il faut définir un seuil sur cette dernière, pour séparer le silence de la parole [26] et [27] ;
- du Taux de Passages par Zéro (TPZ), qui consiste à calculer le nombre de fois pour lequel le signal coupe l'axe des amplitudes nulles (en montée ou descente du signal), par unité de temps. Le silence sera alors caractérisé par un TPZ élevé. Cette technique a été utilisée et couplée à une méthode de détection de début et de fin de phrase [28].

Comme la méthode précédente, cette approche nécessite la détermination d'un seuil au delà duquel le TPZ indiquera la présence d'un silence.

Par conséquent, la détection des silences est liée à deux paramètres :

- un seuil de décision (la frontière entre la parole et le silence). Ce seuil dépend des conditions d'enregistrement. Pour cela, il peut être différent d'un document audio à un autre et voire même aussi au sein d'un même document. De ce fait, il est difficile de le déterminer automatiquement ;
- une durée minimale des segments de silences. Cette durée est aussi déterminée empiriquement et ajustée de manière qu'on puisse détecter les silences entre les interventions de locuteurs différents.

Malgré que la méthode de détection des silences peut être efficace dans le cas de la parole préparée, où il existe souvent des silences entre les interventions de locuteurs différents ; cette méthode peut ne pas être adaptée aux enregistrements de conversations téléphoniques ou les débats, car dans ce type de documents, les locuteurs se coupent souvent la parole.

1.6.1.2. Segmentation par Détection de Changement de Caractéristiques Acoustiques

Cette approche se base sur la détection de changement des caractéristiques acoustiques tout le long du signal audio. La méthode ne détermine pas la nature des caractéristiques présentes dans le signal, mais seulement leurs changements (c'est-à-dire : détection des changements de locuteur, détection des changements parole/musique, parole/silence, parole/ bruit, etc).

Les techniques généralement utilisées dans cette approche, pour détecter ces changements, calculent une mesure de similarité entre deux fenêtres adjacentes du signal audio. La présence ou non d'une rupture à un instant 't' du document audio et qui représente le point de contact

de ces deux fenêtres (figure 1.11) est déterminée en comparant la valeur de la mesure à un seuil donné.

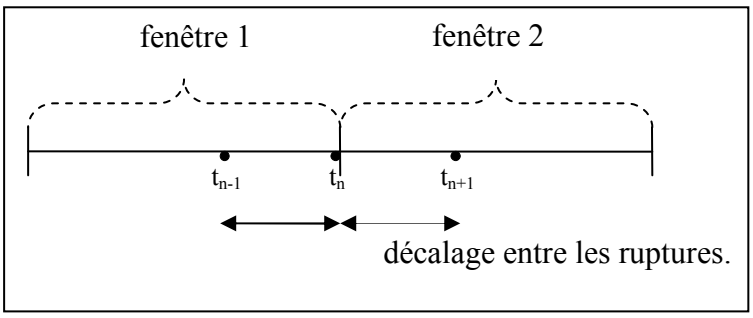

Figure 1.11 : Principe des fenêtres glissantes.

Les paramètres à considérer dans cette approche sont :

- le choix de la mesure. Il s'agit de la distance à utiliser pour discriminer entre les caractéristiques acoustiques ;
- le décalage entre une rupture et une autre, qui est estimé par la durée avec laquelle les deux fenêtres sont glissantes (sur la figure 1.11 : $t_n - t_{n-1}$) ;
- la durée de la fenêtre de part et d'autre de la rupture. En effet, cette durée dépend essentiellement de la mesure choisie (exemple : pour les méthodes statistiques, la durée minimale pour une estimation optimale des caractéristiques acoustiques du locuteur est de 2 secondes) ;
- le seuil de décision, qui représente la valeur de la mesure à partir de laquelle, on décide s'il y a une rupture ou non. Ce seuil reste variable même en utilisant la même mesure et dépendant du type d'enregistrement audio (bruit de fond, parole téléphonique, etc).

1.6.1.3. Segmentation par Identification de la Nature des Segments

Dans cette approche, le système d'indexation possède en mémoire des caractéristiques acoustiques des différentes classes du signal (parole, silence, musique, etc). Ces caractéristiques sont obtenues à partir de la phase d'apprentissage du système d'indexation. Cela permettra par la suite d'étiqueter les différents segments, en comparant leurs caractéristiques à celles des modèles de référence et d'identifier la nature du segment par la classe du modèle de référence le plus proche.

Contrairement aux deux précédentes méthodes (détection des silences et détection des changements de caractéristiques acoustiques), cette approche ne nécessite aucun seuil de décision (utilisation de la règle du plus proche voisin). Toutefois, une extension de cette approche est possible pour la détection des changements entre les locuteurs, dans le cas de l'indexation par connaissances a priori des locuteurs.

1.7. Regroupement en Locuteurs

A l'issue de la phase de segmentation, un ensemble de segments homogènes est disponible (chaque segment contient la parole d'un seul locuteur). La tâche de regroupement consistera à regrouper tous les segments d'un même locuteur en un seul groupe, étape par étape, jusqu'au balayage de tout le flux audio.

Le nombre final des groupes doit être égal au nombre des locuteurs participant à la conversation, puisque chaque groupe représente l'intervention d'un locuteur présent dans le flux audio. Toutefois, dans le cas de l'indexation sans connaissances a priori des locuteurs, le nombre et l'identité des différents intervenants dans le document sonore traité ne sont pas connus au départ, pour cela, la classification (regroupement) est qualifiée de non-supervisée. Cependant, plusieurs techniques de regroupement ont été introduites comme celles du **regroupement hiérarchique** (*hierarchical clustering*) ou du **regroupement séquentiel**, pour résoudre le problème de la classification non-supervisée (pas de connaissances des locuteurs et de leur nombre dans le document audio).

En revanche, dans le cas de l'indexation avec connaissances des locuteurs, la tâche de regroupement devient plus simple, car les groupes sont bien déterminés par les différentes identités des locuteurs et leur nombre est bien connu au départ. Ainsi, le problème du regroupement peut être résolu par un système d'Identification Automatique du Locuteur (IAL) [29] et [30].

1.7.1. Techniques du Regroupement

Il existe dans la littérature certaines techniques pour le regroupement d'un ensemble d'éléments de manière itérative. Ce sont les techniques de regroupement hiérarchique, qui ont été étudiées dans différents domaines, comme le domaine de traitement d'images ou dans certaines branches de mathématiques, et les techniques de regroupement séquentiel qui prennent en considération le voisinage entre les éléments.

1.7.1.1. Regroupement Hiérarchique

Nous distinguons dans ce type de regroupement deux techniques, le regroupement par :

- agglomération (*Agglomerative clustering*), appelé aussi regroupement ascendant (*top-bottom clustering*) ;
- division (*divisive clustering*), appelé aussi regroupement descendant (*top-down clustering*).

Le regroupement par agglomération considère chaque élément ou objet (segment de parole, dans notre cas) comme un groupe (*cluster*). A chaque itération du processus de regroupement, les deux plus proches groupes seront regroupés suivant un critère de regroupement (*merging criterion*). Ce processus est répété jusqu'à ce qu'un critère d'arrêt (*stopping criterion*) soit satisfait (figure 1.12). Ce critère peut être un seuil significatif, pour un calcul de distance de

similarité entre les groupes, auquel cette distance sera comparée et à partir duquel, on peut décider si les deux groupes appartiennent à la même classe (locuteur) ou non.

- distance ($groupe_i$, $groupe_j$) ≤ Seuil alors même locuteur.
- distance ($groupe_i$, $groupe_j$) > Seuil alors locuteurs différents.

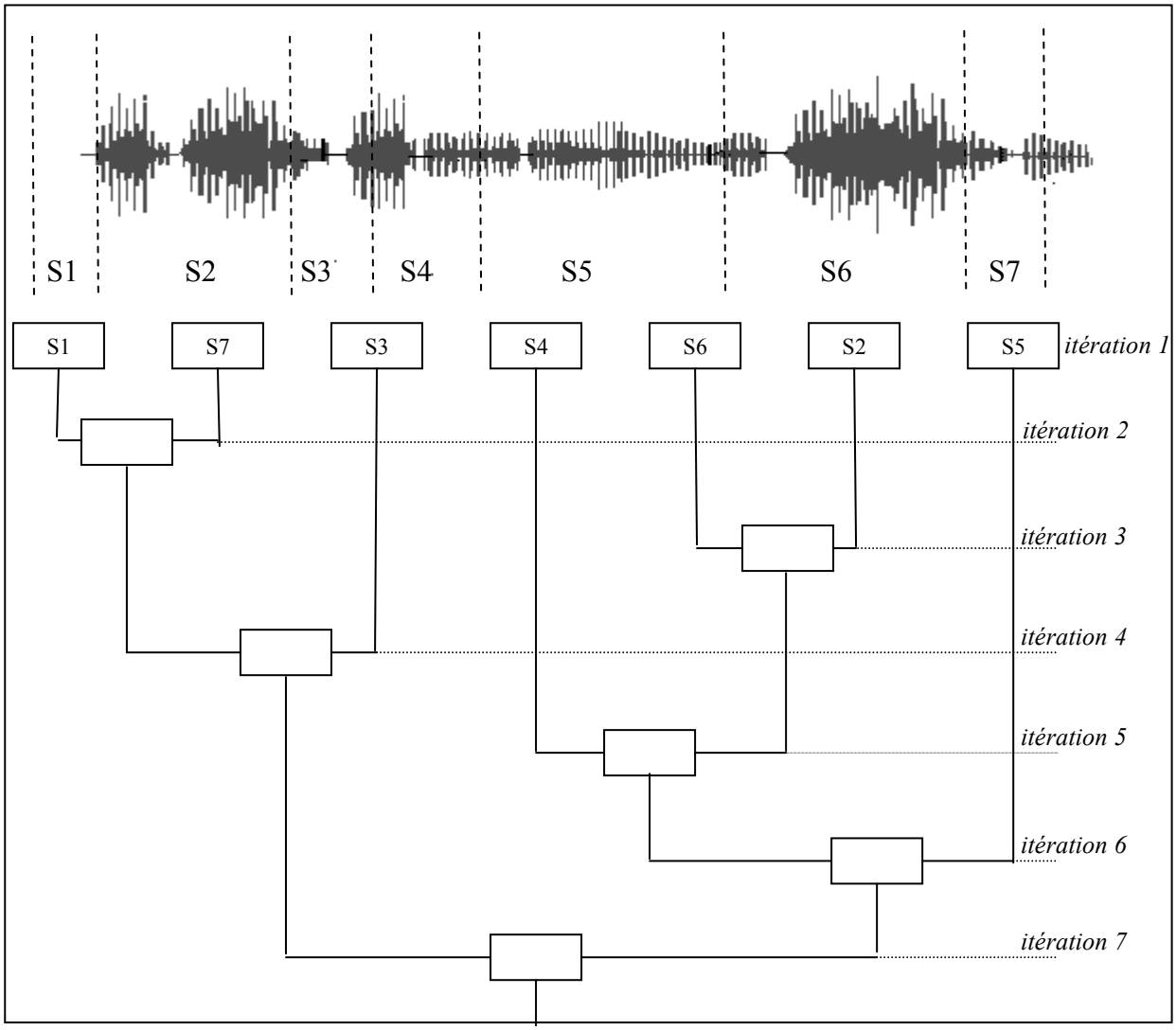

Figure 1.12 : Exemple de regroupement par agglomération.

A l'inverse du regroupement par agglomération, le regroupement par division considère au début les différents éléments (segments de parole) ne formant qu'un seul groupe et à chaque itération, on divise un des groupes selon un critère de division (*splitting criterion*). Le processus est répété jusqu'à la satisfaction d'un certain critère d'arrêt.

Deux paramètres doivent être considérés dans le regroupement hiérarchique, le critère :

- de regroupement : qui peut être le calcul de distance entre les différents groupes à chaque itération. La similarité entre les groupes se fera selon cette distance (valeur minimale de la distance à chaque itération entre 2 groupes regroupe ces derniers) ;

- d'arrêt : dans le cas de calcul de distance, on peut définir le critère d'arrêt comme étant un seuil auquel cette distance minimale sera comparée à chaque itération, pour décider le regroupement ou non de ces deux groupes.

Notons aussi, que ces deux paramètres dépendent du type du document traité, ainsi que de l'application envisagée.

1.7.1.2. Regroupement Séquentiel

Les méthodes de regroupement hiérarchique ne prennent pas en considération le "voisinage" entre les segments, alors qu'il s'avère que la variabilité intra-locuteur entre deux segments appartenant à un même locuteur proches temporellement est moins forte qu'entre deux segments éloignés temporellement, appartenant à ce même locuteur. Pour cela, le principe du regroupement séquentiel est basé sur le traitement de ces segments séquentiellement (par ordre temporel). Donc, à la première itération, on considère que le premier élément (segment), forme la première classe. A chaque itération, on compare le segment prochain et voisin temporellement aux classes déjà existantes et il sera assigné au groupe d'éléments (segments) le plus proche au sens du critère de regroupement. En revanche, si cet élément est différent des groupes d'éléments déjà existants, une nouvelle classe est créée contenant cet élément. Ceci implique que le critère de regroupement soit contraint pour que de nouvelles classes puissent être créées (figure 1.13).

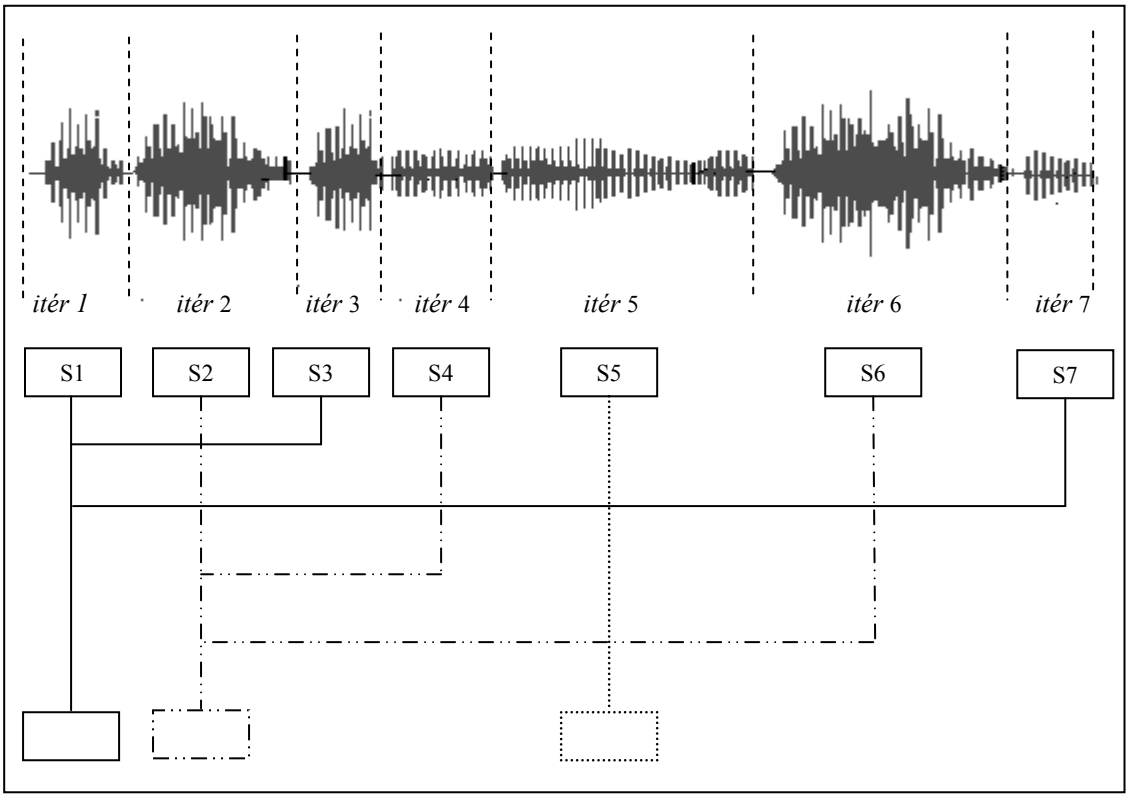

Figure 1.13 : Exemple de regroupement séquentiel.

Le critère d'arrêt dans ce cas de regroupement est plus simple, car le processus s'arrête une fois que tous les éléments (segments) ont été traités.

Ce type de regroupement peut convenir aux applications temps réel, où il est important de pouvoir traiter les segments au fur et à mesure et non d'effectuer le regroupement une fois que tous les segments sont collectés [7].

Dans le cadre de notre travail, deux expériences sont effectuées, une :

- première expérience qui consiste en une indexation avec connaissances préalable des locuteurs et du fait que le système dispose en mémoire des modèles des différents locuteurs à regrouper, le regroupement est réalisé par une simple identification de locuteur ;
- deuxième expérience qui consiste en une indexation sans connaissances a priori des locuteurs, dans laquelle nous avons opté pour la phase de regroupement séquentiel cité précédemment.

1.8. Conclusion

Dans ce chapitre, nous avons défini la reconnaissance automatique du locuteur avec ses différentes spécialités, notamment l'identification, la vérification, la détection du locuteur et l'indexation automatique des documents audio (avec ou sans connaissances a priori des locuteurs). Concernant cette dernière discipline, nous avons donné des explications assez détaillées sur ses deux étapes principales : il s'agit de la segmentation et du regroupement. Différentes techniques utilisées dans la littérature, relatives à chaque étape, ont aussi été présentées dans ce chapitre pour apporter plus de richesse dans ce domaine au lecteur.

Chapitre 2 :

PRINCIPALES TECHNIQUES UTILISEES EN INDEXATION DES DOCUMENTS AUDIO

2.1. Introduction

Dans ce chapitre, nous exposons les principales techniques utilisées durant les dernières années dans le domaine de l'indexation automatique des documents audio (i.e. segmentation et regroupement des segments en locuteurs) et des caractéristiques acoustiques. D'abord, un survol des caractéristiques qui ont montré leur convenance dans l'indexation par locuteurs est donné. Par la suite, nous introduisons les algorithmes et les systèmes généralement utilisés pour la tâche considérée.

La première section dans ce chapitre résume les caractéristiques qui ont prouvé leur bonne modélisation du locuteur (comme dans l'indexation par locuteurs). Nous nous concentrons sur les caractéristiques du locuteur qui permettent de mieux discriminer entre les locuteurs présents dans l'enregistrement et aident à les identifier. Dans la seconde section, nous présentons les principales techniques qui ont été élaborées en segmentation et en regroupement en locuteurs.

2.2. Domaines d'Application de l'Indexation de Documents Audio

L'indexation en locuteurs consiste en première étape à découper le flux audio en segments homogènes (chaque segment contient un seul locuteur à la fois) : il s'agit de la segmentation en locuteurs. En deuxième étape, regrouper les différents segments homogènes selon les locuteurs intervenant dans le document audio dans le but d'obtenir l'intervention globale de chaque locuteur dans le document : c'est le regroupement en locuteurs.

L'indexation répond généralement à la question : *"Qui parle? et Quand ?"* (figure 2.1).

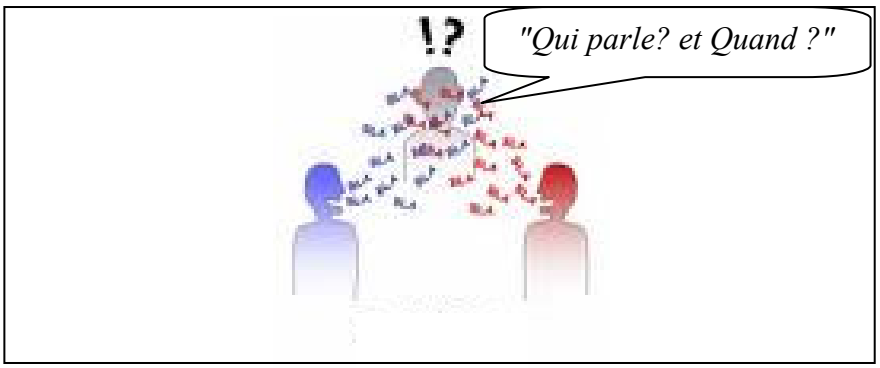

Figure 2.1 : Problème d'indexation.

Selon D. A. Reynolds et P. Torres-Carrasquillo [31], il existe 3 principaux domaines d'application de l'indexation en locuteurs qui ont attiré l'attention des chercheurs actuels, soient :

- l'archivage des informations TV et radio (Broadcast news) : les programmes de Radio et de TV, contenant souvent de la musique et des spots publicitaires, rendent cette tâche difficile ;
- le filtrage par locuteur des réunions, débats, conférences ou rencontres enregistrées (meetings) : où plusieurs personnes interviennent dans la même salle ou à travers le téléphone. Les enregistrements de ce type se font généralement avec plusieurs microphones. Le problème majeur des débats est dû à l'intervention de plusieurs locuteurs en même temps "cocktail party" [32] ;
- la sélection du message téléphonique d'un locuteur donné : les conversations téléphoniques entre deux locuteurs ou un message déposé sur un répondeur via canal téléphonique ont un intérêt particulier en indexation par locuteur. Leur problème majeur est que le signal téléphonique est très détérioré et ne préserve qu'une partie des caractéristiques du locuteur.

Les algorithmes de segmentation et de regroupement en locuteurs nécessitent une caractérisation du locuteur qui représente bien les données acoustiques et définissent une mesure / méthode de distance pour attribuer chaque vecteur caractéristiques à une classe (dans notre cas groupe).

De ce fait, l'utilisation de modèles acoustiques appropriés, une dimension et des algorithmes d'apprentissage optimaux permettent d'identifier correctement les différences acoustiques entre les locuteurs.

2.3. Paramétrisation Acoustique

Le processus de segmentation en locuteurs commence par l'étape de paramétrisation acoustique. Le signal n'est pas directement exploité.

Les caractéristiques acoustiques extraites du signal traduisent l'information concernant les locuteurs présents dans la conversation dans le but de permettre aux systèmes de les séparer.

Comme dans les systèmes de RAL et les systèmes de RAP, les caractéristiques les plus appropriées dans l'indexation en locuteurs sont les coefficients *Mel Frequency Cepstral Coefficients* (MFCC), *Linear Frequency Cepstral Coefficients* (LFCC), *Perceptual Linear Predictive* (PLP) et *Linear Predictive Coding* (LPC). Cependant la paramétrisation la plus employée dans le domaine de la parole est la représentation cepstrale, qui a l'intérêt de séparer l'excitation glottique et les résonances du conduit vocal. Par filtrage, seule la contribution du conduit vocal est conservée [25].

Bien que beaucoup de travaux sont effectués avec les coefficients cepstaux, nous avons opté pour les paramètres MFSC (*Mel Frequency Spectral Coefficients*) que nous allons utiliser pour modéliser les locuteurs.

2.3.1. Vecteurs Acoustiques

Le processus de paramétrisation effectue une analyse temps / fréquence à court terme du signal et rend un ensemble de vecteurs acoustiques.

Deux types de coefficients cepstaux sont retenus, les :

- vecteurs acoustiques issus d'une analyse en banc de filtres à échelle linéaire appelés LFCC ;
- coefficients obtenus à partir d'une échelle de Mel appelés MFCC [25].

2.3.2. Suppression des Informations sur le Canal de Transmission

Les coefficients cepstraux contiennent, en plus des caractéristiques de l'appareil phonatoire, des informations sur le canal de transmission. Suivant la tâche envisagée, ces informations causent des problèmes pour des documents contenant différents canaux de transmission (conversation téléphonique). Le système risque de reconnaître autant le canal de transmission que le locuteur [25].

Des méthodes de compensation sont appliquées sur les paramètres pour atténuer les distorsions engendrées par le canal de transmission [24], [34]. La méthode la plus fréquemment employée est la suppression de la moyenne cepstrale (*Cepstral Mean Subtraction*, CMS), calculée a posteriori sur les vecteurs ou à l'aide d'une fenêtre glissante.

2.3.3. Coefficients Différentiels et Energie

Les vecteurs acoustiques sont complétés, si nécessaire, par les dérivées premières et secondes des vecteurs [35]. Ces coefficients différentiels introduisent des informations dynamiques modélisant l'évolution des coefficients cepstraux, qualifiés alors de "statique". Les dérivées premières et secondes sont communément nommées "Delta" et "Delta Delta". L'énergie du signal et les dérivées de l'énergie sont aussi des coefficients utilisés [25].

Nous citons quelques travaux de recherche concernant la segmentation en locuteurs effectués en utilisant ces coefficients. D. Moraru a proposé deux systèmes utilisant dans le premier 20 coefficients LFCC plus l'énergie [36], D. A. Reynolds a utilisé pour son système 24 MFCC [37], tandis que dans A. Adami, le locuteur est caractérisé par 24 *Line Spectral Pairs* (LSP) [38].

Dans le but d'éviter l'influence des bruits de fond et d'autres évènements indépendants du locuteur, des techniques de "*feature warping*" ont été proposées pour changer la forme de la densité de probabilité des caractéristiques en une forme gaussienne avant leur modélisation.

Elles ont été appliquées avec succès pour l'indexation en locuteurs dans respectivement, les informations télédiffusées et les meetings [24], [39] et [40].

Dans le domaine de la détection parole / silence (*Silence Activity Detection*, SAD), plusieurs caractéristiques ont aussi été proposées pendant les dernières années. Quelques caractéristiques récentes ont été proposées (basées sur l'autocorrélation du signal ou les caractéristiques spectrales) [41].

Dans certains travaux, les systèmes d'indexation sont proposés en construisant un espace du locuteur à partir des données et en projetant les vecteurs caractéristiques en cet espace avant l'étape de regroupement [42]. De la même façon, d'autres travaux proposent une technique utilisant des modèles d'ancrage (introduite par D. Sturim, D. A. Reynolds, E. Singer et J.P.Campbell [43]) où les paramètres acoustiques sont projetés dans un espace des modèles d'ancrage (déjà défini à partir de données extérieures). La tâche du suivi du locuteur est exécutée ensuite en utilisant le vecteur paramètre résultant. Les résultats montrent que cette technique a bien amélioré la robustesse contre les parasites d'interférence extérieurs [44].

Quand les enregistrements sont collectés par plus d'un microphone (par exemple dans les meetings) il est utile dans ce cas d'utiliser les retards temporels entre les microphones [45].

Finalement, des caractéristiques de la source vocale pour la tâche de segmentation en locuteurs sont utilisées, un algorithme temps réel à 2 étapes est proposé par l'utilisation d'une fusion bayésienne des caractéristiques : LSP, MFCC et le pitch [46].

2.4. Segmentation en Locuteurs

La segmentation en locuteurs est définie comme étant la détection de changement de locuteurs et elle est étroitement liée à la détection de changement des caractéristiques acoustiques. Pour un flux audio donné, les systèmes de segmentation en locuteurs (détection des changements de locuteurs) consistent à trouver les instants où il y a eu vraiment un changement de locuteur dans le document audio.

La détection de changements acoustiques consiste à trouver les moments où il y a un changement des caractéristiques acoustiques dans l'enregistrement, qui peut s'agir d'un changement parole / non parole, musique / parole ou autres.

Bien que le terme "segmentation en locuteurs" est, quelquefois, utilisé à la place de l'indexation en locuteurs pour les systèmes exécutant les deux tâches de segmentation et de regroupement des segments. Comme nous allons le voir plus tard, plusieurs systèmes obtiennent l'indexation en locuteurs par le moyen de la segmentation en locuteurs en première étape et puis le regroupement des segments similaires en deuxième étape.

Dans d'autres systèmes, cette distinction n'est pas claire du fait que la segmentation et le regroupement en locuteurs sont mélangés.

Dans cette thèse, le système d'indexation proposée repose sur deux tâches : la segmentation du document en segments homogènes (contenant chacun un seul locuteur à la fois) et le regroupement de ces segments en groupes homogènes contenant chacun l'intervention globale de chaque locuteur dans l'enregistrement.

Deux types de systèmes de segmentation en locuteurs peuvent être trouvés dans la littérature. Le premier type est le système qui exécute la segmentation en une seule phase à partir de laquelle les points de changements sont obtenus. Le deuxième type de systèmes utilise un traitement à plusieurs phases, raffinant ainsi la décision de la détection de ces points de changements lors des itérations successives.

Ce second type de systèmes utilise des algorithmes à deux phases : dans la première phase, beaucoup de points de changements sont suggérés (avec un grand taux de fausses alarmes) et dans la deuxième phase, ces changements sont réévalués et certains sont supprimés. Aussi, une partie des systèmes du deuxième type utilise un traitement itératif convergeant en une segmentation optimale.

A un autre niveau, une classification générale des méthodes destinées à la segmentation en locuteurs seront citées dans cette section pour décrire les différents algorithmes concernés. Trois types d'algorithmes sont définis : algorithmes basées sur une métrique donnée, le silence et le modèle [47] et [48].

2.4.1. Segmentation Basée sur la Métrique

La segmentation basée sur la métrique est probablement la technique la plus utilisée jusqu'à ce jour. Elle est basée sur le calcul de distance entre deux segments acoustiques dans le but de déterminer s'ils appartiennent au même locuteur ou à des locuteurs différents, et par conséquent s'il existe ou pas un changement de locuteur au point du document audio analysé. Les deux segments sont d'habitude voisins (avec recouvrement ou pas) et le point de changement considéré est compris entre les deux.

La plupart des distances utilisées pour la détection de changements acoustiques peuvent aussi être appliquées au regroupement en locuteurs dans le but de comparer si deux groupes de locuteurs appartiennent ou pas au même locuteur.

Considérons deux segments audio "i" et "j" dont les vecteurs acoustiques sont respectivement X_i et X_j de longueurs N_i et N_j, et dont la moyenne et la variance sont notées par μ_i, σ_i pour le segment i et μ_j, σ_j pour le segment j. Les segments sont modélisés en utilisant les processus Gaussiens $M_i(\mu_i, \sigma_i)$ et $M_j(\mu_j, \sigma_j)$ (respectivement), qui peuvent comprendre une seule gaussienne ou un mélange de gaussiennes (*Gaussian Mixture Model,* GMM). D'autre part, le regroupement de ces deux segments est présenté par le vecteur acoustique final X, avec une moyenne μ, une variance σ, et un processus gaussien correspondant $M(\mu, \sigma)$.

En général, il y a deux types de distances qui peuvent être utilisées entre n'importe quelle paire de tels segments audio. Le premier type compare les statistiques des deux vecteurs

acoustiques sans considérer aucun modèle appliqué au préalable sur les données, et qui est appelé distance basée sur les statistiques. Ces distances sont normalement très rapides à calculer et donnent de bonnes performances. Le second type de distances est basé sur l'évaluation de la vraisemblance des données selon les modèles qui la présentent. Ces distances sont lentes à calculer (puisque les modèles nécessitent un apprentissage et une évaluation) mais peuvent donner de meilleurs résultats que les distances basées sur les statistiques. Ces distances sont appelées les techniques basées sur la vraisemblance.

2.4.1.1. Critère d'Information Bayésien (BIC)

Le critère de BIC est probablement la métrique la plus utilisée en segmentation et en regroupement en locuteurs vues sa simplicité et son efficacité. Le critère BIC est un critère de vraisemblance pénalisé par la complexité du modèle (i.e. le nombre de paramètres du modèle), introduit par G. Schwarz [49], comme un critère de sélection du modèle [7].

Soit X_i le vecteur de données à modéliser et M_i le modèle paramétrique envisagé. La fonction de vraisemblance $L(X_i, M_i)$ est alors maximisée pour le modèle.

Si m désigne le nombre de paramètres du modèle. Le critère BIC pour le modèle M_i est alors défini par :

$$BIC(M_i) = \log L(X_i, M_i) - \lambda \frac{m}{2} \log(N_i) \tag{2.1}$$

Le premier terme reflète l'ajustement du modèle aux données et le deuxième terme correspond à la complexité du modèle. λ est un poids de pénalité, en théorie, il est égal à 1.

Le critère BIC permet de sélectionner un modèle parmi plusieurs modèles pour les mêmes données : c'est le modèle qui maximise ce critère, donc il correspond le plus aux données en terme de vraisemblance et dont la complexité reste raisonnable.

Dans le but d'utiliser le critère de BIC pour évaluer s'il existe un point de changement entre deux segments, celui-ci évalue l'hypothèse que :

- X modélise mieux les données ;
- contre l'hypothèse que : (X_i+X_j) le font à la place,

et ceci, comme dans la métrique de rapport de vraisemblance généralisé (*Generalize Likelihood Ratio*, GLR), en calculant l'expression :

$$\Delta BIC(i, j) = -R(i, j) + \lambda P \tag{2.2}$$

Le terme $R(i, j)$ peut être écrit dans le cas des modèles composés d'une seule gaussienne comme suit :

$$R(i, j) = \frac{N}{2} \log\left|\Sigma_X\right| - \frac{N_i}{2} \log\left|\Sigma_{X_i}\right| - \frac{N_j}{2} \log\left|\Sigma_{X_j}\right| \tag{2.3}$$

où Σ : désigne matrice de covariance.

Où P est le terme de pénalité, qui est une fonction du nombre de paramètres du modèle. Pour une matrice de covariance entière, il est donné par :

$$P = \frac{1}{2}\left(p + \frac{1}{2} p(p+1) \right) \log(N) \tag{2.4}$$

p est la dimension de l'espace des vecteurs acoustiques.

Le terme de pénalité compte l'augmentation de la vraisemblance des grands modèles contre les petits modèles.

Bien qu'il soit absent dans la formulation originale, le paramètre λ est introduit pour ajuster l'effet du terme de pénalité dans la comparaison, qui constitue un seuil caché pour la différence BIC. Un tel seuil nécessite d'être ajusté aux données. Plusieurs chercheurs proposent des méthodes pour la sélection automatique de λ [7], [50] et [51].

Plusieurs implémentations utilisant le BIC comme métrique de segmentation ont été proposées :

Au départ, un algorithme de détection multiple de points de changement en deux phases, est proposé [52], et après plusieurs travaux ont suivi avec des algorithmes à une ou deux phases [53] et [54]. Ils ont tous proposé un système utilisant une fenêtre croissante avec des segments d'analyse à longueurs variables pour trouver itérativement les points de changements. Dans d'autres travaux, quelques méthodes sont proposées pour rendre l'algorithme plus rapide et permettant de détecter les plus petits changements de locuteurs [55].

Même avec les différents efforts pour accélérer le traitement du BIC, ce dernier reste plus coûteux en terme de temps de calcul que les autres métriques statistiques dans le cas d'analyse du signal avec une grande résolution, mais ses bonnes performances lui ont permis d'être l'algorithme choisi dans beaucoup d'applications.

C'est pourquoi, plusieurs chercheurs utilisent le BIC dans la deuxième phase (raffinement) pour les systèmes de segmentation en locuteurs à 2 phases. Dans cette direction, P. Delacourt et C. J. Wellekens proposent le système DISTBIC (Speaker based Segmentation for Audio Data Indexing), où le GLR est utilisé dans la première phase de l'algorithme de segmentation [56]

Dans d'autres travaux, un algorithme à deux phases utilisant le critère BIC dans les deux phases, est proposé. Dans la première phase l'algorithme BIC essaye de minimiser les fausses alarmes et dans la seconde phase, le BIC est utilisé pour trouver le reste des points de changements de locuteurs non détectés lors de la première phase [57], [58].

2.4.1.2. Rapport de Vraisemblance Généralisé (*Generalized Likelihood Ratio*, GLR)

Le GLR (proposé pour la détection de changements par A.S. Willsky et H. L. Jones [59], puis U. Appel et A. Brandt [60]) est une métrique basée sur la vraisemblance qui propose un

rapport entre deux hypothèses : d'une part, H_0 (hypothèse 0) considère que les deux segments sont prononcés par le même locuteur, par conséquent $X = X_i \cup X_j \sim M(\mu,\sigma)$ représente mieux les données. Et d'autre part, H_1 (hypothèse 1) considère que chaque segment est prononcé par un locuteur différent, par conséquent $X_i \sim M_i(\mu_i,\sigma_i)$ et $X_j \sim M_j(\mu_j,\sigma_j)$ ensemble, conviennent mieux aux données. Le test du rapport est calculé comme un rapport de vraisemblance entre les deux hypothèses :

$$GLR(i,j) = \frac{H_0}{H_1} = \frac{L(X, M(\mu,\sigma))}{L(X_i, M_i(\mu_i,\sigma_i))L(X_j, M_j(\mu_j,\sigma_j))} \tag{2.5}$$

La distance $D(i, j) = -log(GLR(i, j))$ décide, alors, si les deux segments appartiennent au même locuteur ou pas.

Les densités de probabilité dans le GLR sont inconnues et doivent être estimées directement des données sans considération d'aucun segment. Dans la segmentation en locuteurs le GLR est d'habitude utilisé avec deux segments adjacents de même taille, et le seuil est soit fixé au préalable ou adapté dynamiquement.

Le GLR a été utilisé pour segmenter le signal en locuteurs en une seule phase de traitement pour la tâche de suivi du locuteur. Le seuil est mis de telle sorte que les erreurs de détections manquées soient minimisées (impliquant un taux de Fausses Alarmes élevé) [12].

En outre, dans un système de segmentation en deux locuteurs, exécuté en deux phases, le GLR est utilisé dans la première phase pour une sur segmentation des données [61].

Dans la même tâche (détection de deux locuteurs), la première seconde est considérée prononcée par le premier locuteur et le deuxième locuteur est trouvé en déterminant les points de changements via l'algorithme GLR [38].

Probablement, l'algorithme le plus représentatif de l'utilisation de la métrique GLR pour la segmentation en locuteurs est le DISTBIC [7], où la métrique GLR est proposée comme une première étape dans le processus de segmentation à deux étapes (utilisant le BIC comme seconde métrique). La distance GLR n'est pas utilisée seule, un faible filtrage adaptatif est appliqué dans le but de réduire les ondulations dans la fonction de distance calculée (qui générerait des faux points maximum / minimum) et donc la différence entre chaque maximum et le minimum adjacent est utilisée pour affirmer les points de changements [58].

2.4.1.3. Distance de Gish

C'est une métrique basée sur la vraisemblance, obtenue comme une variante de la distance GLR [62]. La fonction GLR est divisée en deux parties (λ_{cov} et λ_{mean}) et la partie dépendante de l'environnement est ignorée, conduisant à l'équation suivante :

$$D_{Gish}(i,j) = -\frac{N}{2}\log\left(\frac{|S_i|^{\alpha}|S_2|^{(1-\alpha)}}{|W|}\right) \tag{2.6}$$

où *Si* et *Sj* représentent les matrices de covariance de chaque segment, $\alpha = \frac{N_1}{N_1 + N_2}$ et *W* est leur moyenne pondérée :

$$W = \frac{N_1}{N_1 + N_2} S_1 + \frac{N_2}{N_1 + N_2} S_2 \tag{2.7}$$

Dans T. Kemp, la distance de Gish est comparée à d'autres techniques pour la tâche de segmentation en locuteurs [63].

2.4.1.4. Distance de Kullback-Leibler (KL ou KL2)

Les distances KL et KL2 [64], [65] sont bien utilisées vus leur rapidité de calcul et leurs résultats acceptables.

Soientt deux variables aléatoires *X* et *Y*, la distance KL (appelée aussi divergence) est définie par :

$$KL(X,Y) = E_X\left(\log \frac{P_X}{P_Y}\right) \tag{2.8}$$

où E_X est la valeur prévue par rapport à la densité de probabilité de *X*.

Quand les deux variables sont gussiennes, on peut obtenir une forme de solution proche d'une telle expression [66] :

$$KL(X,Y) = \frac{1}{2} tr\left[(\Sigma_X - \Sigma_Y)(\Sigma_Y^{-1} - \Sigma_X^{-1})\right] + \frac{1}{2} tr\left[(\Sigma_Y^{-1} - \Sigma_X^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T\right] \tag{2.9}$$

Où *tr* : désigne trace et *T* : désigne la transposée.

La distance KL2 peut être obtenue en symétrisant la distance KL de la manière suivante:

$$KL2(X, Y) = KL(X, Y) + KL(Y, X) \tag{2.10}$$

Comme précédemment, si les deux variables aléatoires *X* et *Y* sont considérées gaussiennes, on peut obtenir une forme de solution proche pour KL2 en fonction de leurs matrices de covariances et leurs moyennes.

Soient deux segments acoustiques quelconques X_1 et X_2 pouvant être considérés statistiquement comme *X* et *Y*, alors nous pourrons obtenir la distance entre eux en utilisant simplement ces distances.

Dans P. Delacourt et C. J. Wellekens, la distance KL2 est considérée comme une première étape dans un système de détection de changement de locuteurs à deux étapes [7].

2.4.1.5. Distance de Divergence de Forme (*Divergence Shape Distance,* DSD)

D'une manière très similaire à la distance de Gish [62], le DSD est dérivé de la distance KL en éliminant la partie affectée par la moyenne, du fait que la moyenne est dépendante des conditions de l'environnement. Par conséquent, la distance DSD correspond à l'expression :

$$D(i,j) = \frac{1}{2} tr\left[\left(\Sigma_i - \Sigma_j\right)\left(\Sigma_j^{-1} - \Sigma_i^{-1}\right)\right] \quad (2.11)$$

Le DSD est utilisé dans un algorithme à une seule étape et ses résultats sont comparés au critère de BIC [67]. Il est aussi utilisé comme une première étape dans un système de segmentation à deux étapes, utilisant le BIC dans la phase de raffinement [68].

2.4.1.6. Autres Distances

Beaucoup d'autres techniques peuvent définir une distance entre deux ensembles de caractéristiques acoustiques ou deux modèles. Certaines d'entre elles ont été appliquées à la tâche de segmentation en locuteurs. D'autres ont été appliquées avec succès pour la tâche d'identification du locuteur, comme celle que nous proposons d'utiliser dans notre étude. Il s'agit des mesures statistiques du second ordre (*Second Order Statistical Measures*, SOSM), ces dernières sont basées sur le calcul de la matrice de covariance et du vecteur moyenne caractérisant chaque segment de parole. La théorie des différentes mesures SOSM est explicitée en détail au paragraphe 3.2.

Toutes ces techniques, basées sur les métriques, calculent une fonction dont les maximums/ minimums nécessitent d'être comparés à un seuil pour déterminer les points de changements. Dans plusieurs cas, le seuil est déterminé empiriquement selon les données traitées (le seuil est dépendant des données). Le seuil demande à être recalculé à chaque fois que la nature des données change. Ce problème a été étudié dans le cadre de l'identification du locuteur [66].

2.4.2. Segmentation non Basée sur des Métriques

Dans cette section les deux autres classes de la segmentation en locuteurs sont critiquées, notamment : les techniques basées sur les silences et sur le modèle.

2.4.2.1. Segmentation Basée sur les Silences

Ces techniques détectent les changements de locuteurs mettant une hypothèse que la plupart des changements entre les locuteurs sont intercalés par des segments de silence. Les systèmes de cette catégorie sont basés sur l'énergie. Un silence (s'il n'est pas trop bruité) étant caractérisé par un faible niveau d'énergie, la détection de silences repose en général sur le calcul de l'énergie du signal.

Un seuil est habituellement utilisé pour les déterminer [63] et [69].

2.4.2.2. Segmentation Basée sur le Modèle

Les premiers modèles (par exemple les GMM) sont créés pour un ensemble proche des classes acoustiques en utilisant les données d'apprentissage.

Dans L. Lu, S. Z. Li et H. -J. Zhang [70] les Machines à Support de Vecteurs (SVM) sont utilisés comme un classifieur à la place des modèles GMM [58].

Parmi les méthodes présentées, celle qui a retenu notre attention, est la segmentation basée sur la recherche des points de changements de locuteurs appelés aussi points de rupture et utilisant la comparaison entre les caractéristiques acoustiques des locuteurs. La détection de ces points est effectuée une fois par le calcul d'une distance statistique (SOSM) et une autre fois, en utilisant un MLP et un SVM. Elle présente l'intérêt particulier de fonctionner avec un système de discrimination entre deux segments de parole. Ce système est appliqué tout le long du flux audio, quel que soit le type du signal audio. Les techniques de segmentation basées sur des mesures de distance entre portions de signal peuvent a priori améliorer la détection des changements de locuteurs et de la rendre systématique pour tous les types de signaux audio.

2.5. Regroupement en Locuteurs

On peut différencier entre deux types de systèmes : les systèmes *off-line* (opérant en temps différé) et les systèmes *on-line* (opérant en temps réel). Les systèmes *off-line* ont l'accès à tout l'enregistrement avant de commencer son traitement. Ces systèmes sont les plus rencontrés dans la bibliographie. Les systèmes *on-line* ont l'accès seulement aux données enregistrées jusqu'à l'instant de traitement. D'habitude, de tels systèmes, commencent par un seul locuteur (celui qui commence la parole au début de l'enregistrement) et augmente itérativement le nombre de locuteurs au fur et à mesure qu'ils interviennent.

Nous citons quelques systèmes existants, traitant les données *on-line*, un :

- algorithme de regroupement basé sur la mesure de distorsion à VQ est proposé pour cette tâche [71]. Ils commencent le traitement avec un locuteur dans le dictionnaire de références et additionnent progressivement de nouveaux locuteurs pour lesquels la distorsion VQ dépasse la valeur du seuil du dictionnaire de références [72] ;
- système basé sur les GMM est proposé, utilisant une distance KL entre les modèles. Les points de changements sont détectés quand la parole devient disponible et les données sont attribuées soit à un locuteur présent dans la base de données ou bien, alors, un nouveau locuteur est créé selon un seuil dynamique [73].

Tous les systèmes présentés précédemment sont basés sur un traitement *on-line*, bien que quelques techniques peuvent potentiellement être aussi utilisées dans une implémentation *off-line*. Ces systèmes peuvent être classés en deux principaux groupes, les :

- techniques de regroupement hiérarchique qui atteignent le regroupement optimal par un traitement itératif des différents nombres de groupes possibles, obtenus en mélangeant ou divisant les groupes existants ;
- autres techniques de regroupement qui estiment d'abord le nombre de groupes et qui obtiennent leur résultat sans dérivation à partir de grands ou petits groupes [58].

Dans ce qui suit, nous présentons des méthodes de regroupement existant dans la littérature. Nous nous intéressons aux techniques de regroupement hiérarchique par agglomération puis aux algorithmes de regroupement séquentiel.

2.5.1. Techniques de Regroupement Hiérarchique

La plupart des algorithmes de regroupement utilisent la structure hiérarchique, où les segments de parole ou les groupes sont itérativement mélangés ou divisés jusqu'à atteindre le nombre optimal de locuteurs. Les deux techniques couramment utilisées dans le regroupement en locuteurs sont :

- le regroupement ascendant qui commence avec un grand nombre de segments /groupes puis mélange ces groupes pour converger vers le nombre optimal de groupes (figure 2.2) ;
- dans un autre côté, les systèmes descendants commencent par un ou un nombre petit de groupes et via des procédures de division ils obtiennent le nombre de groupes optimal (figure 2.2).

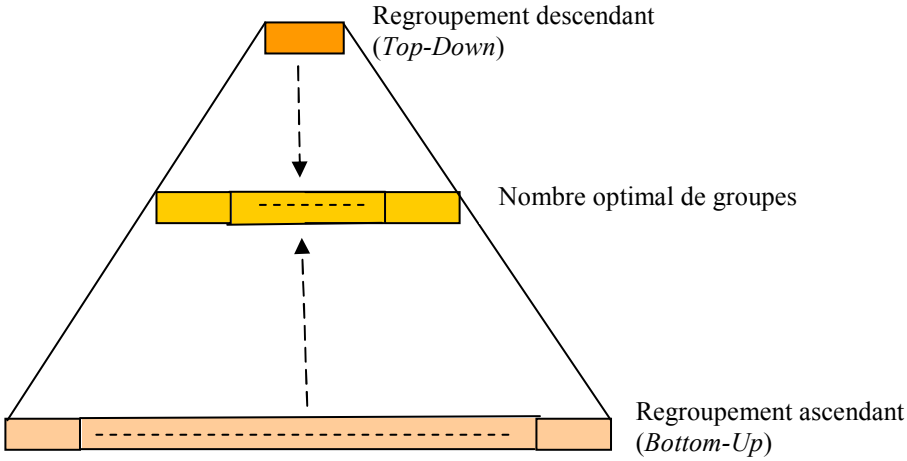

Figure 2.2 : Représentation graphique du regroupement ascendant et descendant.

Dans le design des deux systèmes, deux éléments nécessaires sont à définir :

- la distance entre les groupes / segments pour déterminer la similarité acoustique. Au lieu de définir une paire de valeurs, individuelle, d'habitude une matrice de distances est décrite, et qui est constituée des distances entre n'importe quelle paire possible. Dans plusieurs cas, les métriques de distance utilisées pour le regroupement en locuteurs r0essemble à ceux utilisées pour la segmentation en locuteurs calculées en utilisant les métriques présentées dans la section 2.4.1 ;
- un critère d'arrêt pour stopper le mixage / division itératif au nombre optimal des groupes.

Les techniques les plus représentatives décrites dans la littérature sont classées par leur type de regroupement [58] :

2.5.1.1. Techniques de Regroupement Ascendant

Actuellement, les techniques de regroupement ascendant sont les approches de regroupement en locuteurs les plus utilisées, en utilisant des techniques de segmentation pour

définir le point de départ du regroupement. Elles sont connues aussi sous le nom de regroupement agglomératif, bien qu'elles soient utilisées pendant des années dans la classification des formes [74].

Normalement, une matrice de distances entre tous les groupes (distance de chaque groupe avec tous les autres groupes) est calculée et la plus proche paire est mixée itérativement jusqu'à atteinte un critère d'arrêt.

Une des premières recherches faites sur le regroupement en locuteurs pour la RAP, utilisant la distance de Gish [62] comme matrice des distances, avec un poids pour favoriser le mixage des voisins [75].

Dans le même contexte, la distance KL2 est utilisée comme une métrique de distance et le critère d'arrêt est déterminé avec un seuil de mixage. Ils montrent l'efficacité de la distance KL2 pour le regroupement en locuteurs [64]. Aussi dans un autre travail, la métrique KL2 est utilisée comme une distance de groupe. Dans ce travail, une séparation des segments de parole en féminin / masculin est exécutée en premier, ensuite, le regroupement est exécuté sur chaque groupe indépendamment ; ceci réduit le calcul (le nombre de combinaisons de paires de groupes est plus petit) et donne de meilleurs résultats [76].

D'autres recherches sont faites en utilisant la distance GLR pour un regroupement pilote contrôleur [77] et pour l'indexation de meetings (utilisant le BIC comme critère d'arrêt) [78].

L'une des distances et critères d'arrêt les plus utilisés est le critère de BIC, qui était initialement déjà proposé pour le regroupement [79]. La matrice de distance est calculée pour chaque itération et la paire dont la valeur du ΔBIC est la plus grande est mixée. Le processus s'achève quand toutes les paires ont un $\Delta BIC < 0$.

Ces systèmes utilisent, alors, le regroupement agglomératif standard via le BIC, avec une valeur de pénalité λ posée pour obtenir plus de groupes que l'optimal. Dans la partie indexation par locuteurs, ils classent premièrement chaque groupe par sexe et par bande passante (BP) (dans les informations télédiffusées)

2.5.1.2. Techniques de Regroupement Descendant

Dans la littérature actuelle, il y a peu de travaux où on commence par un seul groupe et le divise itérativement jusqu'à atteindre les segments optimaux par rapport à un critère d'arrêt donné. Cette technique s'appelle Technique de Regroupement Descendant.

Dans certains travaux, une méthode de regroupement descendant est proposée pour le regroupement en locuteurs [80]. Dans d'autres, il est appliqué pour l'indexation par locuteurs [6] [81]. L'algorithme divise les données itérativement en quatre sous groupes et permet le réassemblage des groupes qui sont très similaires.

2.5.1.3. Combinaison des Méthodes de Regroupement

Bien que le regroupement ascendant soit plus populaire que le regroupement descendant, il n'est pas clair quel est celui qui donne les meilleurs résultats et dans quelles conditions. Dans le thème de la transcription des informations télédiffusées, deux techniques sont comparées [82]. D'un côté, le regroupement ascendant utilise une mesure de distance dite "*divergence-like*" et un comptage de groupes comme critère d'arrêt. D'un autre côté, le regroupement descendant utilise une distance sphérique harmonique arithmétique et aussi un comptage de groupes comme critère d'arrêt.

Etant donné que les techniques de regroupement ascendant et descendant peuvent éventuellement être complémentaires, certains ont proposé des systèmes qui peuvent combiner des systèmes multiples, tout en obtenant une indexation en locuteurs améliorée.

Un algorithme de vote de groupes est présenté pour permettre une amélioration de la sortie de l'indexation en regroupant deux systèmes d'indexation en locuteurs différents. Les tests sont procédés en utilisant deux systèmes ascendants et deux autres descendants [83].

D'autre part, deux approches de combinaison différentes sont présentées pour combiner une sortie ascendante et une sortie descendante et sont appliquées pour le traitement des informations télédiffusées et des meetings. La première technique, appelée hybridation, propose un premier système comme initialisation au deuxième système. La seconde technique est appelée Fusion et procède à la recherche des meilleurs segments résultants communs, et une re-segmentation des données pour attribuer les segments non communs [84], [85], [58].

2.5.2. Regroupement Séquentiel

Dans cette section, nous présentons deux méthodes de regroupement séquentiel. Nous en rappelons brièvement le principe :

Le premier segment forme une première classe de locuteur. Les segments suivants sont examinés au fur et à mesure qu'ils sont détectés. Pour chaque segment, un critère de regroupement contraint est calculé par rapport à chaque classe de locuteur déjà existante. Si le critère est vérifié pour l'une des classes et qu'il est optimum par rapport à l'ensemble des classes, alors le segment est ajouté au groupe de segments correspondant. A l'inverse, si le critère n'est vérifié pour aucune des classes alors un nouveau groupe de segments est créé avec ce segment.

2.5.2.1. Utilisation de Sous-Espace Propre du Locuteur (SEP)

Le critère de regroupement s'appuie sur des techniques de vérification du locuteur. Chaque groupe de segments créé est modélisé par un Sous-Espace Propre (SEP).

Soit $\left\{x_t^{(i)}\right\}$ la séquence de vecteurs acoustiques correspondant aux segments contenus dans la classe du locuteur *i*, $\left\{x_t^{(i)}\right\}$ correspond à la concaténation des séquences de vecteurs acoustiques formant chaque segment du groupe *i*. Le SEP du locuteur *i* est calculé de la

manière suivante : soit la matrice $X^{(i)}$ dont les lignes sont formées par les vecteurs où $\mu^{(i)}$ est le vecteur moyenne des vecteurs acoustiques $x_t^{(i)}$ $(1 \le t \le M)$. La matrice $X^{(i)}$ est Décomposée en Valeurs Singulières (DVS) :

$$X^{(i)} = U^{(i)} \Sigma^{(i)} V^{(i)\,T} \tag{2.12}$$

$U^{(i)}$ et $V^{(i)}$ sont les matrices dont les colonnes sont respectivement les vecteurs propres de $X^{(i)} X^{(i)T}$ et $X^{(i)T} X^{(i)}$. $\Sigma^{(i)}$ est la matrice des valeurs singulières de $X^{(i)}$. Les vecteurs propres de la matrice de corrélation de $X^{(i)T} X^{(i)}$ sont les vecteurs de base des données de parole $X^{(i)}$;

Si r est le nombre de valeurs singulières sélectionnées dans $\Sigma^{(i)}$, alors $V^{(i)}$ devient une matrice $N \times r$ formée avec les vecteurs $\left\{v_1^{(i)} ... v_r^{(i)}\right\}$, et caractérise le SEP du locuteur i.

Pour définir le critère de regroupement associé au locuteur i, la distance d'un vecteur acoustique x_t au SEP i est calculée comme suit :

$$dist\left(V^{(i)}, x_t\right) = \left\| x_t - \left\{ \sum_{j=1}^{r} \left(\left(x_t - \mu^{(i)}\right)^T v_j^{(i)} \right) v_j^{(i)} + \mu^{(i)} \right\} \right\|^2 \tag{2.13}$$

Le critère de regroupement du SEP i pour un segment $X^{(k)}$ composé de la séquence de vecteurs acoustiques $x_t^{(k)}$ $(1 \le t \le N)$ est alors défini comme la moyenne des distances des vecteurs acoustiques au SEP :

$$dist\left(V^{(i)}, X^{(k)}\right) = \frac{1}{N} \sum_i dist\left(V^{(i)}, x_t^{(k)}\right) \tag{2.14}$$

Quand un nouveau segment est ajouté au SEP alors le SEP est mis à jour : une nouvelle DVS est effectuée sur tous les vecteurs acoustiques composant le groupe de segments correspondant.

Cette méthode permet de segmenter et indexer un journal télévisé par locuteurs en temps réel [26], [27]. La segmentation repose sur une détection de silences. Les segments de parole correspondent alors à une section de parole entre deux silences. L'hypothèse que les locuteurs soient séparés par des silences significatifs (i.e. facilement détectables) est implicitement faite [7].

Des expériences ont été menées sur 150 mn de journaux télévisés NHK. Le but est d'extraire les paroles du présentateur [26] et [27]. Pour évaluer le processus d'indexation, les auteurs définissent deux taux :

$$\mathrm{TEx} = \frac{\mathrm{SPCI}}{\mathrm{TSP}} \tag{2.15}$$

$$\mathrm{TPr} = \frac{\mathrm{SPCI}}{\mathrm{SIP}} \tag{2.16}$$

avec :

TEx : Taux d'Extraction ;

SPCI : nombre de Segments du Présentateur Correctement Identifiés ;

TSP : nombre Total de Segments du présentateur ;

TPr : Taux de Précision ;

SIP : nombre de Segments Identifiés comme Présentateur.

Le taux d'extraction obtenu est de 93.4% et le taux de précision est de 98.7%. Les auteurs signalent cependant qu'il est préférable d'avoir des segments assez longs, en particulier le premier, sinon l'indexation échoue. Ceci est un des points faibles de la méthode : les SEP ne sont pas représentatifs, étant donné le faible nombre de vecteurs utilisés. Par ailleurs, le seuil d'acceptabilité d'un nouveau segment dans un SEP est sensible à la longueur des segments.

D'autres expériences sont menées sur des débats télévisés ou des téléfilms, et dans les deux cas, les résultats se dégradent de manière significative. De plus, les seuils sont choisis de manière complètement empiriques et semblent peu robustes [7].

2.5.2.2. Utilisation du BIC (*Bayesian Information Criterion*)

Les segments sont traités au fur et à mesure de leur formation pour répondre à des contraintes de temps réel. Des gains de regroupement sont calculés pour tous les couples de segments (i, j) possibles. Ce gain est une différence de critère d'information Bayésien.

En théorie, cet algorithme *on-line* est sous-optimal comparé à l'algorithme *off-line*. En effet, les maxima dans l'algorithme *on-line* sont locaux alors qu'ils sont globaux dans l'autre algorithme. Cependant, les réunions optimales de segments concernent des segments proches temporellement parlant. En pratique, il se trouve que l'algorithme *on-line* prend mieux en compte ces relations que l'algorithme *off-line*. Par ailleurs, les segments trop petits ne sont pas considérés ici : ils sont collectés dans un groupe de segments "poubelle" (en Anglais : *garbage*). Une comparaison des résultats est faite sur des données extraites de la base de données HUB-4 1997 [86].

Même si les deux algorithmes fournissent de bons résultats, l'algorithme *on-line* se montre plus performant que l'algorithme *off-line* en terme de rapidité, de pureté des groupes de segments (respectivement 98.58% contre 96.7%) et de nombre de groupes de segments [7].

2.5.3. Utilisation de l'Information de Support dans l'Indexation

Pour certaines applications, il est possible d'obtenir une amélioration de l'indexation en locuteurs en utilisant les informations non acoustiques. Dans cette section l'utilisation des transcriptions à partir de l'enregistrement et les techniques de retard temporel entre les canaux multi-microphones sont cités [58].

2.5.3.1. Aide de l'Indexation en Utilisant la Transcription de Parole

Une technique très intéressante pour améliorer l'indexation en locuteurs dans certaines conditions, est l'utilisation des transcriptions à partir du signal acoustique dans le but

d'extraire l'information qui peut aider l'attribution de chaque locuteur à chaque groupe. De telles transcriptions peuvent être obtenues via un système de RAP.

L'utilisation d'une telle information linguistique est étudiée pour le domaine des informations télédiffusées, où les personnes se présentent normalement et dialoguent avec d'autres locuteurs en les appelant par leurs noms [87] et [88]. Les auteurs proposent un ensemble de règles pour identifier le locuteur qui se présente, et les locuteurs qui le précèdent et qui parle après lui. Les règles sont appliquées aux changements de locuteurs générés par un système basé sur un décodeur qui est la sortie d'un système de RAP [58].

2.5.3.2. Indexation en Locuteurs Utilisant l'information Multi-Canaux

Une caractéristique exceptionnelle propre au domaine des meetings est la disponibilité de microphones multiples pour l'enregistrement. La différence temporelle entre les microphones peut être considérée comme une caractéristique pour identifier les locuteurs dans la salle par leurs positions, du fait que la parole prononcée par chaque locuteur prend des temps différents pour atteindre chacun des microphones, selon la position du locuteur dans la salle. Une telle caractéristique a deux inconvénients principaux à partir des caractéristiques acoustiques. D'un côté, c'est une source d'erreurs quand les locuteurs sont situés en symétrie par rapport aux microphones. D'un autre côté, ils deviennent moins malléables quand deux locuteurs se déplacent dans la salle, ce qui cause alors l'utilisation des algorithmes de suivi du locuteur.

Pour la tâche de segmentation en locuteurs, une approche de suivi de locuteur est proposée en utilisant seulement les différences de canaux [89]. La même approche est étendue au regroupement en locuteurs [90], [91] et [92].

Les retards entre les canaux ne peuvent pas donner de meilleurs résultats que les caractéristiques acoustiques. Toutefois, certains travaux on montré que la combinaison entre ces retards et les paramètres MFCC peut réellement améliorer le regroupement [93], [45] et [58].

2.6. Conclusion

L'indexation de documents audio en locuteurs fait appel à deux disciplines : la segmentation et le regroupement en locuteurs. Plusieurs travaux relatifs à ces deux disciplines sont effectués utilisant différentes paramétrisations du signal de parole, ainsi que différentes métriques et différents modèles. Pour cela, nous avons présenté dans ce chapitre les principales techniques utilisées dans la littérature pour la segmentation et le regroupement en locuteurs. En premier, nous avons présenté les principales paramétrisations utilsées pour la modélisation des locuteurs. Ensuite, un certain nombre de travaux de segmentation sont résumés, commençant par la segmentation basée sur des métriques, où nous avons expliqué les métriques les plus utilisées dans ce domaine, comme le critère de BIC et la distance de Kullback-Leibler. Nous avons, après cela, exposé des techniques de segmentation qui ne sont pas basées sur les métriques, en l'occurrence : la segmentation basée sur les silences et sur le

modèle du locuteur. De la même manière, le regroupement en locuteurs est présenté avec ses différents types : hiérarchique et séquentiel. Quelques recherches liées à chaque type de regroupement sont citées permettant de mieux comprendre ce dernier.

Nous avons remarqué que la paramétrisation MFCC est très utilisée pour la modélisation des locuteurs. D'autre part, le BIC et les GMM sont très employés dans les techniques de segmentation. Ceci est confirmé par le grand nombre de travaux utilisant ces techniques. Concernant le regroupement, plusieurs chercheurs optent pour le regroupement hiérarchique ascendant, vue son efficacité. Cependant, le critère d'arrêt du regroupement est déterminé d'une manière empirique dans la plupart des travaux et il reste dépendant de la nature de la BD traitée.

Nous avons aussi vu l'indexation *off-line* et l'indexation *on-line*, et comment le regroupement hiérarchique semble mieux adapté pour la première, tandis que le regroupement séquentiel est bien adapté à la deuxième. Ce dernier prend en considération, les relations de voisinage entre les segments.

Chapitre 3 :

METHODES PROPOSEES POUR L'INDEXATION

3.1. Introduction

Ce chapitre présente les différentes méthodes proposées dans le cas de l'indexation avec connaissances du locuteur (modèles des locuteurs disponibles) et dans le cas de l'indexation sans connaissances a priori du locuteur (indexation aveugle). Premièrement, nous définissons les différents classifieurs utilisés, notamment : la mesure statistique μ_{Gc}, le multi-layer perceptron (MLP) et les machines à vecteurs de support (SVM). Après, nous présentons une caractéristique du locuteur réduite que nous avons développée pour réduire les entrées des deux classifieurs MLP et SVM. Deuxièmement, nous détaillons les différentes architectures utilisées pour fusionner les classifieurs précédents. En dernier, nous donnons les différents algorithmes correspondant aux deux cas d'indexation (avec et sans connaissances a priori du locuteur).

Dans le cas de l'indexation du locuteur avec connaissances des locuteurs, le classifieur utilisé est le classifieur statistique, vu que nous possédons les modèles des locuteurs. Tandis que dans le cas de l'indexation aveugle, les trois différents classifieurs (détaillés ci-dessous) sont utilisés ainsi que certaines techniques de fusion.

3.2. Classifieur Mono-Gaussien ou "*Second Order Statistical Measures*" (SOSM)

Les mesures statistiques du second ordre sont des mesures basées sur le calcul de la moyenne et de la matrice de covariance à partir d'un modèle gaussien de chaque locuteur.

3.2.1. Propriétés du Modèle Gaussien

Soit $\{x_t\}_{1\le t\le M}$ une suite de M vecteurs résultant de l'analyse acoustique de dimension p d'un signal de parole prononcé par le locuteur $\boldsymbol{x}$. Les coefficients composant ces vecteurs sont obtenus soit par bancs de filtres, par prédiction linéaire ou par cepstre.

Sous l'hypothèse d'un modèle Gaussien du locuteur [94], la suite des vecteurs $\{x_t\}$ peut être résumée par son vecteur moyenne $\bar{x}$ et sa matrice de covariance X, tels que :

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \tag{3.1}$$

et $$X = \frac{1}{M}\sum_{t=1}^{M} (x_t - \bar{x})(x_t - \bar{x})^T \tag{3.2}$$

avec T qui représente la transposée.

De même, pour un autre locuteur ***y***, la suite $\{y_t\}$ de *N* vecteurs peut être modélisée par $\bar{y}$ et Y , avec :

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \tag{3.3}$$

et $$Y = \frac{1}{N}\sum_{t=1}^{N} (y_t - \bar{y})(y_t - \bar{y})^T \tag{3.4}$$

Les vecteurs moyenne $\bar{x}$ et $\bar{y}$ sont de dimension p, tandis que les covariances X et Y sont des matrices symétriques de dimension p*p.

Ainsi, le locuteur ***x*** (respectivement ***y***) sera représenté par $\bar{x}$, X et *M*, (respectivement $\bar{y}$, Y et *N*).

3.2.2. Notion de Mesure de Similarité

La mesure de similarité $\mu(\boldsymbol{x}, \boldsymbol{y})$ entre les locuteurs ***x*** et ***y*** peut être exprimée comme la fonction Φ suivante :

$$\mu(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\bar{x}, X, M, \bar{y}, Y, N) \tag{3.5}$$

Elle est non-négative, c'est- à- dire :

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ 0 \le \mu(\boldsymbol{x}, \boldsymbol{y}), \tag{3.6}$$

et elle satisfait la propriété (3.7) :

$$\forall\, \boldsymbol{x},\ \mu(\boldsymbol{x}, \boldsymbol{x}) = 0 \tag{3.7}$$

Dans leur forme fondamentale, ces types de mesures ne sont pas symétriques, mais il y a plusieurs méthodes pour les rendre symétriques, bien que :

$$\forall\, \boldsymbol{x},\ \forall\, \boldsymbol{y},\ \mu(\boldsymbol{x}, \boldsymbol{y}) = \mu(\boldsymbol{y}, \boldsymbol{x}) \tag{3.8}$$

3.2.3. Différentes Mesures Statistiques du 2ème Ordre

Les mesures statistiques les plus courantes sont la mesure :

- de Vraisemblance Gaussienne notée "μ_G" ;
- de Vraisemblance Gaussienne à Covariance notée "μ_{Gc}" ;
- Arithmétique-Géométrique Sphérique notée "μ_{Sc}" ;
- de Déviation Absolue notée "μ_{Dc}".

3.2.3.1. Mesure de Vraisemblance Gaussienne

En supposant que tous les vecteurs acoustiques extraits du signal de parole prononcé par le locuteur $\boldsymbol{x}$ sont distribués selon une loi gaussienne, la vraisemblance d'un vecteur acoustique seul y_t prononcé par le locuteur $\boldsymbol{y}$ est donnée par la fonction $G(y_t/\boldsymbol{x})$, suivante :

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} \times \exp(-\frac{1}{2}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x})) \tag{3.9}$$

Si nous supposons que tous les vecteurs y_t sont indépendamment observables, la moyenne du log-vraisemblance de $\{y_t\}_{1\le t\le N}$ peut être décrite par :

$$\bar{G}_{\boldsymbol{x}}(y_1^N) = \frac{1}{N}\log G(y_1 \ldots y_N/\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t/\boldsymbol{x}) =$$

$$= -\frac{1}{2}\left[p\log 2\pi + \log(\det X) + \frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{x})^T X^{-1}(y_t - \bar{x}) \right] \tag{3.10}$$

Par ailleurs, en remplaçant $y_t - \bar{x}$ par $y_t - \bar{y} + \bar{y} - \bar{x}$ et en utilisant la propriété mathématique (3.11) :

$$\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^T X^{-1}(y_t - \bar{y}) = tr(YX^{-1}) \tag{3.11}$$

Nous aurons alors

$$\bar{G}_{\boldsymbol{x}}(y_1^N) + \frac{p}{2}\log 2\pi = -\frac{1}{2}\left[\log(\det X) + tr(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] \tag{3.12}$$

De même l'expression suivante :

$$\frac{2}{p}\bar{G}_{\boldsymbol{x}}(y_1^N) + \log 2\pi + \frac{1}{p}\log(\det Y) + 1$$

$$= \frac{1}{p}\left[\log\left(\frac{\det Y}{\det X}\right) - tr(YX^{-1}) - (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x}) \right] + 1 \tag{3.13}$$

Si nous définissons la mesure de vraisemblance gaussienne μ_G comme :

$$\mu_G(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x}) \right] - 1 \tag{3.14}$$

$$= \frac{1}{p}\left[tr(\Gamma) - \log(\det\Gamma) + \delta^T X^{-1}\delta \right] - 1 \tag{3.15}$$

$$= a - \log g + \frac{1}{p}\delta^T X^{-1}\delta - 1 \tag{3.16}$$

Alors nous aurons :

$$\textbf{Argmax}_{\boldsymbol{x}}\{\bar{G}_{\boldsymbol{x}}(y_1^N)\} = \textbf{Argmin}_{\boldsymbol{x}}\{\mu_G(\boldsymbol{x}, \boldsymbol{y})\} \tag{3.17}$$

Par conséquent cette fonction μ_G peut être assimilée à une mesure de distance inversement proportionnelle à la probabilité de vraisemblance.

Sachant que :

$$a(\lambda_1, \lambda_2, ..., \lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i \tag{3.18}$$

représente la moyenne arithmétique,

$$\Gamma = X^{-\frac{1}{2}} Y X^{-\frac{1}{2}}, \tag{3.19}$$

λ_i : les valeurs propres de Γ ;

X : la matrice de covariance représentant **x**

Y : la matrice de covariance représentant **y**,

$$g(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\prod_{i=1}^{p}\lambda_i\right)^{1/p} \tag{3.20}$$

représente la moyenne géométrique,

et $$\delta = \bar{y} - \bar{x} \tag{3.21}$$

3.2.3.1.1. Propriété de la μ_G

La mesure gaussienne de vraisemblance μ_G est non symétrique. C'est-à-dire que, si on considère son terme dual μ_G (***y, x***),

$$\mu_G(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g + \frac{1}{p}\delta^T Y^{-1}\delta - 1 \tag{3.22}$$

Alors on remarque que μ_G (***x***, ***y***) ≠ μ_G (***y, x***)

Sachant que :

$$h(\lambda_1, \lambda_2, ..., \lambda_p) = \left(\frac{1}{p}\sum_{i=1}^{p}\left(\frac{1}{\lambda_1}\right)_i\right)^{-1} \tag{3.23}$$

représente la moyenne harmonique.

3.2.3.1.2. Inconvénient de la μ_G

Quand on traite un signal de parole distordu ou bruité, les vecteurs moyenne $\bar{x}$ et $\bar{y}$ peuvent être fortement influencés par les caractéristiques de l'environnement ou du canal de transmission, tandis que les matrices de covariance X et Y sont habituellement plus robustes aux variations entre les conditions d'enregistrement et les lignes de transmission du canal. Ainsi la différence $\delta = \bar{y} - \bar{x}$ peut devenir un terme d'erreur pour la μ_G.

Par conséquent, la mesure de vraisemblance gaussienne par covariance μ_{Gc} peut donc être dérivée de la mesure précédente par l'équation (3.24), en supposant que la variabilité interlocuteur de la moyenne est nulle.

Soit :

$$\mu_{Gc}(\boldsymbol{x}, \boldsymbol{y}) = a - \log g - 1 \tag{3.24}$$

Cette mesure (mesure de vraisemblance gaussienne par covariance) est aussi non symétrique :

$$\mu_{Gc}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{h} + \log g - 1 \neq \mu_{Gc}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.25}$$

3.2.3.2. Mesure Arithmétique- Géométrique Sphérique

la mesure Arithmétique- Géométrique Sphérique [94] est donnée par l'équation suivante :

$$\mu_{Sc}(\boldsymbol{x}, \boldsymbol{y}) = \log\left(a/g\right) \tag{3.26}$$

De même, cette mesure n'est pas symétrique :

$$\mu_{Sc}(\boldsymbol{y}, \boldsymbol{x}) = \log\left(g/h\right) \neq \mu_{Sc}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.27}$$

3.2.3.3. Mesure de Déviation Absolue

Elle est donnée par [94] :

$$\mu_{Dc}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{p}\sum_{i=1}^{p}\left|\lambda_i - 1\right| \tag{3.28}$$

Cette mesure est aussi non symétrique car :

$$\mu_{Dc}(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{p}\sum_{i=1}^{p}\left|\frac{1}{\lambda_i} - 1\right| \neq \mu_{Dc}(\boldsymbol{x}, \boldsymbol{y}) \tag{3.29}$$

3.2.4. Procédures de Symétrisation

Toutes les mesures vues précédemment ont une propriété commune d'être non symétriques, autrement dit, les rôles joués par les données de l'apprentissage et celles du test ne sont pas interchangeables. Cependant, notre intuition logique serait que la mesure de similarité devrait être symétrique.

L'asymétrie de μ_G, μ_{Gc}, μ_{Sc} et μ_{Dc} peut être expliquée par le fait que ces mesures sont basées sur des essais statistiques qui supposent que la référence (le modèle du locuteur $\boldsymbol{x}$) est exacte, tandis que le modèle test (le locuteur $\boldsymbol{y}$) est une estimation. Mais en pratique, les deux modèles de référence et de test sont des estimations.

De plus, il est clair que la fiabilité d'un modèle de référence est dépendante du nombre de données qui a été employé pour estimer ces paramètres. Ceci est confirmé lors de nos

expériences par les différences en performance qui peuvent être observées dans les tests d'identification entre $\mu(\boldsymbol{x}, \boldsymbol{y})$ et $\mu(\boldsymbol{y}, \boldsymbol{x})$, surtout si M et N (le nombre de vecteurs de référence et celui de test) sont disproportionnés (c'est- à- dire si $\rho = \frac{N}{M}$ est très différent de 1).

La première possibilité pour la symétrisation de la mesure $\mu(\boldsymbol{x}, \boldsymbol{y})$, est de construire la moyenne $\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y})$ entre la mesure et son terme dual :

$$\mu_{[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\mu(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2}\mu(\boldsymbol{y}, \boldsymbol{x}) = \mu_{[0.5]}(\boldsymbol{y}, \boldsymbol{x}) \tag{3.30}$$

La mesure de vraisemblance gaussienne symétrisée de cette façon, devient :

$$\mu_{G[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[a + \frac{1}{h} + \frac{1}{p}\delta^T\left[X^{-1} + Y^{-1}\right]\delta\right] - 1 \tag{3.31}$$

Tandis que la mesure de vraisemblance à covariance devient :

$$\mu_{Gc[0.5]}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\left[a + \frac{1}{h}\right] - 1 \tag{3.32}$$

Cette procédure de symétrisation peut améliorer nettement les performances de classification, en comparaison avec les deux termes non symétriques pris individuellement.

Cependant, nous observons, en pratique, une dégradation des performances quand les longueurs diffèrent considérablement ($\rho \neq 1$).

Ainsi, si $\rho \leq 1 \Rightarrow M \geq N$ alors $\mu(\boldsymbol{x}, \boldsymbol{y})$ sera meilleure que $\mu(\boldsymbol{y}, \boldsymbol{x})$ et inversement si $\rho \geq 1$.

Sous le manque d'un cadre théorique rigoureux, les vérifications ont été limitées aux expériences empiriques. Une forme arbitraire a été postulée pour la généralité des mesures symétriques, lestées par des coefficients qui sont fonction du nombre de vecteurs d'apprentissage et de test (respectivement M et N).

Cette nouvelle symétrisation notée $\mu_\beta(\boldsymbol{x}, \boldsymbol{y})$ est donnée par l'expression 3.33.

$$\mu_\beta(\boldsymbol{x}, \boldsymbol{y}) = (M*\mu(\boldsymbol{x}, \boldsymbol{y}) + N*\mu(\boldsymbol{y}, \boldsymbol{x})) / (M+N) \tag{3.33}$$

Cette nouvelle mesure est bien adaptée au cas où la durée de test et celle de l'apprentissage sont très différentes.

3.3. Classifieur à Base de Perceptron Multi Couches (ou *Multi-Layer Perceptron* : MLP)

Les Réseaux de Neurones Artificiels (RNA) sont des circuits électroniques dont chaque élément est sensé simuler le fonctionnement d'une cellule élémentaire du cerveau humain qu'est le neurone. Bien souvent en pratique, les chercheurs ne font pas appel à de véritables neurones électriques mais simulent, une nouvelle fois, les RN à l'aide d'un simple programme.

3.3.1. Aspect Biologique des Réseaux de Neurones (RN)

Les cellules nerveuses, appelées neurones, sont les éléments de base du système nerveux central, leur nombre est égal à environ cent milliards [95].

Les neurones possèdent de nombreux points communs dans leur organisation générale et leur système biochimique avec les autres cellules. Cependant, ils présentent des caractéristiques propres et qui se retrouvent au niveau des cinq fonctions spécialisées qu'ils assurent :

- recevoir des signaux en provenance de neurones voisins ;
- intégrer ces signaux ;
- engendrer un influx nerveux ;
- conduire ce flux ;
- transmettre le flux à un autre neurone capable de le recevoir.

Un neurone est constitué de trois parties : le corps cellulaire, les dendrites et l'axone (figure 3.1).

Le corps cellulaire contient le noyau du neurone et effectue les transformations biochimiques nécessaires à la synthèse des enzymes et des autres molécules qui assurent la vie du neurone. Sa forme est pyramidale ou sphérique dans la plupart des cas.

Chaque neurone possède une chevelure de dendrites, qui sont de fines extensions tubulaires. Elles se ramifient, en formant une espèce d'arborescence autour du corps cellulaire. Elles sont les récepteurs principaux du neurone pour capter les signaux qui lui parviennent.

L'axone est la fibre nerveuse. Il sert de moyen de transport pour les signaux émis par le neurone. Il se distingue des dendrites par sa forme et sa longueur en se ramifiant à son extrémité, là où il communique avec les autres neurones [95].

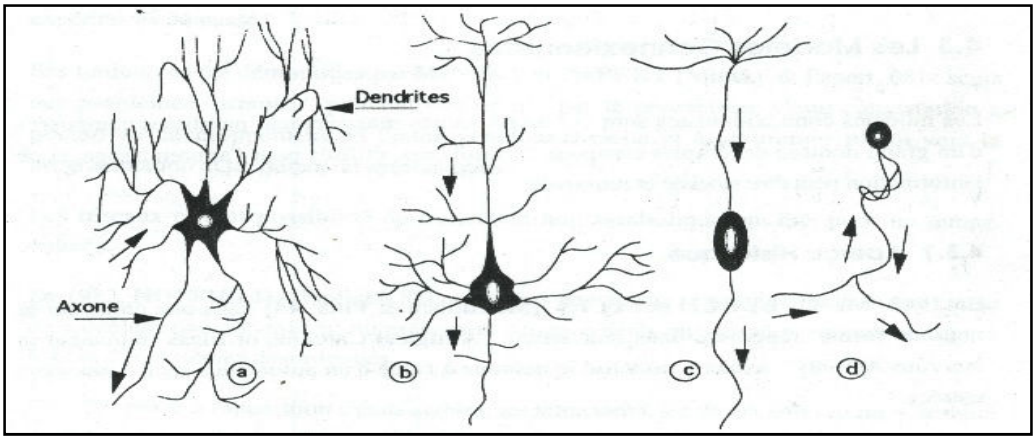

Figure 3.1 : Quelques silhouettes de neurones.
a) motoneurone de la corne antérieure de la moelle ; b) cellule pyramidale du cerveau ; c) cellule olfactive ; d) neurone en T du ganglion spinal [96].

3.3.2. Fonctionnement et Modélisation du RN

Quand un neurone a reçu suffisamment de signaux lancés par des milliers d'autres neurones à travers les synapses (figure 3.2), il envoie à son tour une décharge d'énergie vers les autres neurones qui lui sont connectés par l'intermédiaire de son axone (figure 3.3).

En ce sens, le neurone ressemble à une batterie à seuil. Quand la batterie atteint une charge donnée, elle délivre une impulsion de courant et revient à une charge plus faible. Dans cette description du fonctionnement du neurone, la fonction d'activation est définie comme étant une somme pondérée des signaux reçus par chacune des dendrites.

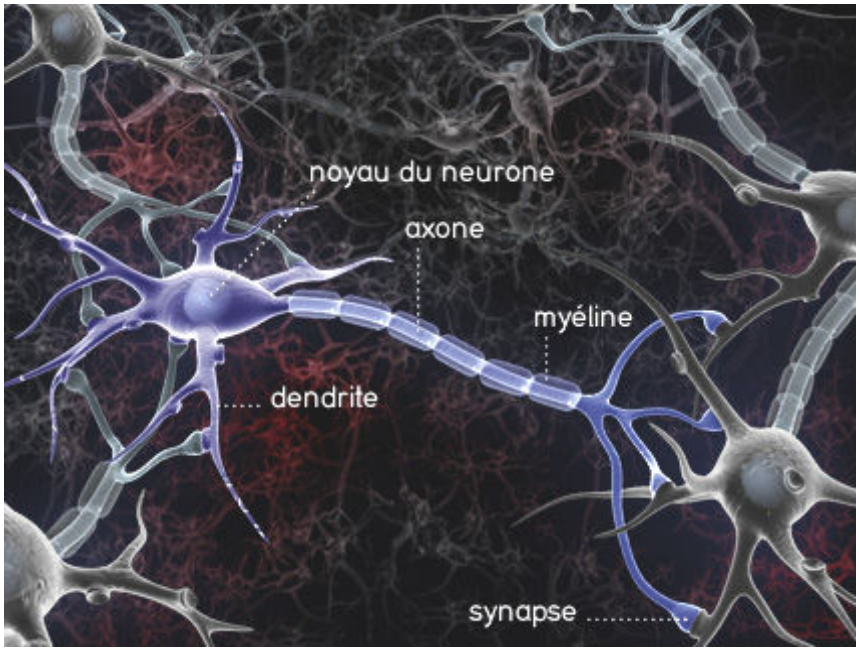

Figure 3.2 : Communication neuronale à l'aide des synapses [96].

Les réseaux connexionnistes peuvent être définis comme un ensemble d'unités (ou automates) réalisant des calculs élémentaires. Elles sont structurées en couches successives capables d'échanger des informations au moyen des connexions qui les relient. Chaque unité effectue un traitement local d'informations.

Il existe plusieurs types d'unités :

- d'entrée, auxquelles sont transmises les données à traiter, en provenance de sources externes au réseau ;
- de sortie, qui contiennent l'information traitée utilisable par d'autres systèmes connectés au réseau ;
- cachées, dont les entrées et les sorties sont reliées aux autres unités du réseau, et sont donc non "visibles" par des systèmes extérieurs. Ces dernières servent à coder, de façon interne au système, la structure des formes présentées à l'entrée.

Ce type de neurone calcule d'abord une fonction d'entrée [97] :

$$a_i = \sum_j Wij * Xj \tag{3.34}$$

Puis son activité est donnée par :

$$S_i = f(a_i) = f\left(\sum_j Wij * Xj\right) \tag{3.35}$$

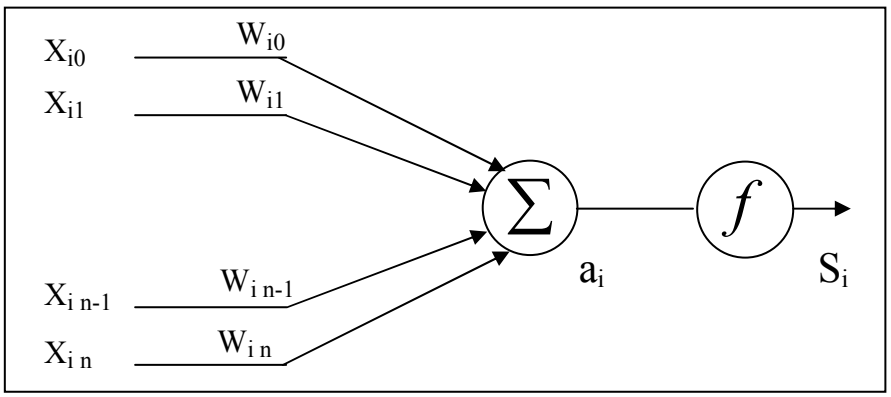

Figure 3.3 : Exemple d'unité neuronale.

X_j : entrée provenant du neurone j.

W_{ij} : poids de la connexion de j vers i.

Où f peut être une fonction identité (purelin), une fonction à seuil (hardlim) ou une fonction sigmoïde (logsig) (figure 3.4).

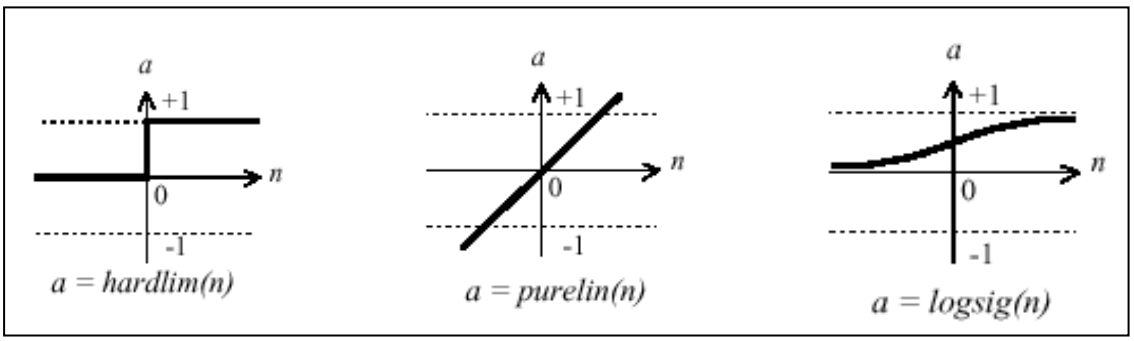

Figure 3.4 : Quelques types de fonctions de transfert.

La fonction sigmoïde possède les particularités suivantes [98] :

- elle est continue sur l'ensemble des valeurs réelles ;
- sa dérivée est toujours positive ;
- elle est bornée : les intervalles des valeurs de sortie sont généralement]0,1[ou]-1,1[.

3.3.3. Construction des RN

Un RN peut être représenté par un graphe direct composé d'un ensemble de nœuds ou éléments processeurs, fortement interconnectés par des liens orientés ou connexions.

Les RNs peuvent se distinguer par leur architecture et leur mode d'apprentissage.

3.3.3.1. Architecture des Réseaux Monocouches

Dans les réseaux monocouches, l'ensemble des unités d'entrée est connecté à l'ensemble des unités de sortie par une couche de connexions modifiables. Cette architecture ne comporte pas de cellule cachée (figure 3.5)

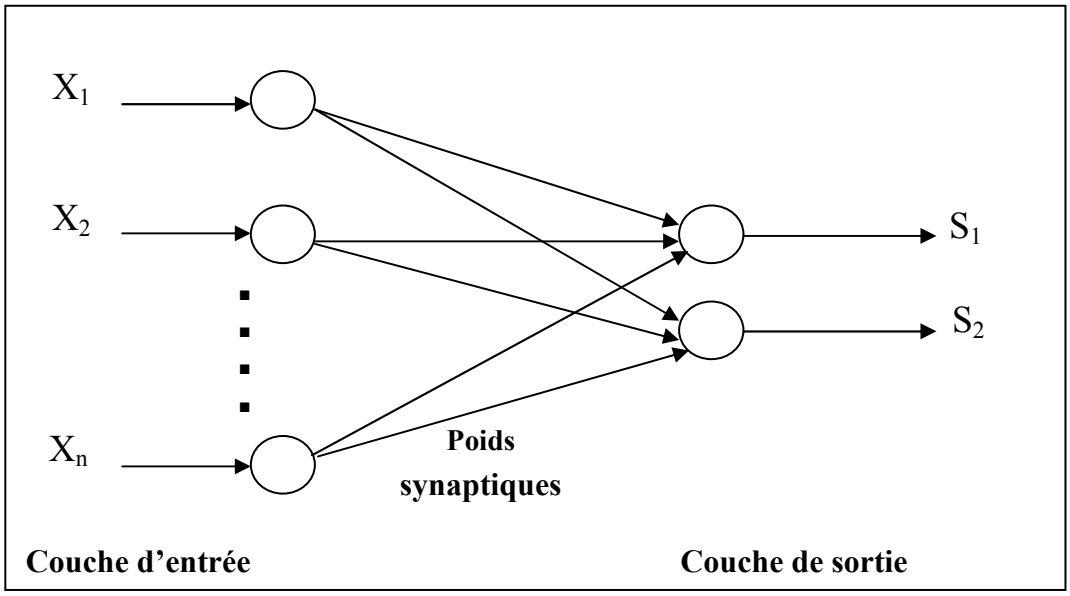

Figure 3.5 : Architecture d'un réseau monocouche à deux sorties.

Un réseau multicouche est un réseau de neurones structuré en cascade de groupes de neurones (ou couches). Chaque couche est connectée à la suivante (figure 3.6).

Cependant, il existe plusieurs façons de relier les unités d'un réseau, la plus simple est la connectivité totale : chaque unité est connectée à toutes les unités de la couche suivante.

Les unités de la même couche ne sont pas connectées entre elles. La première couche s'appelle alors couche d'entrée, la dernière : couche de sortie et les autres sont dites couches cachées.

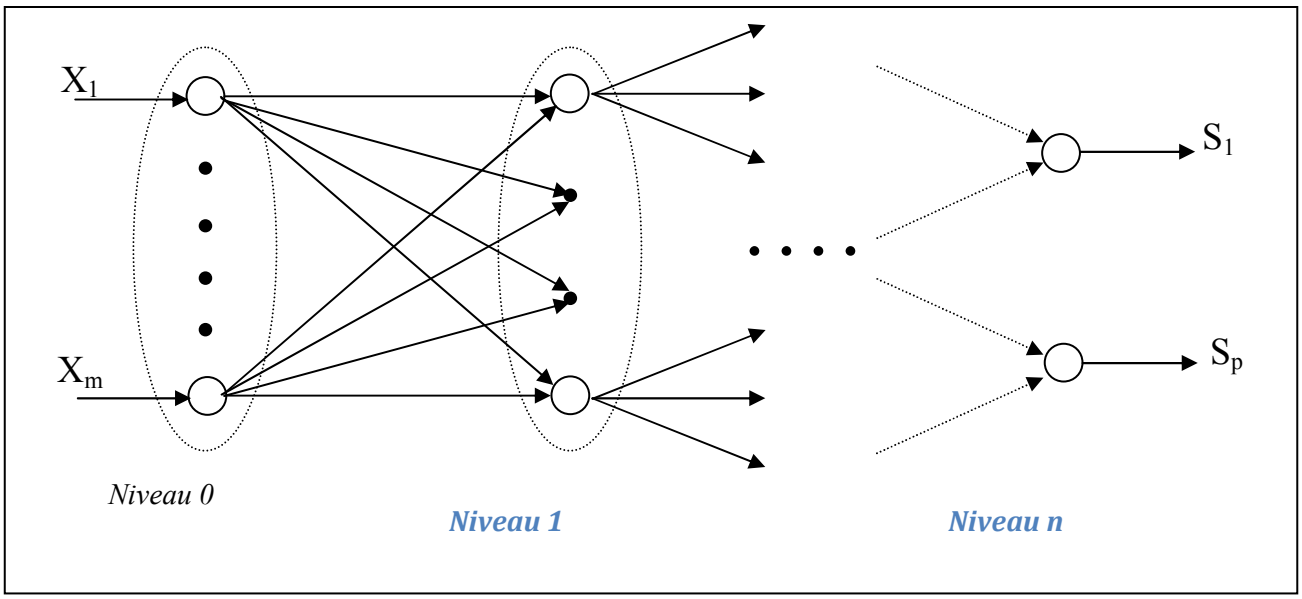

Figure 3.6 : Architecture d'un réseau multicouche à (n+1) niveaux.

3.3.3.2. Mode d'Apprentissage

Il existe deux types classiques d'apprentissage supervisé et non supervisé, appelé également apprentissage sans professeur.

3.3.3.2.1. Apprentissage Supervisé

Le but de l'apprentissage supervisé est d'inculper un comportement de référence au réseau. On suppose alors qu'à chaque patron d'entrée est associée une sortie désirée qui spécifie les valeurs de sortie. L'apprentissage se déroule de la façon suivante :

Un patron est présenté aux neurones d'entrée, puis l'activation est propagée à travers le réseau, et enfin, la réponse des neurones de sortie est alors comparée aux valeurs désirées. Ceci détermine l'erreur du réseau pour le patron donné. Il s'agit alors de répartir cette erreur à chaque poids du réseau en fonction de la part qu'il a jouée dans la production de la réponse erronée. On procède alors à une modification des poids qui vise à réduire l'erreur ainsi calculée.

L'ensemble d'apprentissage est donc constitué de paires de vecteurs {(Xi,Yi)}où : i=1,…, N.

N : le nombre d'exemples d'apprentissage.

Xi : le vecteur d'entrée.

Yi : le vecteur de la sortie désirée.

La procédure d'apprentissage consiste à :

- appliquer le vecteur d'entrée ;

- calculer le vecteur de sortie ;
- comparer le vecteur de sortie au vecteur de la sortie désirée. La différence qui est l'erreur, est répliquée au réseau en modifiant les poids suivant un algorithme qui tend à minimiser cette erreur [99].

3.3.3.2.2. Apprentissage non Supervisé

Parallèlement et de façon complémentaire au développement des techniques d'apprentissage supervisé, ont été développées des techniques dites non supervisées. Elles visent à faire apprendre certaines informations sur des données non étiquetées. Il existe de nombreux cas où on ne possède pas d'informations sur les classes de l'ensemble d'apprentissage. Ce manque de connaissance peut avoir plusieurs causes : manque d'informations sur les données, volume d'informations trop important pour pouvoir être étiqueté à la main [97], etc.

Dans cette catégorie d'apprentissage, la règle n'est pas fonction du comportement de sortie du réseau, mais plutôt du comportement local des neurones [95].

3.3.4. Perceptron

Un perceptron est un réseau de cellules composé de plusieurs modules, disposés en couches, la :

- rétine ou première couche contient les cellules d'entrée. Chaque cellule se contente de recopier la valeur qu'elle reçoit de l'extérieur sur sa sortie ;
- couche d'association ou deuxième couche est composée de cellules associatives. Chaque cellule a des connexions entrantes pouvant provenir de toutes ou d'une partie des cellules de la rétine. Les fonctions de transfert de ces cellules f_i sont fixées à priori et sont en général différentes d'une cellule à une autre ;
- couche de cellules de décision est composée d'automates à seuil. Chaque automate est connecté à toutes les sorties de la couche précédente. Les coefficients linéaires (les poids) de ces cellules sont déterminés par apprentissage. Chaque cellule de décision calcule donc sa sortie selon l'expression suivante :

$$Y_j(X) = H\left[\sum_i W_i f_i(X)\right] \tag{3.36}$$

où $Y_j(X)$ représente la sortie de la $j^{ième}$ cellule.

X désigne la forme présentée en entrée sur la rétine R et H désigne la fonction de seuillage. Seule cette seconde couche est donc adaptative et soumise à l'apprentissage.

Le perceptron est un réseau qui a pour tâche la classification des formes $X_1, X_2,\ldots,X_m$ présentées sur la rétine en p classes $C_1, C_2,\ldots, C_p$.

L'apprentissage est supervisé et permet l'adaptation des poids des connexions de la couche de sortie.

Considérons, un problème de classification de p modèles dans un espace de représentation égal à R^n.

Soit E = {(Xi, Yi)} l'ensemble d'apprentissage où :

Xi $\in R^n$ vecteur d'entrée.

Yi $\in \{1, -1\}^p$ vecteur de la sortie désirée.

Par convention si Xi appartient à la classe ou modèle j, tel que $1 \leq j \leq p$, alors on souhaite lui associer un vecteur de sortie $Yi = (Y_{i1}, Y_{i2},\ldots, Y_{ip})^t$ tel que :

$$Y_{ik} = \begin{cases} +1 & si\ \ k = j \\ -1 & .si\ \ k \neq j \end{cases} \tag{3.37}$$

L'apprentissage est effectué selon les étapes suivantes :

étape 1 : initialiser aléatoirement les poids des connexions de la couche de sortie ;

étape 2 : choisir aléatoirement un couple (Xi, Yi) de l'ensemble E ;

Présenter Xi en entrée sur la rétine, puis calculer la sortie du réseau à l'instant t.

$$Y(Xi, t) = (Y_1(Xi), Y_2(Xi),\ldots, Y_p(Xi))^t \tag{3.38}$$

étape 3 : Pour tout automate K, pour lequel la sortie effective est différente de la sortie désirée ($Y_k(Xi, t) \neq Y_{ik}$) modifier les poids par la formule :

$$W_k(t+1) = W_k(t) + \varepsilon * \delta_{ik} * f_k(Xi) \tag{3.39}$$

où : ε = constante positive.

$\delta_{i\,k} = Y_{i\,k} - Y_k(Xi, t)$ est le signal d'erreur sur le $k^{ième}$ automate de sortie pour l'entrée Xi à l'étape t.

$f_k(Xi)$: sortie de la $k^{ième}$ cellule associative.

étape 4 : Incrémenter **t :** t = t+1 ;

Si condition d'arrêt non remplie, aller à l'étape 2 ;

Sinon fin.

3.3.5. Perceptron Multicouche MLP

Le perceptron multicouche est un réseau de groupes de neurones ou couches. Chaque couche est connectée à la suivante. Il comporte en général une couche d'entrée, une de sortie et une ou plusieurs couches dites cachées.

La fonction d'activation des neurones de ce réseau est la fonction sigmoïde. Dans ce type de réseau, le superviseur fournit à l'entrée un ensemble de couples (entrée, sortie désirée).

L'information circule et se propage de l'entrée vers la sortie, et le réseau calcule sur sa couche de sortie un résultat qui doit être le plus proche possible de la sortie désirée pour toutes les entrées.

L'apprentissage est supervisé, réalisé avec l'algorithme de rétropropagation du gradient de l'erreur entre la sortie calculée et celle désirée. Le principe de cet algorithme est que, de même que l'on est capable de propager un signal provenant des cellules d'entrées vers la couche de sortie. On peut, en suivant le chemin inverse, rétropropager l'erreur commise en sortie vers les couches internes, afin d'ajuster les poids synaptiques du réseau en commençant par les dernières couches jusqu'aux premières, afin que le réseau converge vers un état qui permettra à tous les modèles d'apprentissage d'être codés [100].

3.3.6. Algorithme de Rétropropagation du Gradient (RPG)

La rétropropagation du gradient est certainement l'un des plus simples et des plus efficaces algorithmes d'apprentissage pour les réseaux multicouches [97].

Mathématiquement, cet algorithme utilise simplement les règles de dérivations composées et ne présente aucune difficulté particulière. Le principe de cet algorithme est que, de même que l'on est capable de propager un signal provenant des cellules d'entrées vers la couche de sortie, on peut, en suivant le chemin inverse, rétropropager l'erreur commise en sortie vers les couches internes.

Dans l'apprentissage des réseaux à RPG, on dispose d'un ensemble d'exemples qui sont les couples (entrées, sorties désirées). A chaque étape un exemple est présenté en entrée du réseau. Un signal est propagé de proche en proche à travers chaque couche supérieure à la couche précédente jusqu'à ce qu'une sortie soit générée. Cette dernière est alors comparée à la sortie désirée et un signal d'erreur (somme quadratique des erreurs sur chaque cellule de sortie) est calculé. Ce signal est ensuite rétropropagé de la couche de sortie vers chaque cellule de la couche intermédiaire qui contribue directement à la sortie.

Cependant, chaque unité dans la couche intermédiaire reçoit seulement une portion de l'erreur totale basée sur la contribution relative de cette unité à la génération de sortie.

Ce processus est répété, couche par couche, jusqu'à ce que chaque cellule du réseau ait reçu le signal d'erreur.

En parallèle, les poids des connexions sont alors mis à jour pour chaque cellule, et cela pour permettre au réseau de converger vers un état qui permettra à tous les modèles d'apprentissage d'être codés.

Ce processus est répété, en présentant successivement chaque exemple. Si pour tous les exemples l'erreur est inférieure à un seuil choisi, on dit alors que le réseau a convergé.

Le sens de ce processus est que durant l'apprentissage du réseau, les cellules des couches intermédiaires s'organisent de manière à ce qu'elles apprennent à reconnaître les différentes caractéristiques de tout l'espace d'entrée [99], [95] et [101].

3.4. Classifieur à Base de Machine à Vecteurs de Support (SVM)

Les machines à vecteurs de support ou séparateurs à vaste marge (en Anglais *Support Vector Machine*, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires.

Les SVM ont été développés durant les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la Théorie de Vapnik-Chervonenkis. Les SVM ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyper paramètres, le fait qu'ils soient bien fondés théoriquement, et leurs bons résultats en pratique.

Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance, etc.). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un Modèle de Mixture de Gaussiennes (en Anglais : *Gaussian Mixture Models*, GMM) [102].

3.4.1. Historique des SVM

Les SVM reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau. Ces deux notions existaient depuis plusieurs années avant qu'elles ne soient mises en commun pour construire les SVM.

L'idée des hyperplans à marge maximale a été explorée dès 1963 par Vladimir Vapnik et A. Lerner, et en 1973 par Richard Duda et Peter Hart dans leur livre Pattern Classification. Les fondations théoriques des SVM ont été explorées par Vapnik et ses collègues dans les années 70 avec le développement de la Théorie de Vapnik-Chervonenkis, et par Valiant et la théorie de l'apprentissage "Probablement approximativement Correcte" (en Anglais : *Probably Approximately Correct*, PAC).

L'idée des fonctions noyaux n'est pas non plus nouvelle : le théorème de Mercer date de 1909, et l'utilité des fonctions noyaux dans le contexte de l'apprentissage artificiel a été montrée dès 1964 par Aizermann, Bravermann et Rozoener [102].

Ce n'est toutefois qu'en 1992 que ces idées ont été bien comprises et rassemblées par Boser, Guyon et Vapnik, dans un article, fondateur des SVM. L'idée des variables ressorts, qui permet de résoudre certaines limitations pratiques importantes, ne sera introduite qu'en 1995. À partir de la date de la publication du livre de Vapnik, les SVM gagnent en popularité et sont utilisés dans de nombreuses applications.

Les SVM sont des classifieurs qui reposent sur deux idées clés, permettant de traiter des problèmes de discrimination non-linéaire, et de reformuler le problème de classification comme un problème d'optimisation quadratique.

La première idée clé est la notion de *marge maximale*. La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Ces derniers sont appelés *vecteurs supports*. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge. Ce choix est justifié par la théorie de Vapnik-Chervonenkis (ou théorie statistique de l'apprentissage), qui montre que la frontière de séparation de marge maximale possède la plus petite capacité. Le problème est de trouver cette frontière séparatrice optimale, à partir d'un ensemble d'apprentissage. Ceci est fait en formulant le problème comme un problème d'optimisation quadratique, pour lequel il existe des algorithmes connus [102].

Afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables, la deuxième idée clé des SVM est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension, dans lequel il est probable qu'il existe une séparatrice linéaire. Ceci est réalisé grâce à une fonction noyau, qui doit respecter certaines conditions, et qui a l'avantage de ne pas nécessiter la connaissance explicite de la transformation à appliquer pour le changement d'espace. Les fonctions noyau permettent de transformer un produit scalaire dans un espace de grande dimension, ce qui est coûteux, en une simple évaluation ponctuelle d'une fonction. Cette technique est connue sous le nom de "*kernel trick*" [102].

3.4.2. Principe Général des SVM

Les SVM peuvent être utilisés pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou des problèmes de régression, c'est-à-dire prédire la valeur numérique d'une variable. La résolution de ces deux problèmes passe par la construction d'une fonction f qui à un vecteur d'entrée x fait correspondre une sortie y :

$$y = f(x) \tag{3.40}$$

Pour l'instant, on se limite à un problème de discrimination à deux classes (discrimination binaire), c'est-à-dire $y \in \{-1, 1\}$, le vecteur d'entrée x étant dans un espace X muni d'un produit scalaire. On peut prendre par exemple : $X = \mathbb{R}^N$.

3.4.2.1. Discrimination Linéaire et Hyperplan Séparateur

A titre de rappel, le cas simple est le cas d'une fonction discriminante linéaire, obtenue par combinaison linéaire du vecteur d'entrée $x = (x_1,\ldots, x_N)^T$:

$$h(x) = w^T x + w_0 \tag{3.41}$$

Il est alors décidé que *x* est de classe :

- 1 si h(x) ≥ 0 ;
  et
- -1 dans le cas contraire.

C'est donc un classifieur linéaire.

La frontière de décision $h(x) = 0$ est un hyperplan, appelé *hyperplan séparateur*, ou *séparatrice*. Rappelons que le but des algorithmes à apprentissage supervisé est d'apprendre la fonction h(x) par le biais d'un ensemble d'apprentissage :

$$\{(x_1, l_1), (x_2, l_2), \ldots, (x_p, l_p)\} \in \mathbb{R}^N \times \{-1,1\} \tag{3.42}$$

où les l_k sont les étiquettes, p est la taille de l'ensemble d'apprentissage, N la dimension des vecteurs d'entrée. Si le problème est linéairement séparable, on doit alors avoir :

$$l_k h(x_k) \geq 0 \quad 1 \leq k \leq p, \qquad \text{autrement dit} \qquad l_k(w^T x_k + w_0) \geq 0 \quad 1 \leq k \leq p. \tag{3.43}$$

Prenons un exemple pour bien comprendre le concept de séparation. Imaginons un plan dans lequel sont répartis deux groupes de points. Ces derniers sont associés à un groupe : les points (+) pour y > x et les points (-) pour y < x. On peut trouver un séparateur linéaire évident dans cet exemple, la droite d'équation y=x. Dans ce cas le problème est dit *linéairement séparable* (figure 3.7).

Pour des problèmes plus compliqués, en général, il n'existe pas de séparateur linéaire. Imaginons par exemple un plan dans lequel les points (-) sont regroupés en un cercle, avec des points (+) tout autour (figure 3.8) : aucun séparateur linéaire ne peut correctement séparer les groupes : le problème n'est pas *linéairement séparable* et il n'existe pas d'hyperplan séparateur.

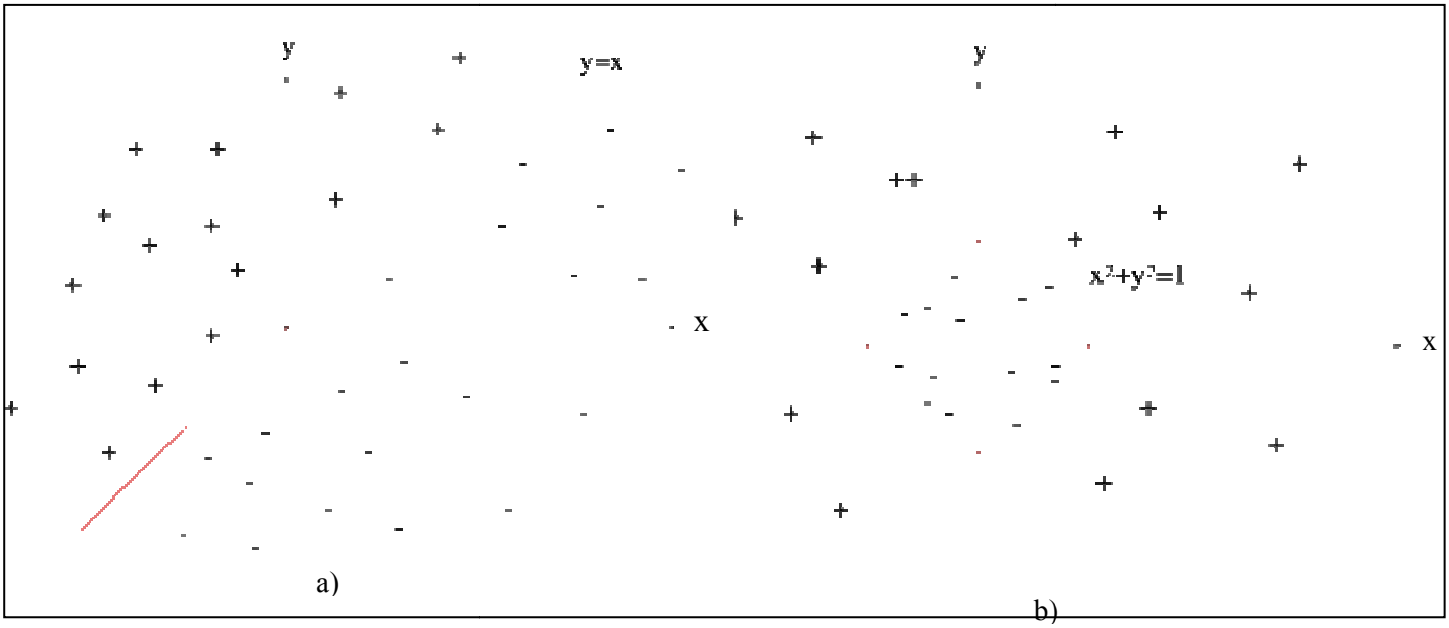

Figure 3.7 : Exemples d'un problème de discrimination à deux classes, avec une séparatrice :
a) linéaire. b) non-linéaire.

3.4.2.2. Marge Maximale

On se place désormais dans le cas où le problème est linéairement séparable. Même dans ce cas simple, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque empirique est le même), mais dont les performances en généralisation peuvent être très différentes (figure 3.8). Pour résoudre ce problème, il a été montré qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur (figure 3.9).

Il existe des raisons théoriques à ce choix. Vapnik a montré que la capacité des classes d'hyperplans séparateurs diminue lorsque leur marge augmente [102].

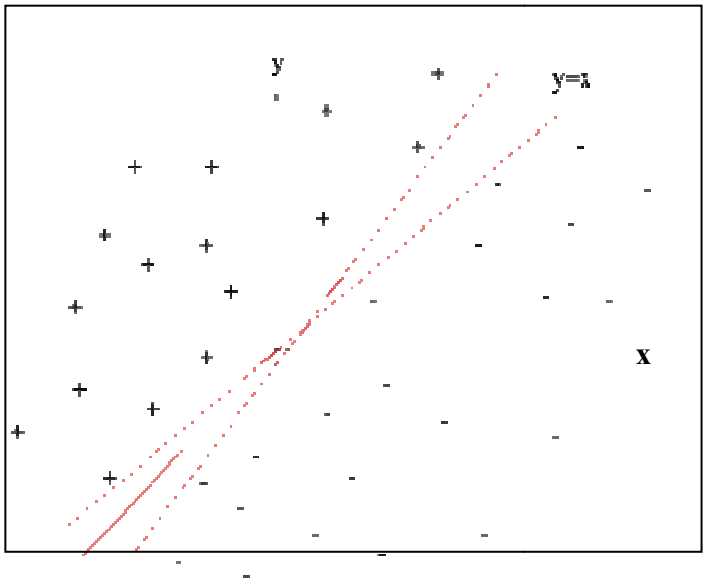

Figure 3.8 : Ensemble de points linéairement séparables, il existe une infinité d'hyperplans séparateurs.

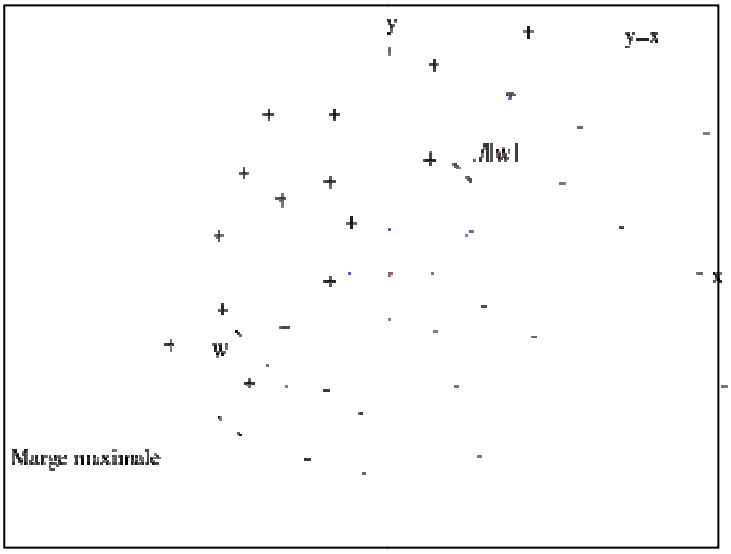

Figure 3.9 : Hyperplan optimal (en rouge) avec la marge maximale. Les échantillons entourés sont des *vecteurs supports*.

La marge est la distance entre l'hyperplan et les échantillons les plus proches. Ces derniers sont appelés **vecteurs supports**. L'hyperplan qui maximise la marge est donné par :

$$\arg\max_{w,w_0} \min_{k} \{||x - x_k|| : x \in \mathbb{R}^N, w^T x + w_0 = 0\} \tag{3.44}$$

Il s'agit donc de trouver w et w_0 remplissant ces conditions, afin de déterminer l'équation de l'hyperplan séparateur :

$$h(x) = w^T x + w_0 = 0 \tag{3.45}$$

3.4.2.3. Recherche de l'Hyperplan Optimal

La marge est la plus petite distance entre les échantillons d'apprentissage et l'hyperplan séparateur qui satisfasse la condition de séparabilité (à savoir $l_k(w^T x_k + w_0) \geq 0$ comme

expliqué précédemment). La distance d'un échantillon x_k à l'hyperplan est donnée par sa projection orthogonale sur l'hyperplan :

$$\frac{l_k(w^T x_k + w_0)}{||w||}$$

L'hyperplan séparateur (w, w_0) de marge maximale est donc donné par :

$$\arg\max_{w,w_0}\left\{\frac{1}{||w||}\min_k\left[l_k(w^T x_k + w_0)\right]\right\} \tag{3.46}$$

Afin de faciliter l'optimisation, on choisit de normaliser w et w_0, de telle manière que les échantillons à la marge (x^+_{marge} pour les vecteurs supports sur la frontière positive, et x^-_{marge} pour ceux situés sur la frontière opposée) satisfassent :

$$\begin{cases} w^T x^+_{marge} + w_0 = 1 \\ w^T x^-_{marge} + w_0 = -1 \end{cases}$$

d'où pour tous les échantillons, $k = 1, \ldots, p$

$$l_k(w^T x_k + w_0) \geq 1 \tag{3.47}$$

Cette normalisation est parfois appelée la forme canonique de l'hyperplan, ou *hyperplan canonique*.

Avec cette mise à l'échelle, la marge vaut désormais $\frac{1}{||w||}$, il s'agit donc de maximiser $||w||^{-1}$. La formulation dite *primale* des SVM s'exprime alors sous la forme suivante :

$$\text{Minimiser} \quad \frac{1}{2}||w||^2 \quad \text{sous les contraintes} \quad l_k(w^T x_k + w_0) \geq 1 \tag{3.48}$$

Ceci peut se résoudre par la méthode classique des Multiplicateurs de Lagrange, où le lagrangien est donné par :

$$L(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum_{k=1}^{p} \alpha_k \left\{l_k(w^T x_k + w_0) - 1\right\} \tag{3.49}$$

Le lagrangien doit être minimisé par rapport à w et w_0, et maximisé par rapport à α.

En annulant les dérivées partielles du lagrangien, selon les conditions de Kuhn-Tucker, on obtient :

$$\begin{cases} \sum_{k=1}^{p} \alpha_k l_k x_k & = w^* \\ \sum_{k=1}^{p} \alpha_k l_k & = 0 \end{cases} \tag{3.50}$$

En réinjectant ces valeurs dans l'équation (3.49), on obtient la *formulation duale* :

$$\text{Maximiser } \tilde{L}(\alpha) = \sum_{k=1}^{p} \alpha_k - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j l_i l_j x_i^T x_j \tag{3.51}$$

Sous les contraintes $\alpha_k \geq 0$, et $\sum_{k=1}^{p} \alpha_k l_k = 0$

Ce qui donne les multiplicateurs de Lagrange optimaux α_k^*.

Afin d'obtenir l'hyperplan solution, on remplace w par sa valeur optimale w^*, dans l'équation de l'hyperplan $h(x)$, c0e qui donne :

$$h(x) = \sum_{k=1}^{p} \alpha_k^* l_k (x \cdot x_k) + w_0 \tag{3.52}$$

Il y a trois remarques intéressantes à faire à propos de ce résultat. La première découle de l'une des conditions de Kuhn-Tucker, qui donne :

$$\alpha_k [l_k h(x_k) - 1] = 0 \quad 1 \leq k \leq p. \tag{3.53}$$

d'où

$$\begin{cases} \alpha_k & = 0 \\ l_k h(x_k) & = 1 \end{cases} \tag{3.54}$$

- les seuls points pour lesquels les contraintes du lagrangien sont actives sont donc les points tels que $l_k h(x_k) = 1$, qui sont les points situés sur les hyperplans de marges maximales. En d'autres termes, seuls les *vecteurs supports* participent à la définition de l'hyperplan optimal ;
- la deuxième remarque découle de la première. Seul un sous-ensemble restreint de points est nécessaire pour le calcul de la solution, les autres échantillons ne participant pas du tout à sa définition. Ceci est donc efficace au niveau de la complexité. D'autre part, le changement ou l'agrandissement de l'ensemble d'apprentissage a moins d'influence que dans un modèle de mélanges gaussiens par exemple, où tous les points participent à la solution. En particulier, le fait d'ajouter des échantillons à l'ensemble d'apprentissage, qui ne sont pas des vecteurs supports, n'a aucune influence sur la solution finale ;
- la dernière remarque est que l'hyperplan solution ne dépend que du produit scalaire entre le vecteur d'entrée et les vecteurs supports. Cette remarque est l'origine de la deuxième innovation majeure des SVM : le passage par un espace de redescription grâce à une fonction noyau [103].

3.4.2.4. Technique du "*Kernel trick*" (Cas non séparable)

La technique du *"Kernel trick"* est utilisée pour résoudre les cas où il n'existe pas d'hyperplan séparateur dans l'espace où sont répartis les deux groupes de points.

Soit l'exemple suivant, illustré par la figure 3.10 :

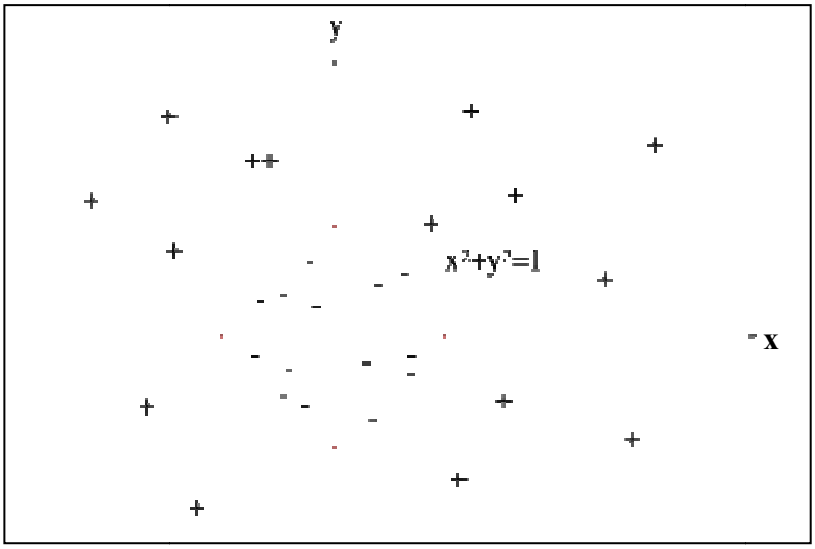

Figure 3.10 : Exemple simple de transformation : le problème n'est pas linéairement séparable en coordonnées cartésiennes, par contre en coordonnées polaires, le problème devient linéaire.

Il s'agit ici d'un exemple très simple, l'espace de redescription étant de même dimension que l'espace d'entrée.

La notion de marge maximale et la procédure de recherche de l'hyperplan séparateur, telles que présentées précédemment, ne permettent de résoudre que des problèmes de discrimination linéairement séparables. C'est une limitation sévère qui condamne à ne pouvoir résoudre que des problèmes très particuliers. Afin de remédier au problème de l'absence de séparateur linéaire, l'idée des SVM est de reconsidérer le problème dans un espace de dimension supérieure, éventuellement de dimension infinie. Dans ce nouvel espace, il est alors probable qu'il existe un séparateur linéaire.

Plus formellement, on applique aux vecteurs d'entrée x une transformation non-linéaire ϕ et l'espace d'arrivée $\phi(x)$ est alors appelé *espace de redescription*. Dans cet espace, on cherchera l'hyperplan $h(x)$:

$$h(x) = w^T \phi(x) + w_0 \tag{3.55}$$

qui vérifie la condition :

$l_k h(x_k) > 0$, pour tous les points x_k de l'ensemble d'apprentissage, c'est- à- dire l'hyperplan séparateur dans l'espace de redescription.

En utilisant la même procédure que dans le cas sans transformation, on aboutit au problème d'optimisation suivant :

$$\text{Maximiser } \tilde{L}(\alpha) = \sum_{k=1}^{p} \alpha_k - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j l_i l_j \phi(x_i)^T \phi(x_j) \tag{3.56}$$

$$\text{Sous les contraintes } \alpha_i \geq 0 \text{ , et } \sum_{k=1}^{p} \alpha_k l_k = 0$$

Le problème de cette formulation est qu'elle implique un produit scalaire entre vecteurs dans l'espace de redescription, de dimension élevée, ce qui est coûteux en terme de calculs. Pour résoudre ce problème, on utilise une astuce connue sous le nom de "*Kernel trick*", qui consiste à utiliser une fonction noyau, qui vérifie la condition suivante :

$$K(x_i, x_j) = \phi(x_i)^T \cdot \phi(x_j) \tag{3.57}$$

d'où l'expression de l'hyperplan séparateur en fonction de la fonction noyau :

$$h(x) = \sum_{k=1}^{p} \alpha_k^* l_k K(x_k, x) + w_0 \tag{3.58}$$

L'intérêt de la fonction noyau est, alors, double :

- le calcul se fait dans l'espace d'origine, ceci est beaucoup moins coûteux qu'un produit scalaire en grande dimension ;
- la transformation ϕ n'a pas besoin d'être connue explicitement : seule la fonction noyau intervient dans les calculs. On peut donc envisager des transformations complexes, et même des espaces de redescription de dimension infinie [102].

3.4.2.5. Choix de la Fonction Noyau

En pratique, on construit directement une fonction noyau qui doit respecter certaines conditions, elle doit correspondre à un produit scalaire dans un espace de grande dimension. Le théorème de Mercer définit les conditions que K doit satisfaire pour être une fonction noyau : elle doit être symétrique et Semi-définie positive.

L'exemple le plus simple de K est le noyau linéaire :

$$K(x_i, x_j) = x_i^T \cdot x_j \tag{3.59}$$

On se ramène au cas d'un classifieur linéaire, sans changement d'espace. L'approche par "*Kernel trick*" généralise donc l'approche linéaire en faisant un cas particulier. Le noyau linéaire est parfois employé pour évaluer la difficulté du problème.

Les noyaux usuels employés pour les SVM sont le noyau :

- polynomial :

$$K(x_i, x_j) = (x_i^T \cdot x_j)^d \tag{3.60}$$

- gaussien ou RBF (*Radial Basis Function*) :

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \tag{3.61}$$

3.5. Caractéristique Relative du Locuteur (RSC)

Dans le but d'optimiser le nombre d'entrées et la convergence du MLP et des SVM, nous avons développé une nouvelle Caractéristique Relative du Locuteur, que nous avons appelée RSC (*Relative Speaker Characteristic*).

3.5.1. Dimension des Entrées des Classifieurs MLP et SVM

Les RN nécessitent un nombre de neurones à leurs entrées, égal à la dimension du vecteur exemple. Alors, dans le cas d'un vecteur de caractéristiques avec P coefficients représentant un segment de parole, le nombre de neurones dans la couche d'entrée doit être 2P, parce que nous avons deux segments de parole différents à faire comparer [104] et [19].

De ce fait, si nous avons 37 caractéristiques acoustiques (dans notre cas : 37 MFSC (*Mel Frequency Spectral Coefficients*)) pour chaque locuteur [105] et [106], nous aurons besoin de 37x37 = 1369 neurones pour la matrice de covariance dans le but de modéliser correctement chaque segment de parole. En pratique, nous considérons la moitié de la matrice (puisqu'elle est symétrique), mais nous avons besoin de deux matrices (puisque nous avons deux segments). Le nombre de neurones d'entrée devient : 2(1369 / 2) =1369 pour chaque exemple d'entrée, tandis que si nous utilisons la caractéristique réduite DRSC (*Diagonal of the Relative Speaker Characteristic*) nous avons besoin d'un vecteur de 37 (ou 74 pour les deux segments) éléments seulement à l'entrée du MLP ou du SVM, ce qui représente 2.7% (ou 5.4% pour les deux segments) de la dimension totale des entrées. Cette réduction permet un temps d'apprentissage plus réduit, et améliore aussi la discrimination.

3.5.2. Notion de DRSC (*Diagonal of* RSC)

Pour modéliser le locuteur, la plupart des systèmes de reconnaissance du locuteur existants utilisent les composantes statistiques comme le vecteur moyenne ou la matrice de covariance, qui sont extraits à partir des caractéristiques acoustiques comme : les coefficients MFSC, les LFCC (*Linear Frequency Cepstral Coefficients*), les MFCC (*Mel Frequency Cepstral Coefficients*), leurs dérivées premières et secondes et les AR (*Auto Regressif*). Ainsi, pour modéliser un locuteur, nous avons seulement besoin de son propre signal de parole, mais, dans la discrimination de locuteurs, nous sommes toujours en présence de deux segments à comparer (généralement petits). Ces segments peuvent appartenir à un même locuteur ou à deux locuteurs différents. Ainsi, nous avons défini une nouvelle caractéristique utilisée à l'entrée des deux classifieurs et modélisant chaque locuteur par rapport à un autre (MLP et SVM), qui permet d'améliorer les performances des classifieurs. Cette caractéristique est appelée : Caractéristique Relative du Locuteur (RSC). De plus, en utilisant cette dernière, nous arrivons à réduire la taille du vecteur d'entrée des deux classifieurs et accélérer la phase d'apprentissage [35].

Soient à comparer deux segments de parole, appartenant aux locuteurs ***x*** et ***y***. Leurs matrices de covariance sont respectivement X et Y et leurs vecteurs moyenne sont respectivement x_m et y_m. La mesure de similarité entre eux est donnée par la formule (3.11).

$$\psi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\frac{\det(Y)}{\det(X)}) + tr(YX^{-1})\right] - 1 \tag{3.62}$$

Mais, du fait que

$$\frac{\det(Y)}{\det(X)} = \det(Y/X) \tag{3.63}$$

Alors

$$\psi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\det(Y/X) + tr(Y/X))\right] - 1 \tag{3.64}$$

Si nous introduisons le rapport de relativité $\Re$ par :

$$\Re(\boldsymbol{x},\boldsymbol{y}) = \frac{X}{Y} = X * Y^{-1}$$

Alors

$$\psi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log(\det(\Re(\boldsymbol{y},\boldsymbol{x})) + tr(\Re(\boldsymbol{y},\boldsymbol{x}))\right] - 1 \tag{3.65}$$

Donc, $\psi(\boldsymbol{x},\boldsymbol{y})$ est une fonction du rapport de relativité $\Re(\boldsymbol{y},\boldsymbol{x})$.

Nous avons appelé le rapport $\Re$: caractéristique relative du locuteur (*Relative Speaker Characteristic RSC*). L'acronyme DRSC représente la diagonale de la matrice RSC.

$$\text{DRSC}(\boldsymbol{x},\boldsymbol{y}) = \text{diag}(\Re(\boldsymbol{x},\boldsymbol{y})) \tag{3.66}$$

La nouvelle caractéristique DRSC, ainsi calculée, contient beaucoup d'informations normalement capables de discriminer entre deux énonciations différentes (figure 3.11). Les expériences effectuées sur des signaux de parole extraits de la base de données Hub4 Broadcast-News ont montré une bonne performance des RN en utilisant cette caractéristique, comparativement aux autres résultats obtenus dans le même corpus en utilisant : le vecteur moyenne, la diagonale de la covariance et les deux premiers vecteurs propres de la covariance.

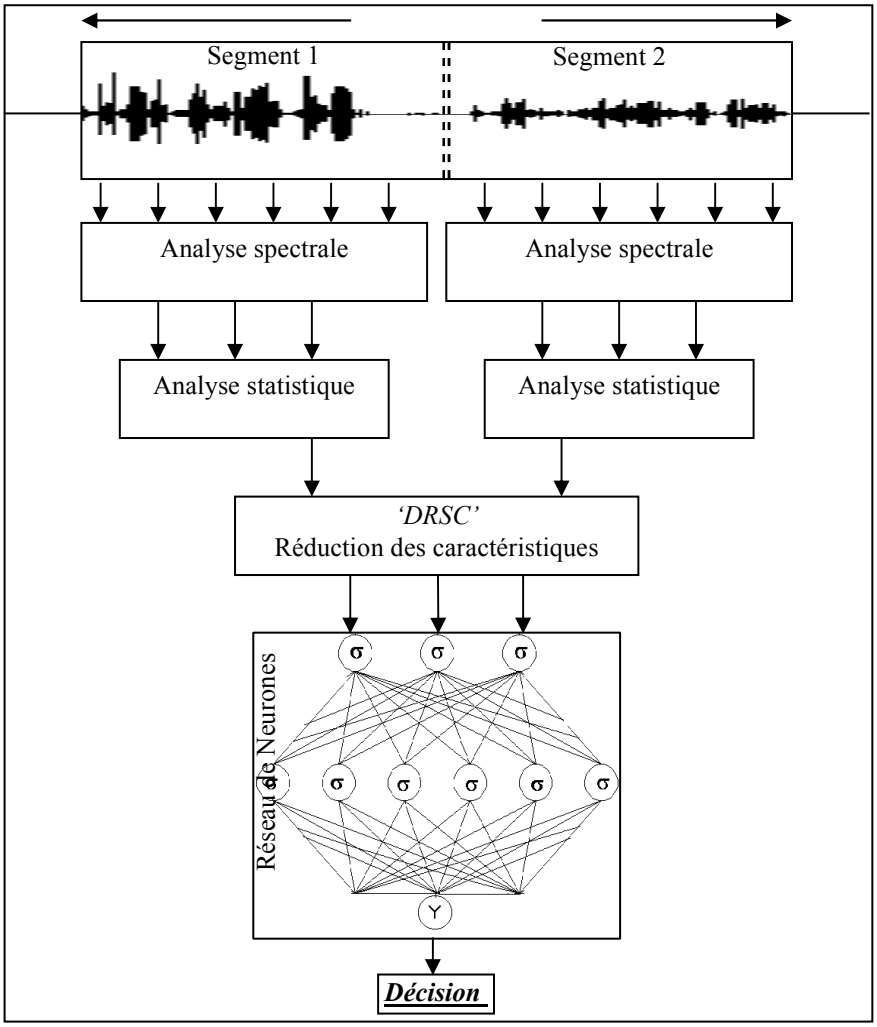

Figure 3.11: Principe du système de discrimination utilisant un RN basé sur la DRSC.

3.6. Fusion des Classifieurs

Dans le but d'améliorer les performances de l'indexation, nous avons combiné les différents classifieurs décrits précédemment en vue de diminuer l'erreur de discrimination et de segmentation. Cette combinaison est appelée Fusion.

En général, la fusion peut être utilisée à plusieurs niveaux de traitement [107], [108] et [109]. La fusion des données la plus courante est donnée par le niveau :

- **des caractéristiques** où les ensembles des caractéristiques des différentes modalités sont combinés. La fusion à ce niveau fournit une grande flexibilité, cependant des problèmes de classification peuvent survenir. Ceci est dû à la grande dimension des vecteurs caractéristiques combinés (concaténés) ;
- **des scores** provenant de plusieurs classifieurs sont d'habitude normalisés et puis combinés (c'est le type de fusion le plus utilisé) ;

- de décision où les sorties des différents classifieurs établissent la décision via des techniques comme le vote. La fusion au niveau de décision est considérée comme rigide pour l'intégration d'informations [110].

La combinaison des décisions partielles des différents classifieurs par une fusion des scores ou des décisions peut être réalisée en utilisant une architecture [107] et [108] :

- sérielle de fusion : qui est une combinaison en série des sorties d'un ensemble de différents classifieurs. Il s'agit en fait de concaténer les classifieurs (la sortie du premier classifieur sert comme une entrée pour le deuxième) ;
- parallèle de fusion : qui consiste à lier les sorties des différents classifieurs en parallèle. Les différentes sorties de ces classifieurs sont combinées en parallèle par un module de fusion [111].

Dans ce travail, nous avons fait des expériences sur quelques types de fusion résultant des trois classifieurs notamment : la μ_{Gc} (mesure statistique), le MLP et les SVM.

3.6.1. Fusion Sérielle

Du fait qu'il a été prouvé que les RN ont une excellente propriété discriminative [112], nous avons pensé ajouter la mesure statistique aux entrées du MLP dans le but d'améliorer ses performances. Ainsi, une entrée supplémentaire est prévue dans le MLP où sera injectée la mesure statistique (μ_{Gc}) pour chaque couple de segments de parole. Cette entrée est injectée avec la caractéristique DRSC, puis l'appprentissage du MLP est lancé avec le nouveau vecteur d'entrée (μ_{Gc} score et la caractéristique DRSC) (figure 3.12).

La fusion sérielle est exécutée comme suit : premièrement, les caractéristiques sont extraites des deux segments de parole à comparer, après la mesure statistique μ_{Gc} et le vecteur DRSC sont calculés et injectés ensemble à l'entrée du classifieur MLP. L'apprentissage est ainsi amélioré avec l'information apportée par l'approche statistique.

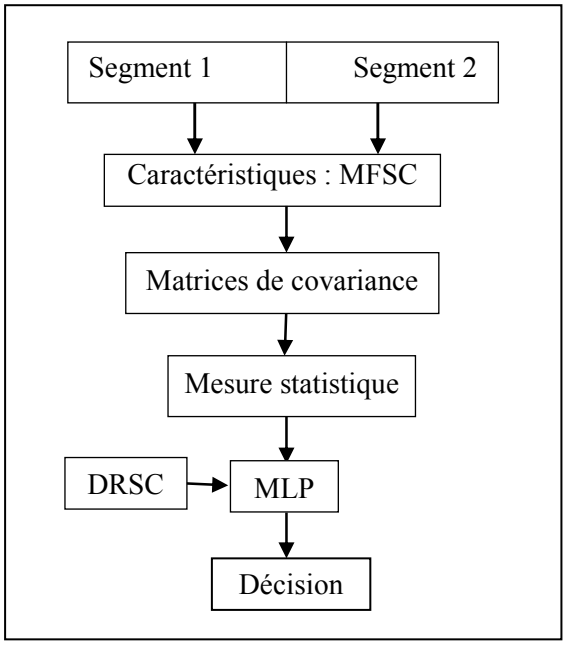

Figure 3.12 : Fusion sérielle entre les classifieurs statistique et neuronal.

3.6.2. Fusion Parallèle

La fusion parallèle est effectuée au niveau des scores obtenus par chaque classifieur à part, en utilisant une somme pondérée de ces derniers [113]. En effet, chaque classifieur est appliqué au couple de segments à comparer, dans le but d'avoir un résultat (score) de discrimination propre à chaque classifieur. Ensuite, les différents scores sont combinés pour donner la décision de discrimination finale.

Si les scores individuels sont notés par S_j, alors le score de fusion Sf est donné par :

$$Sf = \sum_{j=1}^{N} C_j S_j \tag{3.67}$$

où C_j est le coefficient de pondération pour le classifieur "j" et N représente le nombre de classifieurs. Le choix de ces coefficients est expérimental.

avec

$$\sum_{j=1}^{N} C_j = 1 \quad \text{et} \quad C_j \in]0, 1[\tag{3.68}$$

En pratique, C_j représente l'importance du classifieur j, plus le classifieur est précis, plus son coefficient de pondération est élevé.

Dans la figure 3.13, nous décrivons le principe de la somme pondérée utilisée pour combiner les classifieurs statistique et neuronal. Les sorties de ces classifieurs sont représentées respectivement par le score 1 et le score 2. Ces derniers sont multipliés respectivement par les poids C_1 et C_2, et puis la somme des scores pondérés fournit le résultat final de la fusion parallèle.

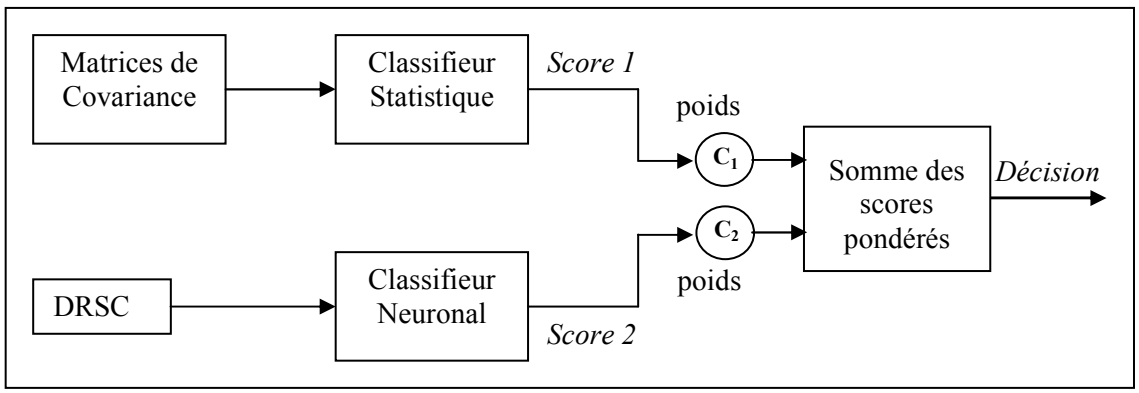

Figure 3.13 : Fusion parallèle (au niveau des scores) de deux classifieurs.

3.6.3. Fusion Sérielle-Parallèle

Dans la fusion sérielle-parallèle, nous considérons la fusion sérielle comme un troisième classifieur qui fournit une autre décision partielle (score sériel).

Après, avec une autre fusion parallèle, nous combinons tous les différents scores : le statistique, le neuronal et le résultat de la fusion sérielle, dans le but d'obtenir le résultat de la combinaison parallèle.

Le principe de la fusion sérielle-parallèle est résumé comme suit :

Premièrement, chaque classifieur est appliqué au couple de segments de parole pour obtenir la décision de la discrimination entre les deux segments à comparer; nous aurons donc deux décisions représentées par le score 1 pour le classifieur statistique et le score 2 pour le classifieur neural. Deuxièmement, la mesure statistique est injectée avec le vecteur DRSC à l'entrée du MLP dans le but d'obtenir la fusion sérielle qui sera représentée par le score 3. Troisièmement, les différents scores : score 1, score 2 et score 3 sont multipliés par leurs coefficients de pondération correspondants représentés respectivement par les coefficients C_1, C_2 et C_3. Les scores pondérés sont sommés pour fournir la décision finale de discrimination (figure 3.14).

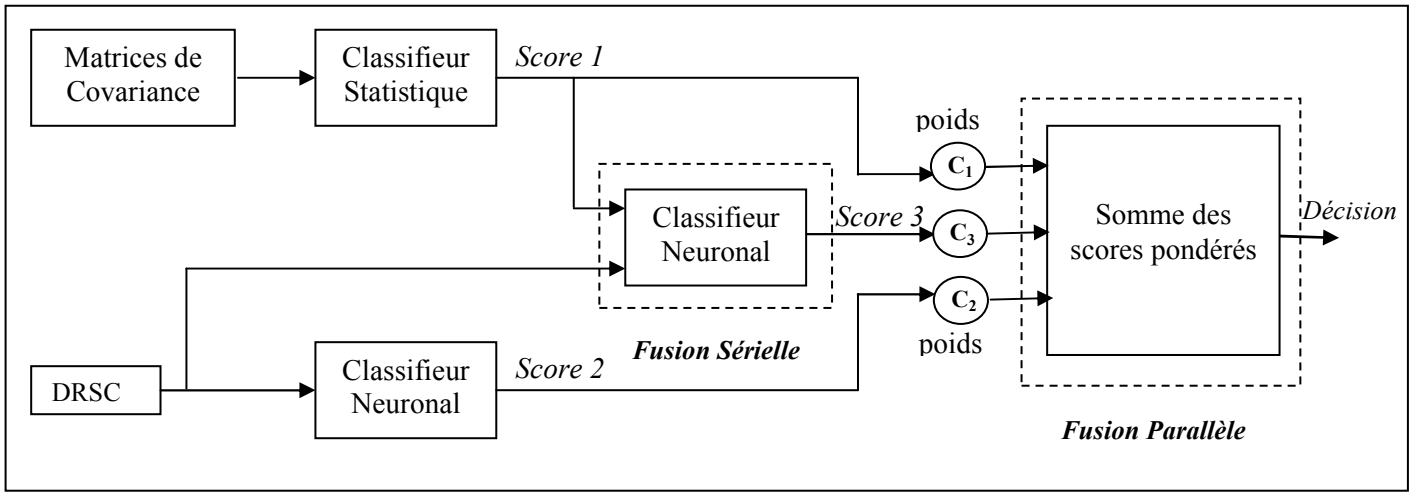

Figure 3.14 : Fusion sérielle-parallèle de deux classifieurs.

3.7. Algorithmes d'Indexation

Nous présentons les différents algorithmes que nous avons développés pour traiter l'indexation avec et sans connaissances a priori des locuteurs.

3.7.1. Indexation Avec Connaissances des Locuteurs

Dans ce type d'indexation, le système dispose des modèles des différents locuteurs présents dans le document audio à traiter. Pour ce faire, nous avons développé un algorithme basé sur une Indexation Entrelacée de la parole appelée ISI (*Interlaced Speech Indexing*).

3.7.1.1. Segmentation et Etiquetage du Signal Audio

Dans notre application, nous divisons le signal de parole en deux groupes de segments uniformes (équidistants). Chaque segment a une durée de 2 s. Le deuxième groupe de

segments est décalé par rapport au premier avec un retard de 1 s, c'est-à-dire les segments sont recouverts de 50% (figure 3.15). Ces deux groupes de segments, appelés respectivement séquence impaire et séquence paire, forment la segmentation entrelacée.

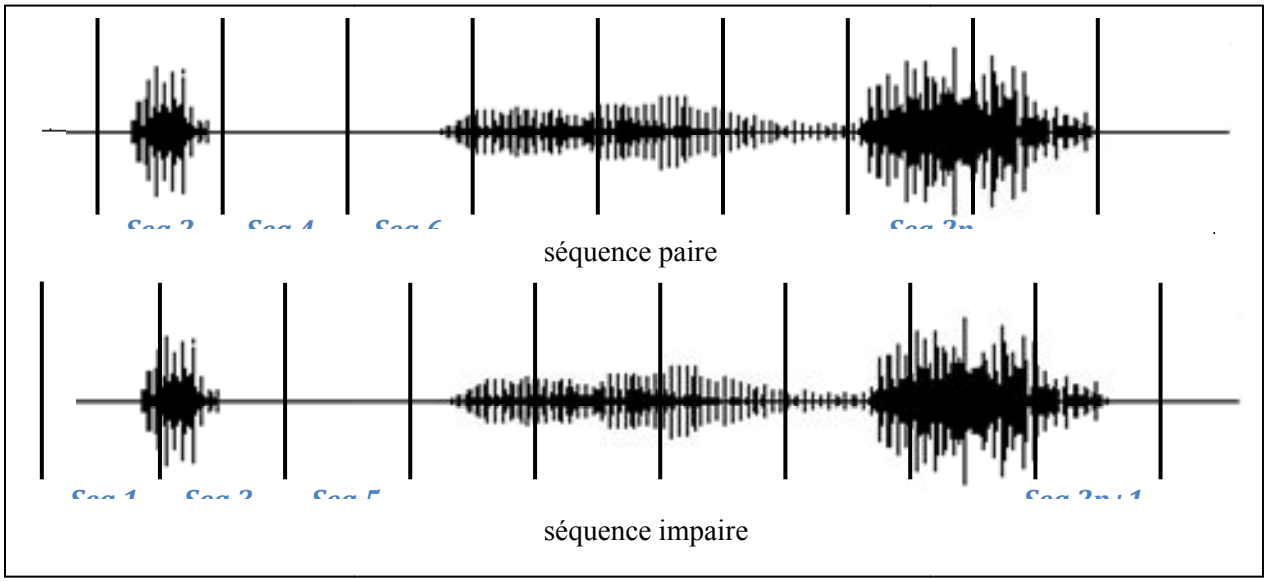

Figure 3.15 : Segmentation entrelacée de la parole (*Seg* signifie segment).

Le pré-traitement est constitué des étapes d'analyse du signal de parole précédant l'apprentissage et le test d'identification. Ainsi, chaque phrase est analysée selon les étapes suivantes :

- fenêtrage du signal de parole sur des trames de 32 ms (du type Rectangulaire) avec recouvrement de 50% (la fréquence d'échantillonnage Fe=16 kHz) ;
- évaluation du spectre d'énergie par une Transformée de Fourier Rapide "FFT" (*Fast Fourier Transform*) sur des trames d'analyse de 512 échantillons ;
- le spectre d'énergie passe ensuite à travers un banc de filtres constitué de plusieurs filtres du type Hamming (figure 3.16) ;
- la sortie de chaque filtre est calculée en sommant toutes les composantes fréquentielles à l'intérieur de la bande passante du filtre, pondérées par les coefficients du filtre, comme ce qui suit :

$$Y_k(m) = \sum_i c_{ik} Y^i(m), \tag{3.69}$$

où $Y_k(m)$ est la sortie du k^e canal du banc de filtres pour la $m^{ème}$ fenêtre. $Y^i(m)$ est l'énergie du signal (carré du module du spectre) dans le $i^{ème}$ coefficient de Fourier pour la $m^{ème}$ fenêtre. c_{ik} est le gain du $k^{ème}$ filtre dans le $i^{ème}$ échantillon fréquentiel et i est sommé sur la bande passante du filtre ;

- finalement les coefficients $Y_k(m)$ pour chaque trame "m" sont stockés dans des vecteurs de dimensions K, appelés les MFSC ou *Mel Frequency Spectral Coefficients*. Ce sont les caractéristiques du locuteur utilisées dans cette étude (figure 3.17).

Une fois la covariance est calculée pour chaque segment, quelques mesures de distance sont utilisées pour trouver la plus proche référence de chaque segment (dans un espace à 24 dimensions) (figures 3.18, 3.19 et 3.20).

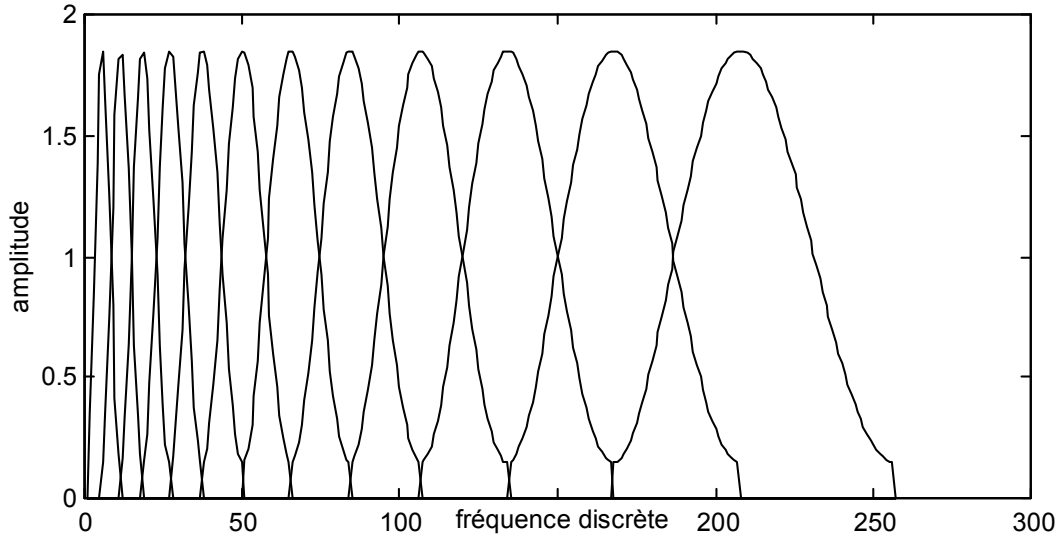

Figure 3.16 : Banc de Filtres du type Hamming (dans ce cas : 12 filtres).

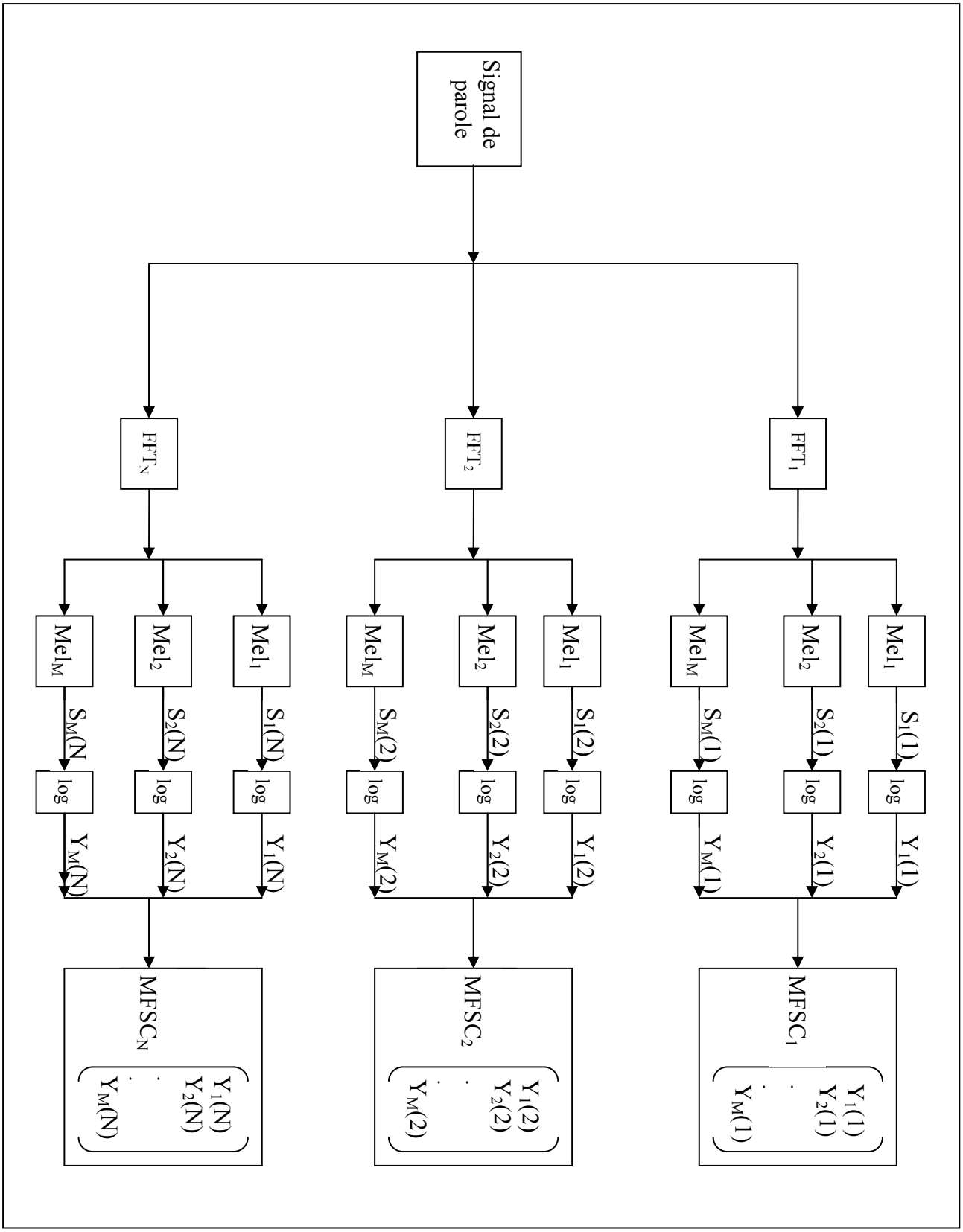

Figure 3.17 : Principe d'extraction des coefficients MFSC.

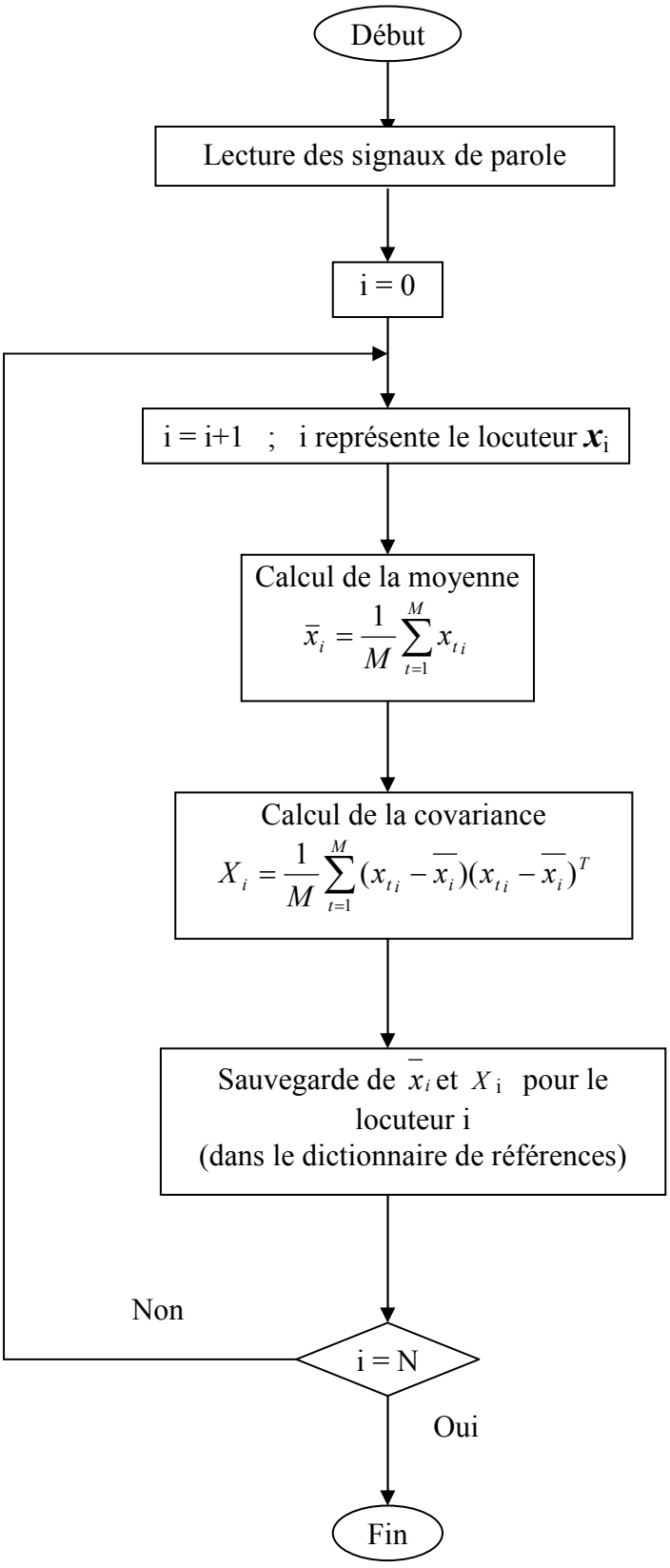

Figure 3.18 : Organigramme d'apprentissage dans le cas de l'indexation avec connaissances a priori des locuteurs.

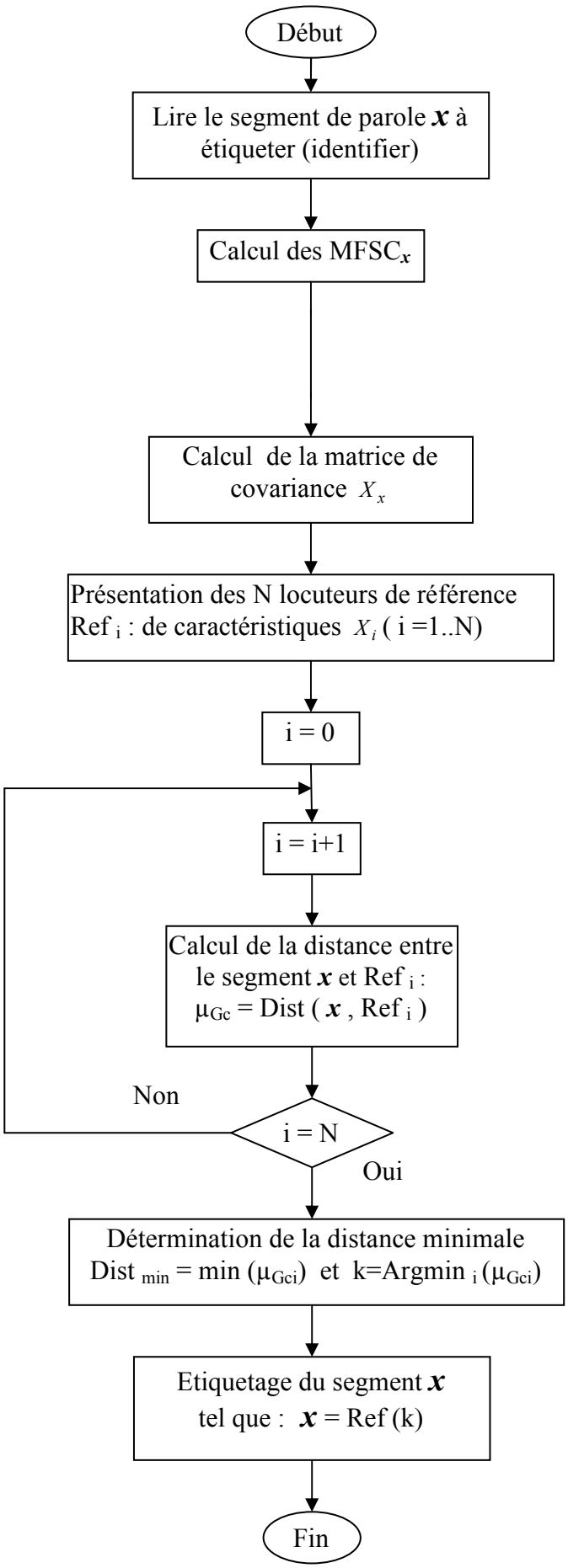

Figure 3.19 : ***Organigramme d'étiquetage des segments.***

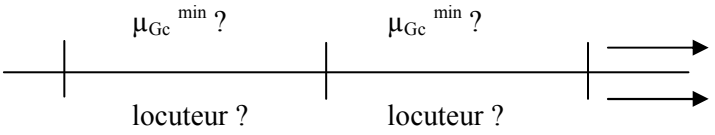

Figure 3.20 : Calcul de la distance minimale.

Une fois la distance minimale entre les caractéristiques du segment et celles de la référence, est trouvée (exemple correspondant au locuteur L_j). Ainsi, ce processus continue jusqu'au dernier segment du fichier de parole, le segment est étiqueté par l'identité de cette référence (locuteur L_j) (figure 3.21). Finalement, nous obtenons deux séquences étiquetées correspondant à un étiquetage pair et un étiquetage impair.

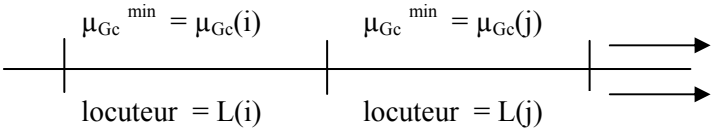

Figure 3.21 : Étiquetage des segments.

3.7.1.2. Indexation Entrelacée de la Parole : ISI

Nous avons développé une nouvelle technique appelée "Algorithme ISI". Cette technique basée sur deux segmentations (une déplacée par rapport à l'autre), utilise une combinaison logique intelligente pour trouver les étiquettes les plus appropriées au locuteur, en combinant les deux séquences de segmentation (figure 3.22) [114] et [115].

Ayant deux séquences de segmentation différentes, nous essayons alors de donner un compromis d'étiquetage raisonnable entre les deux séquences précédentes d'étiquetage. Ainsi, nous divisons chaque segment en deux autres segments similaires (de 1 s chacun), appelés sous-segments, donc nous obtenons "2n" étiquettes paires (notées par $L'^{1/2'}_{paire}$) pour les sous-segments pairs et "2n+1"étiquettes impaires (notées par $L'^{1/2'}_{impaire}$) pour les sous-segments impairs. Dans ce cas $L'^{1/2'}_{paire}$ et $L'^{1/2'}_{impaire}$ sont appelées sous-étiquettes.

Etant données que la sous-étiquette paire et la sous-étiquette impaire d'un même segment devrait être la même : pour cela, nous devrons comparer $L'^{1/2'}_{paire}(j)$ avec $L'^{1/2'}_{impaire}(j)$ pour chaque sous-segment j (pour j=2, 3, …, 2n+1). Ainsi , deux cas sont possibles :

- **si** $L'^{1/2'}_{paire}(j) = L'^{1/2'}_{impaire}(j)$ **alors** l'étiquette est correcte :

  étiquette nouvelle =étiquette correcte = $\mathbf{L'^{1/2'}(j) = L'^{1/2'}_{paire}(j) = L'^{1/2'}_{impaire}(j)}$ (3.69)

  où $L'^{1/2'}$ représente une sous-étiquette.

- **si** $L'^{1/2'}_{paire}(j) \neq L'^{1/2'}_{impaire}(j)$ **alors** confusion d'étiquette :

  étiquette nouvelle = $\mathbf{L'^{1/2'}(j) = Cf}$ (3.70)

  Où Cf signifie une confusion d'étiquetage.

Dans le cas de confusion (nouvelle étiquette = Cf), nous avons développé un nouvel algorithme de correction appelé "correction ISI".

Dans le cas de confusion, nous divisons le sous-segment (de 1 s) en deux autres sous-segments de 0.5 s chacun, appelés micro-segments. Leurs étiquettes, appelées micro-étiquettes, sont notées par $L^{'1/4'}$.

L'algorithme de correction est donc donné par :

- **si** { $L^{'1/4'}(j) = Cf$ **et** $L^{'1/4'}(j+1) = Cf$
  et $L^{'1/4'}(j-1) \neq Cf$ }
  alors
  $\mathbf{L^{'1/4'}(j) = L^{'1/4'}(j-1)}$ (3.71)

  Ceci est appelé une correction à gauche (figure 3.22),

- **si** { $L^{'1/4'}(j) = Cf$ **et** $L^{'1/4'}(j-1) = Cf$
  et $L^{'1/4'}(j+1) \neq Cf$ }
  alors

  $\mathbf{L^{'1/4'}(j) = L^{'1/4'}(j+1)}$ (3.72)

  Ceci est appelé une correction à droite (figure 3.22).

  Où, $L^{'1/4'}$ représente une micro-étiquette pour un micro-segment de 0.5 s.

La correction ISI peut être utilisée plusieurs fois (plusieurs itérations) pour raffiner progressivement la précision de l'indexation. Dans notre application, nous avons utilisé cet algorithme avec 2 et 4 itérations.

Les expériences ont montré que la correction ISI permet de trouver la meilleure décision d'étiquetage, à partir des deux séquences entrelacées étiquetées, dans la réduction efficace de l'erreur d'indexation. De plus, la résolution de segmentation (résolution de L (j)), qui était 2 s, est réduite à seulement 0.5 s (résolution de $L^{1/4}(j)$), ainsi la performance apportée par la technique ISI est observable, en même temps, dans la précision de l'indexation et dans la résolution de la segmentation.

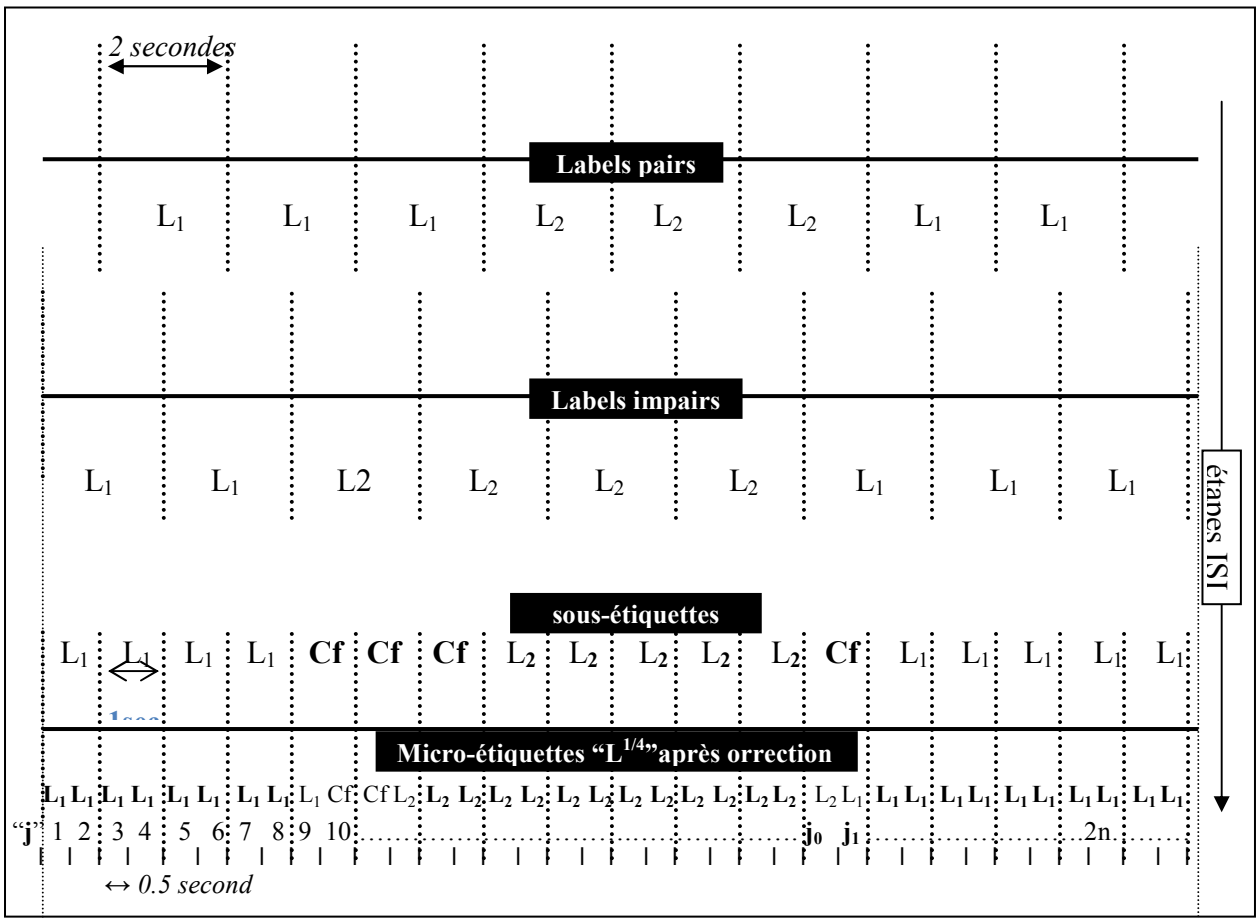

Figure 3.22 : Etapes de l'algorithme ISI avec une itération.
L_j : le locuteur "j" et Cf : Confusion.

3.7.1.3. Regroupement en Locuteurs

L'algorithme de regroupement dans ce cas est simple, puisqu'il consiste à regrouper les segments ayant les mêmes étiquettes, et le nombre de groupes est connu d'avance puisqu'il est égal au nombre de locuteurs connus déjà par le système.

3.7.2. Indexation Sans Connaissances a Priori du Locuteur

Cette indexation est appelée aveugle puisque le système ne dispose d'aucun modèle des locuteurs.

3.7.2.1. Segmentation du Signal Audio

L'algorithme de segmentation est basé sur une détection des points de rupture ou des points de changement de locuteur dans le document audio. Dans l'algorithme que nous avons développé, une fenêtre glissante de durée 8 s est appliquée tout au long du signal de parole. La fenêtre représente deux segments équidistants de 4 s chacun. Cette durée est choisie pour permettre un bon rendement des classifieurs.

L'idée est de discriminer entre ces deux segments en utilisant un classifieur donné (parmi les classifieurs cités auparavant) pour savoir s'il y a eu un changement de locuteur ou non :

ainsi, le point de rupture, s'il existe, sera le milieu de cette fenêtre (figure 3.2 3). Toutefois, le décalage entre deux points d'analyse consécutifs est 4 s (erreur de segmentation de 4 s), pour réduire cette durée, la fenêtre glisse avec une durée de 1 s ce qui correspond à un recouvrement des fenêtres de 75%. Ainsi, l'erreur de segmentation est divisée par 4 (1 s au lieu de 4 s). À la fin de la segmentation, nous obtenons des segments homogènes : chacun d'eux contient un seul locuteur à la fois, mais les segments ne sont pas encore étiquetés.

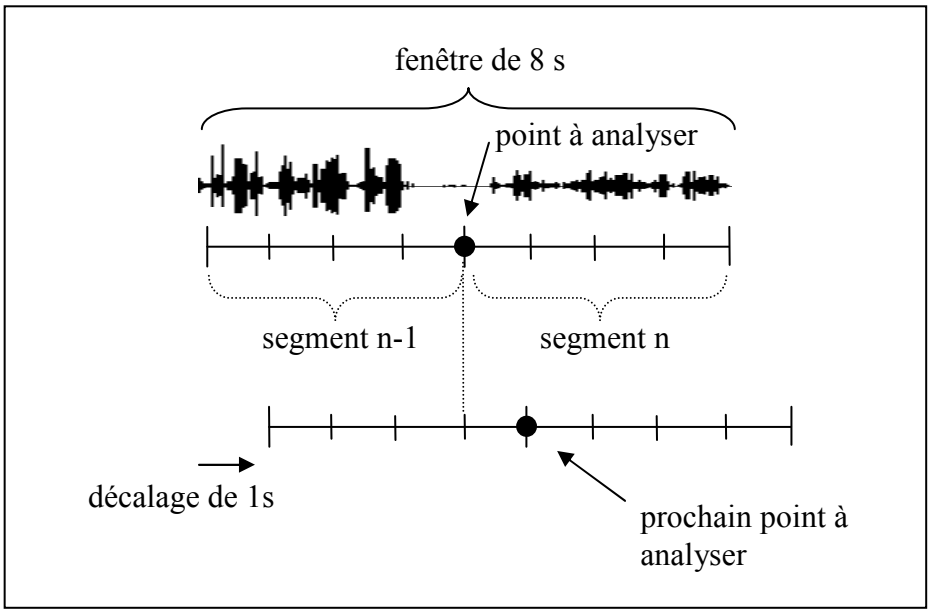

Figure 3.23 : Fenêtre glissante d'analyse.

L'algorithme de segmentation est basé sur une détection des points dans le document où il y a un changement de locuteur. Ces points sont aussi appelés points de ruptures. Selon le classifieur utilisé, la recherche de ces points consiste à comparer les caractéristiques acoustiques entre deux segments voisins et dire s'ils appartiennent à un même locuteur ou non. Chaque segment est choisi avec une durée 4 s, et codé par 37 coefficients MFSC. Le processus est exécuté tout le long du signal de parole.

Dans le cas du classifieur statistique, la discrimination entre les deux segments revient à un calcul d'une distance statistique mono gaussienne (μ_{Gc}) et la comparer à un seuil.

Tandis que dans le cas des classifieurs MLP et SVM, les deux classifieurs passent d'abord par une phase d'apprentissage leur permettant de reconnaître si les deux segments à comparer (présentés à l'entrée du classifieur) appartiennent au même locuteur ou non. Pour cela, nous avons préparé un nombre d'exemples contenant des segments provenant d'un même locuteur et d'autres de locuteurs différents. Une fois l'apprentissage terminé, le classifieur est appliqué tout le long du fichier audio pour la recherche des points de ruptures. Nous rappelons que pour modéliser les segments, nous avons développé une caractéristique réduite que nous avons appelée DRSC [116].

Une fusion de ces trois classifieurs, qui consiste à combiner les scores des classifieurs simples pour chaque couple de segments à comparer, est prévue pour améliorer davantage les résultats.

3.7.2.2. Regroupement en Locuteurs

L'algorithme choisi est basé sur le regroupement séquentiel, car il prend en considération les propriétés du voisinage, pour rassembler les différents segments homogènes obtenus lors de la tâche de segmentation. De plus, en utilisant les classifieurs MLP et SVM, le problème du seuil de décision est résolu du fait que ces classifieurs ne nécessitent pas de seuil. Toutefois, dans le cas de la mesure statistique [117], ce seuil demande à être réglé au préalable.

3.8. Conclusion

Dans ce chapitre, nous avons détaillé la théorie des différents classifieurs implémentés pour effectuer la tâche d'indexation, commençant par le classifieur mono-gaussien, le MLP et les SVM. Nous avons ensuite présenté la théorie de la RSC, la nouvelle caractéristique que nous avons développée pour optimiser l'apprentissage des classifieurs MLP et SVM. Par la suite nous avons expliqué les différentes architectures que nous avons proposées afin de fusionner les différents classifieurs. Dans cet objectif, nous avons implémenté une fusion sérielle, une fusion parallèle et sérielle-parallèle. Le dernier point dans ce chapitre a été consacré à l'exposition des différents algorithmes développés dans les deux cas d'indexation : avec et sans connaissance a priori des locuteurs.

Chapitre 4 :

EXPERIENCES ET RESULTATS DES DEUX TYPES D'INDEXATION

4.1. Introduction

Nous avons expérimenté deux types d'indexation, organisés en cinq séries d'expériences. Ces dernières sont effectuées dans le but d'arriver à une indexation performante des signaux audio, en segments de locuteurs homogènes et avec la meilleure précision possible. Ces expériences sont organisées de la façon suivante :

- **Indexation avec connaissances a priori des locuteurs**, qui correspond à l'expérience N° 1 (Exp1), où nous avons proposé une nouvelle technique d'indexation appelée technique ISI [118] ;
- **Indexation sans aucune connaissance des locuteurs**, correspondant à quatre expériences :
  - Exp 2 concerne le test des différentes caractéristiques réduites en caractérisation du locuteur ;
  - Exp 3, dans laquelle une discrimination automatique des locuteurs est appliquée sur des signaux de parole différents ;
  - Exp 4 effectue une segmentation, sans connaissances a priori des locuteurs, sur un flux audio télédiffusé réel ;
  - Exp 5 concernant le regroupement des différents segments appartenant à un même locuteur (ayant été trouvés dans l'expérience N° 4) en un unique segment appelé "final cluster ".

4.2. Indexation Avec Connaissances a Priori des Locuteurs (Exp 1)

En indexation avec connaissances a priori du locuteur, le système d'indexation possède les modèles de chaque locuteur à indexer, dans un dictionnaire de références. Cette indexation est moins complexe que l'indexation sans connaissances de modèle. Parmi les applications connues de ce type d'indexation, nous citons le suivi du locuteur [119], qui permet de localiser les paroles d'une personnalité particulière appelée "cible", dans un flux audio. Toutefois, en segmentation, nous rencontrons souvent, un problème de confusion surtout si le signal est bruité ou s'il contient de la parole provenant de différents canaux de transmission (microphone, téléphone, etc.), comme c'est le cas des signaux audio issus de HUB-4 Broadcast-News 1996. Pour régler ces problèmes de confusion, nous avons développé un algorithme basé sur une indexation entrelacée et utilisant des mesures statistiques du 2ème ordre.

4.2.1. Corpus d'Indexation

Concernant l'Exp1, pour évaluer l'algorithme d'Indexation Entrelacée (ISI) que nous avons proposé, la BD utilisée est extraite de la BD HUB-4 Broadcast-News. Cette dernière contient un enregistrement des informations télédiffusées 1996 [9]. Elle couvre au total, 104 heures d'informations recueillies des chaînes de télévision : ABC, CNN, et des chaînes de Radio : NPR/PRI [19] et [120].

La BD utilisée dans nos expériences consiste en des informations enregistrées de la chaîne CNN (informations et interviews d'environ 30 mn). Notre enregistrement correspond à l'intervention de 19 locuteurs différents avec plusieurs changements de locuteur dans le fichier. Cette BD est très diversifiée parce qu'elle contient : de la parole, du bruit, de la musique, des silences et plusieurs types d'autres sons. D'autre part, elle représente bien l'aspect naturel des discussions en mode multilocuteur, dans une conférence (ou téléconférence) réelle.

4.2.2. Protocole Expérimental

Nous avons testé différentes mesures et différents algorithmes de correction dans le but de les comparer objectivement (tab 4.1 et tab 4.2).

Tableau 4.1 : Taux d'erreurs d'indexation avec différentes mesures (avec 2 corrections ISI).

Durée du segment (s)	Erreur (%)		
	Mesure		
	$\mu_{Gc\beta}$	$\mu_{Gc0.5}$	μ_{Gc1}
1	21,4	36,7	15,4
2	07,8	11,5	11,1
3	07,7	08,6	10,4
4	08,8	08,9	11,9
6	09,5	10,5	14,5

Le tab 4.1 représente l'erreur d'indexation, dans HUB-4, obtenue avec les différentes mesures et pour les différentes durées du segment, avec μ_{Gc1} (mesure non-symétrique), $\mu_{Gc0.5}$ et $\mu_{Gc\beta}$ (mesures symétriques).

D'après ces résultats, nous remarquons que la mesure $\mu_{Gc\beta}$ donne la meilleure performance d'indexation. Par exemple, si la durée du segment est 3 s, les mesures μ_{Gc1} et $\mu_{Gc0.5}$ donnent respectivement une erreur de 10.4% et 8.6%, tandis que la $\mu_{Gc\beta}$ donne la plus petite erreur, qui est égale à 7.7%.

Dans une autre présentation, le tableau 4.2 représente l'erreur d'indexation dans le même corpus, obtenue avec et sans l'utilisation de l'algorithme de correction ISI et pour différentes durées du segment, dans le but d'avoir une comparaison globale. Nous remarquons que

l'erreur d'indexation après la correction ISI est plus petite que celle obtenue sans correction (figure 4.1). Par exemple, pour une durée de segment égale à 3 s, l'erreur sans correction ISI est d'environ 9%, mais elle devient 7.7% quand une correction ISI à deux itérations est appliquée et peut diminuer jusqu'à 7.6% avec une correction à quatre itérations.

Tableau 4.2 : Taux d'erreurs utilisant $\mu_{Gc\beta}$ avec et sans correction ISI.

Durée du segment (s)	Taux d'erreur (%) avec et sans correction ISI		
	Sans	**2 corrections**	**4 corrections**
1	26,0	21.4	19.7
2	09,0	08,0	07,7
3	09,0	07,7	07,6
4	09,9	08,8	08,7
6	09,6	09,5	09,4

La figure 4.1 représente l'erreur d'indexation, dans Hub4, obtenue avec différentes mesures et pour différentes durées du segment.

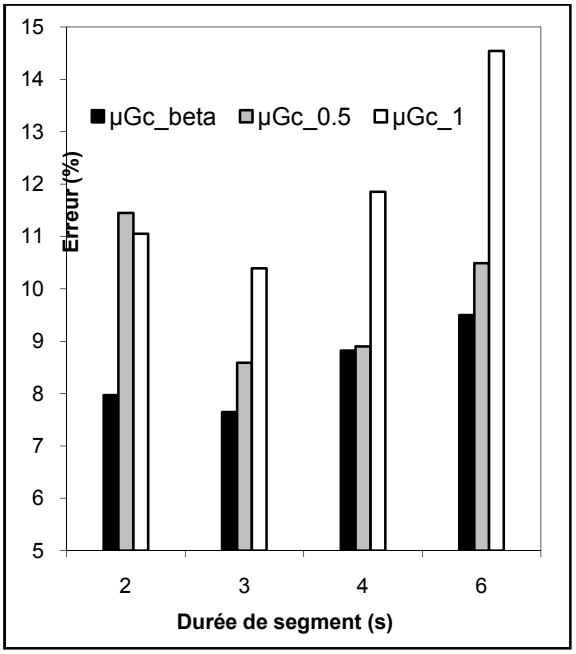

Figure 4.1 : Erreur d'indexation pour différentes mesures statistiques.

Une présentation graphique est donnée dans la figure 4.2 représentant l'erreur d'indexation dans Hub4, obtenue avec et sans correction ISI et pour différentes durées du segment.

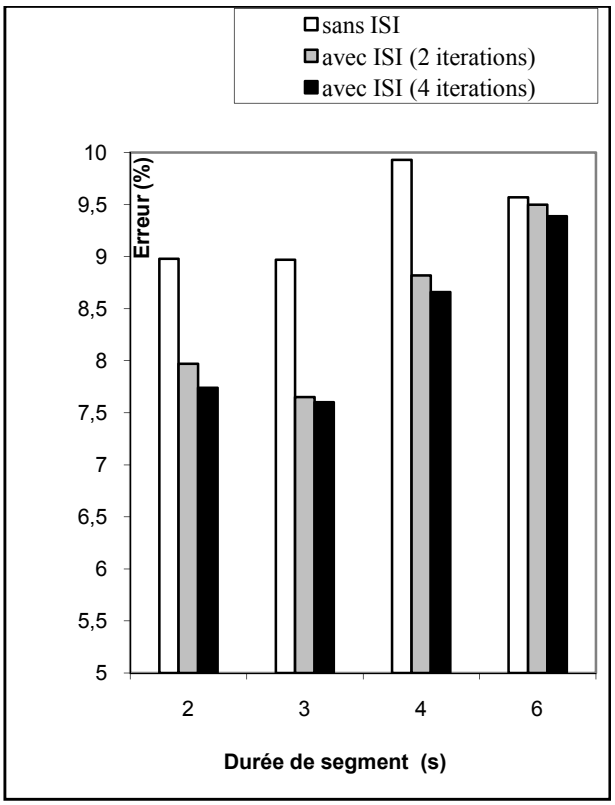

Figure 4.2 : Erreur d'indexation avec et sans correction ISI.

Dans le but de trouver la meilleure durée du segment pour l'indexation, nous avons représenté simultanément les différents taux d'erreurs dans la figure 4.3. Nous voyons une comparaison entre les différents résultats obtenus avec les différentes durées (1, 2, 3, 4 et 6 secondes) et nous remarquons que la plus petite erreur d'indexation est obtenue pour une durée de segment de 3 secondes.

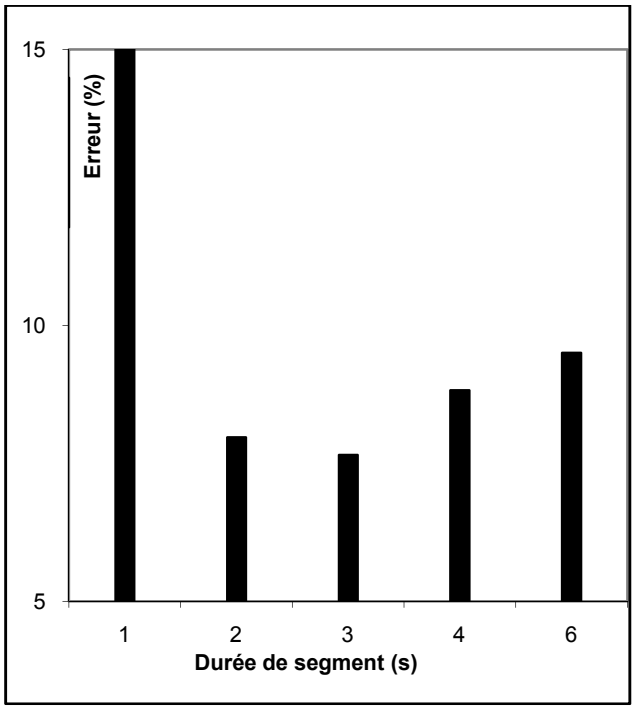

Figure 4.3 : Erreur d'indexation pour différentes longueurs de segment.

4.2.3. Résultats Obtenus de l'Indexation Avec Connaissances des Locuteurs

Durant l'Exp1, nous avons développé une technique d'indexation entrelacée ISI. L'expérience montre que l'association entre les mesures statistiques SOSM et la technique ISI, est utilisée efficacement, comme nous pouvons le constater dans les points suivants :

- bien que les mesures SOSM nécessitent une durée de segments d'au moins 2 s, ce qui veut dire que la résolution de la segmentation est d'environ 2 s, l'association SOSM-ISI permet une résolution plus fine en réduisant l'erreur de segmentation à seulement 0.5 s ;
- de plus, cette nouvelle approche améliore considérablement la précision d'indexation en corrigeant les erreurs de confusion. Quand le nombre de corrections dans l'algorithme de correction ISI (nombre d'itérations) augmente, l'erreur d'indexation diminue continûment.

Par la suite, les expériences effectuées sur HUB-4 Broadcast-News, indiquent que la meilleure mesure statistique est la $\mu_{Gc\beta}$ [94]) et que la meilleure durée des segments pour l'indexation est de 3 s.

Cette méthode fournit des résultats satisfaisants. De plus, la technique ISI est simple à implémenter, non coûteuse en temps de calcul et fournit une assez bonne performance d'indexation.

4.3. Indexation Sans Aucune Connaissance des Locuteurs

Durant ce deuxième type d'indexation, nous avons effectué quatre expériences différentes.

4.3.1. Exp 2 : Test des Différentes Caractéristiques Réduites en Caractérisation du Locuteur

Cette série d'expériences est très importante puisqu'elle permet de trouver les caractéristiques réduites les plus pertinentes pour caractériser un locuteur. Ainsi, nous avons, élaboré une nouvelle caractéristique appelée DRSC que nous comparons à d'autres caractéristiques qui existent déjà.

La BD concernée par cette expérience est un sous-ensemble (DB1) de "HUB-4 Broadcast-News 1996", contenant des enregistrements de la chaîne CNN (*CNN early edition*) et composé de signaux de parole, de la musique, des appels téléphoniques, du bruit, etc. La fréquence d'échantillonnage, Fe est égale à 16 kHz. Les signaux de parole sont extraits et arrangés en segments d'environ 4 s chacun.

DB1 contient 14 locuteurs différents (dont la plupart sont des journalistes) organisés en 259 combinaisons de locuteurs (interlocuteur et intralocuteur) pour la phase d'apprentissage et 195 combinaisons de locuteurs (interlocuteur et intralocuteur) pour le test.

La deuxième BD (appelée TB1) est un sous-ensemble extrait de la BD "Call Home" contenant des enregistrements de communications téléphoniques réels avec une Fe de 8 kHz. La durée de chaque segment est d'environ 10 s.

TB1 contient 24 locuteurs différents : 12 hommes et 12 femmes (parlant par téléphone sur différents sujets). Elle est organisée en 670 combinaisons de locuteurs pour l'apprentissage et 334 combinaisons pour le test.

Dans le but de tester les différentes caractéristiques et de trouver la meilleure caractéristique réduite [96] pour la tâche de discrimination, nous avons effectué différentes expériences de comparaison. Nous avons utilisé pour la tâche d'évaluation, certains taux d'erreurs connus. Les définitions de ces derniers sont données ci-après :

- Fausses Alarmes (*False Alarms FA*) : représente les erreurs en cas où le système décide que les deux segments de parole (à comparer) n'appartiennent pas au même locuteur, alors qu'ils sont réellement prononcés par la même personne. Son expression est donnée par :

$$FA = \frac{\text{NbFA}}{\text{NbCInterloc} + \text{NbFA}} \tag{4.1}$$

avec :

NbFA : Nombre de Fausses Alarmes ;

NbCInterloc : Nombre de Combinaisons Interlocuteurs.

- Détections Manquées (*Missed Detections MD*) : représente les erreurs commises par le système dans le cas où ce dernier ne peut pas détecter la différence entre deux segments de parole appartenant à deux locuteurs différents. Son expression est donnée par :

$$MD = \frac{\text{NbDM}}{\text{NbCInterloc}} \tag{4.2}$$

avec :

NbDM : Nombre de Détections Manquées.

- Taux d'erreur égal (*Equal Error Rates : EER*) : représente l'erreur de discrimination quand le taux des FA est égal au taux des MD. Donc le EER est égal aux deux taux FA et MD en ce moment.

$$EER = FA = MD \text{ (quand FA = MD)}$$

Les résultats sont représentés par les figures 4.4 et 4.5, et les tableaux 4.3 et 4.4.

Le tableau 4.3 expose les taux d'erreurs EER avec leur nombre d'itérations correspondant, nécessaire pour l'apprentissage du MLP, et ayant permis d'obtenir ces valeurs d'erreurs.

Ces taux d'erreurs sont obtenus par comparaison avec plusieurs caractéristiques du locuteur, qui sont : DRSC [116], diagonale de la covariance, vecteur moyenne et les 2 premiers vecteurs propres de la covariance. Les résultats montrent que le MLP utilisant

comme entrée la caractéristique DRSC donne la meilleure performance avec un EER de 7.20% et un nombre minimal d'itérations pour l'apprentissage de l'ordre de 1000 à 1500 itérations (tableau 4.3) et (figure 4.4).

Tableau 4.3 : Taux EER obtenus dans DB1, avec les différentes caractéristiques, (NIA : Nombre d'Itérations Approximatif).

Caractéristique	NIA nécessaire pour l'apprentissage	EER %
DRSC	1000 < NIA <1500	07.20
Diagonale de la covariance	3500 < NIA < 4000	13.90
Vecteur moyenne	6500 < NIA < 7000	25.19
Les 2 premiers vecteurs propres de la covariance	2500 < NIA < 3000	33.67

Nous avons refait les mêmes expériences, mais cette fois-ci sur la BD téléphonique TB1, avec une durée de 10 s pour chaque segment de parole.

Le tableau 4.4 résume les différents résultats, où nous donnons pour chaque expérience, le nombre d'itérations approximatif nécessaire pour l'entraînement du MLP, le coefficient de convergence et l'erreur de discrimination représentée par le EER.

Les résultats confirment la bonne performance du MLP utilisant la caractéristique relative DRSC comme entrée, comparativement aux autres caractéristiques testées dans le même corpus. Cette nouvelle caractéristique associée au MLP à 2 couches cachées, donne un EER de 4.65%. Les autres caractéristiques, testées dans les mêmes conditions, nécessitent un nombre d'itérations beaucoup plus grand pour la phase d'apprentissage du MLP, comme dans le cas du vecteur moyenne.

Tableau 4.4 : Performances obtenues dans TB1, avec les différentes caractéristiques.

Caractéristique	NIA nécessaire pour l'apprentissage	Coefficient de convergence	EER %
DRSC	$1*10^4$	0.010	04.65
Diagonale de la covariance	$2*10^4$	0.005	10.94
Vecteur moyenne	$2*10^5$	0.001	07.01
Les 2 premiers vecteurs propres de la covariance	$1*10^5$	0.001	17.50

Avec une meilleure représentation des résultats de discrimination, les figures 4.4 et 4.5, donnent respectivement, les différentes courbes d'erreurs ROC (*Receiver-Operating-Characteristic*) pour les différentes caractéristiques évaluées sur les bases de données DB1 et TB1. Nous remarquons que le MLP utilisant la caractéristique DRSC donne les meilleures

performances, du fait que le EER a été réduit considérablement, suivie par la diagonale de la covariance (ou le vecteur moyenne). Les 2 vecteurs propres donnent le plus mauvais résultat.

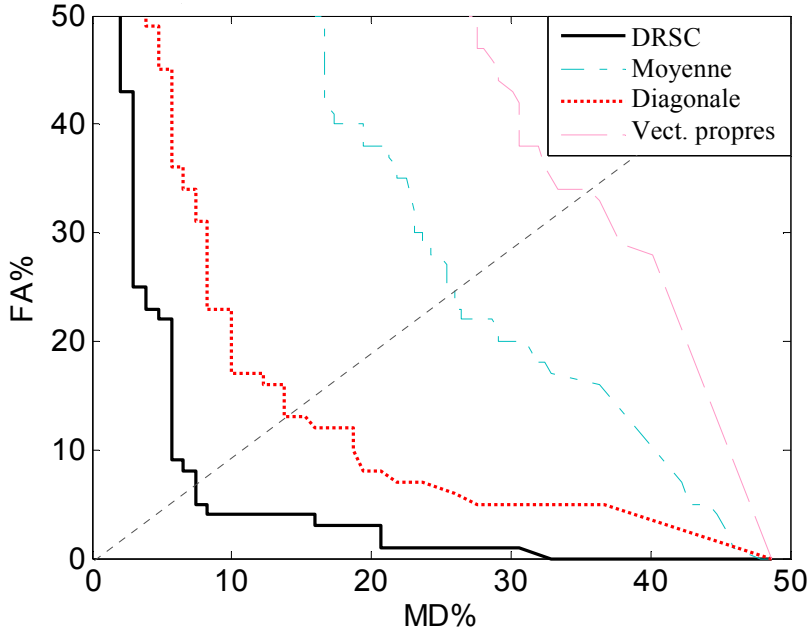

Figure 4.4 : Erreurs de discrimination de locuteurs dans DB1, Comparaison des différentes caractéristiques : DRSC, Vecteur moyenne, Diagonale de la covariance et les 2 premiers vecteurs de propres de la covariance.

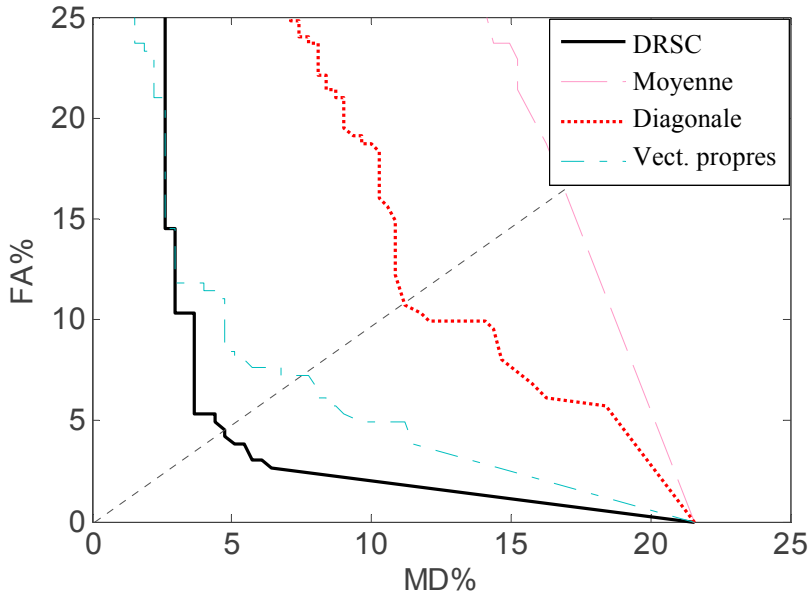

Figure 4.5 : Erreurs de discrimination de locuteurs dans TB1, Comparaison des différentes caractéristiques : DRSC, Vecteur moyenne, Diagonale de la covariance et les 2 premiers vecteurs propres de la covariance.

4.3.2. Exp 3 : Discrimination des Locuteurs

La discrimination consiste à vérifier si deux segments donnés appartiennent au même locuteur ou non.

Nous décrivons les différents résultats obtenus lors de cette expérience (expérience de discrimination) et qui sont effectuées sur 4 BD, en deux sous-ensembles :

- DB1 et DB2 extraits de "HUB-4 Broadcast-News 96", contenant des enregistrements de la chaîne CNN (*CNN early edition*) et composés de parole propre, musique, appels téléphoniques, bruits, etc ;
- TB1 et TB2 des enregistrements de communications téléphoniques.

Dans toutes les BD, les exemples de test sont différents de ceux de l'apprentissage. Les BD sont organisées en plusieurs combinaisons de locuteurs.

Habituellement, la reconnaissance du locuteur est plus difficile en téléphonie, ceci est dû à la limitation de la BP et la distorsion du signal par le canal de transmission. Cependant, la présence de bruit et de musique dans HUB-4 Broadcast-News rend aussi la tâche de discrimination difficile [32].

Les résultats correspondants aux expériences de discrimination du locuteur sont représentés dans le tableau 4.5.

Tableau 4.5 : Erreurs de discrimination obtenues dans les 4 BD : DB1, DB2, TB1 et TB2, avec FP : Fusion Parallèle et FS : Fusion Sérielle.

Type de Classifieur	EER % dans Broadcast-News avec des segments de 4 s		EER % dans les appels téléphoniques avec des segments de 10 s	
	DB1	DB2	TB1	TB2
Classifieur Statistique (μ_{Gc})	11.75	11.75	5.74	5.74
MLP (avec l'entrée DRSC)	07.20	09.04	5.02	3.83
SVM (avec l'entrée DRSC)	08.63	08.84	3.75	3.65
FS (hybride) : MLP/ μ_{Gc}	07.70	09.95	4.65	3.28
FP : MLP/ μ_{Gc}	06.77	07.92	4.29	3.65
Fusion Sérielle-Parallèle	06.77	07.92	4.29	3.28
FP : SVM/ μ_{Gc}	09.04	08.84	3.75	3.65
FP : MLP/ SVM	06.77	08.59	3.75	3.28
FP : MLP/ μ_{Gc} / SVM	06.77	07.92	3.56	3.28

Concernant les performances des classifieurs simples, les résultats montrent que la meilleure performance est donnée par le classifieur SVM, à l'exception de la base DB1. Les différents EERs sont 8.63%, 8.84%, 3.75% et 3.65% (obtenus respectivement sur DB1, DB2,

TB1 et TB2). Le classifieur MLP est moins précis que les SVM mais il donne de meilleurs résultats que le classifieur statistique (tableau 4.5).

Pour les différentes techniques de fusion entre les deux classifieurs μ_{Gc} et MLP [121] [122], nous pouvons faire les remarques suivantes :

- la fusion sérielle (hybride) MLP/μ_{Gc} n'a pas amélioré les performances de discrimination dans les bases de HUB-4 : DB1 et DB2 (EER de 7.70% et 9.95% dans respectivement DB1 et DB2), mais elle a montré un bon comportement, comparativement aux classifieurs simples, dans TB1 et TB2 (4.65% et 3.28% dans TB1 et TB2). D'autre part, la fusion parallèle a diminué l'erreur de discrimination sur toutes les BD, par rapport à celle fournie par les classifieurs individuels ;
- les meilleurs résultats sont donnés par la fusion sérielle-parallèle. Ils sont équivalents à ceux donnés par la fusion parallèle dans DB1, DB2 et TB1, et à ceux obtenus par la fusion sérielle dans TB2 ;
- la comparaison entre les différentes architectures de fusion montre que la fusion parallèle est meilleure que la fusion sérielle entre les deux classifieurs μ_{Gc} et MLP. Dans les trois BD : DB1, DB2 et TB1, tandis que dans TB2 la fusion sérielle est meilleure. Dans la BD téléphonique, la fusion sérielle paraît intéressante, mais dans Broadcast-News (avec des segments courts) la fusion parallèle paraît meilleure. Dans l'ensemble, la fusion sérielle-parallèle reste la plus intéressante.

Concernant la fusion entre les trois classifieurs (μ_{Gc}, MLP et SVM), nous constatons que la fusion améliore la précision de discrimination par rapport aux classifieurs simples, et la fusion parallèle entre les trois classifieurs (μ_{Gc}, MLP et SVM) reste la plus efficace sur toutes les DB testées. Cette dernière donne les meilleures performances (figure 4.6).

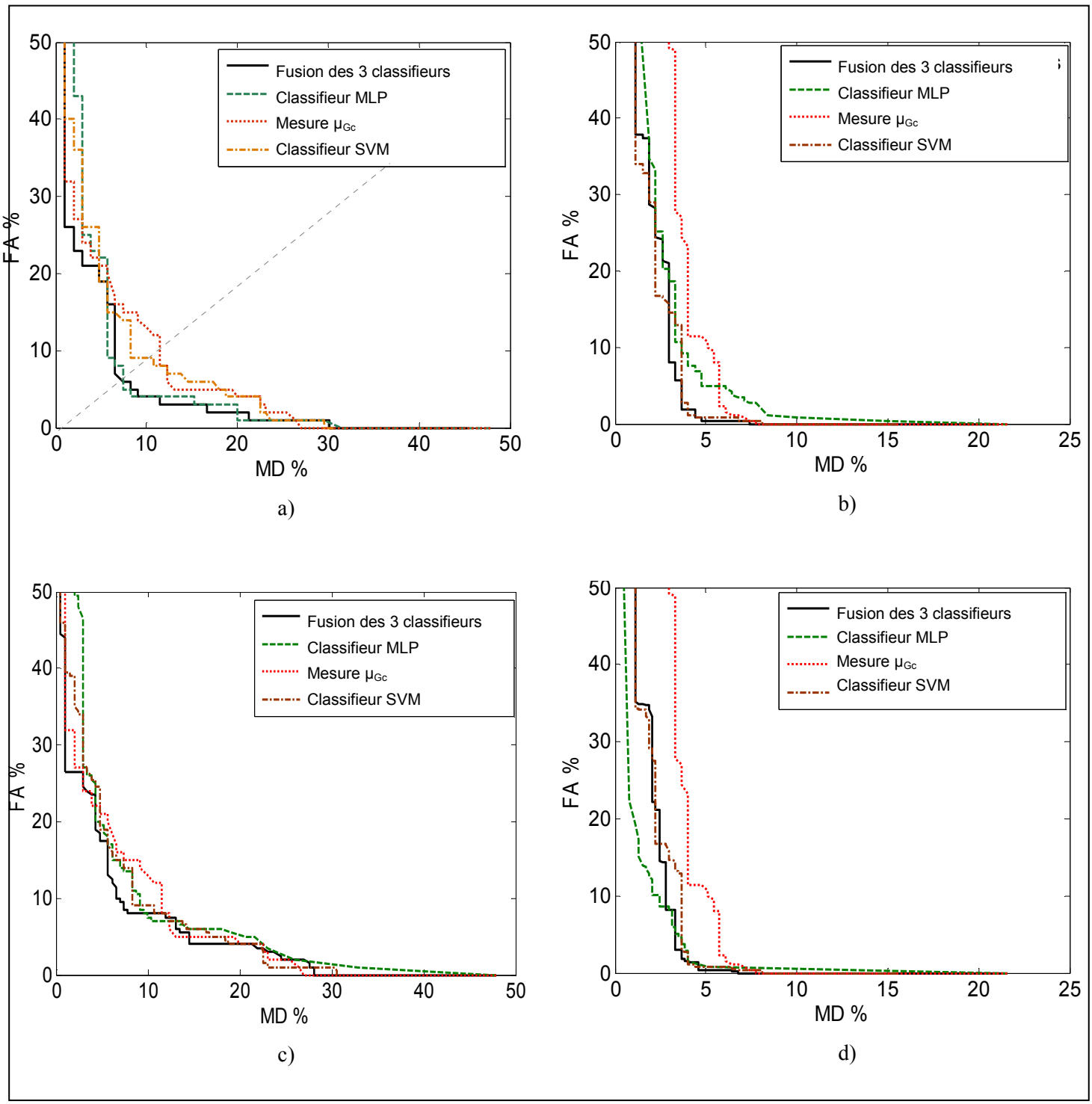

Figure 4.6 : Courbes ROC de discrimination de locuteurs dans : a) DB1, b) DB2, c) TB1, d) TB2.

4.3.3. Exp 4 : Segmentation Sans Connaissances a Priori des Locuteurs

Dans cette expérience, nous essayons de segmenter un flux audio réel en des segments homogènes appartenant à un seul locuteur chacun. Cette segmentation est effectuée sur un enregistrement réel extrait de HUB-4 Broadcast-News 96 [18], sans aucune connaissance au préalable des identités des locuteurs participant à la discussion et ni même leur nombre. Il s'agit des programmes "*ABC News Nightline*" télédiffusés avec une durée d'environ 30 mn, où nous pouvons trouver des conversations, des débats [32], des spots publicitaires, de la musique, et d'autres types de sons. L'enregistrement concerne 12 locuteurs principaux, parlant à tour de rôle (séquentiellement). Nous rencontrons des locuteurs qui parlent avec une bonne, moyenne et mauvaise qualité de parole (c'est-à-dire mixée avec de la musique).

Chaque phase de l'expérience concerne une méthode particulière et les résultats sont représentés dans la figure 4.7. Cette figure illustre les courbes ROC qui représentent le taux de Fausses Alarmes (FA) en fonction du taux de Détections Manquées (MD). Ces courbes sont obtenues par les différents classifieurs et les fusions décrites précédemment.

D'après la figure 4.7, nous pouvons faire les remarques suivantes :

- la comparaison entre les trois classifieurs simples : classifieur statistique (μ_{Gc}), MLP et SVM, montre que la performance de la segmentation obtenue par les SVM (EER=21.24%) (figure 4.7.b) est meilleure que les performances obtenues par le classifieur statistique (EER=30.48%) et le MLP (EER=27.2%) (figure 4.7.a) ;
- la fusion parallèle (figure 4.7.a), effectuée au niveau des scores, est plus intéressante que la fusion sérielle (figure 4.7.c) ;
- la courbe ROC de la fusion parallèle entre les 3 classifieurs est meilleure que les courbes ROC des classifieurs simples (figure 4.7.d), même si le EER donné par les SVMs est le même que celui obtenu par la fusion. Cette dernière courbe est proche des axes, ce qui montre que la technique de fusion cause moins d'erreurs que les classifieurs individuels.

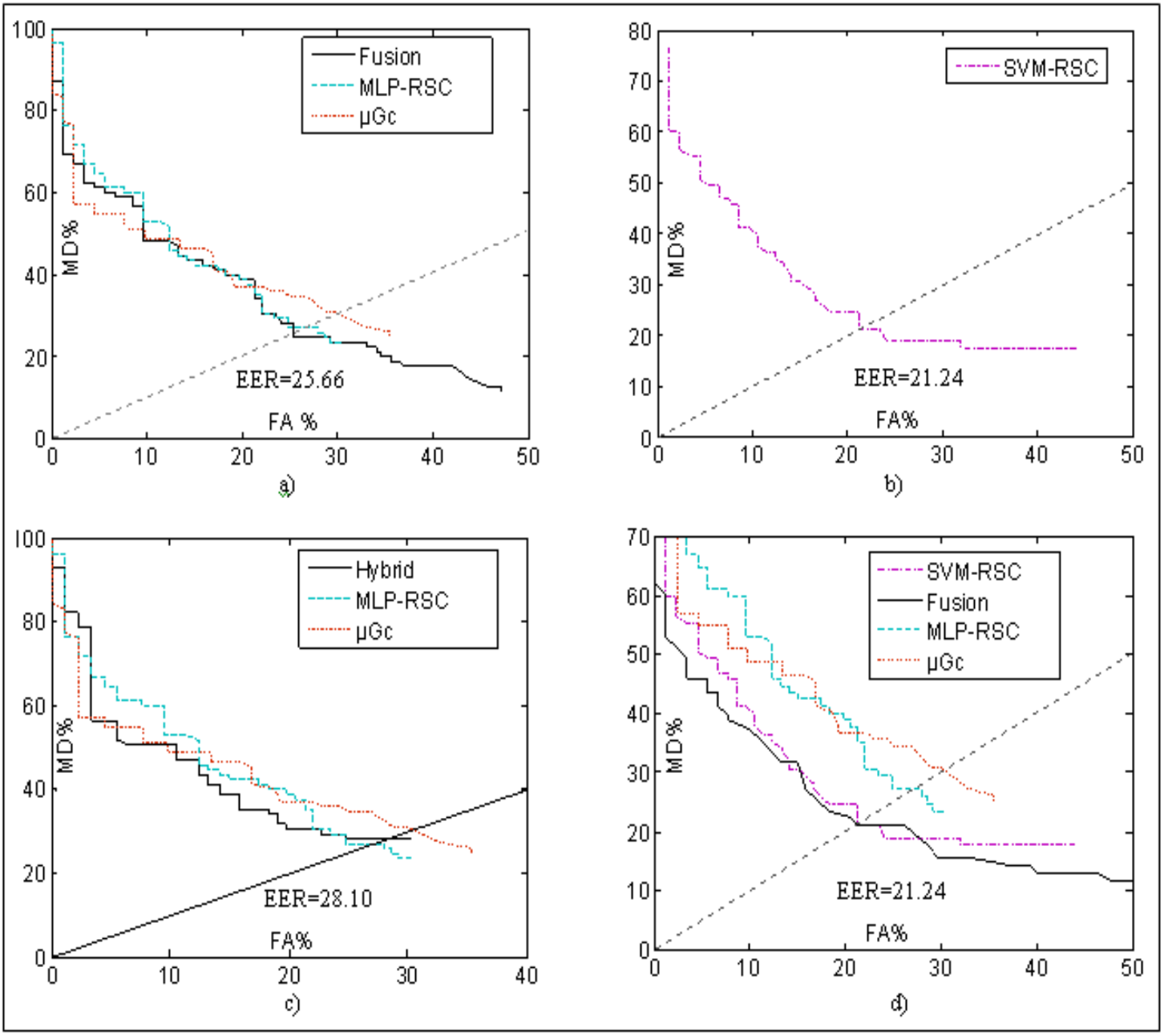

Figure 4.7 : Courbes ROC de segmentation utilisant : a) la fusion parallèle entre la μ_{Gc} et le MLP, b) le classifieur SVM, c) la fusion sérielle (hybride) entre la μ_{Gc} et le MLP, d) la fusion parallèle entre les 3 classifieurs : μ_{Gc}, MLP et SVM.

Les différentes erreurs (EER) de segmentation en locuteurs sont résumées dans le tab. 4.6.

Tableau 4.6 : Erreurs EER données par les différentes méthodes, avec FP : Fusion Parallèle et FS : Fusion Sérielle.

Méthode	EER %
Mesure statistique (μ_{Gc})	30.48
MLP	27.20
SVM	21.24
FP : MLP / μ_{Gc}	25.66
FS : MLP / μ_{Gc}	28.10
FP : SVM / μ_{Gc}	21.24
FP: SVM / MLP	21.60
FP: SVM / MLP / μ_{Gc}	21.24

4.3.4. Exp 5 : Regroupement Séquentiel des Groupes de Locuteurs

A l'issue de la segmentation de l'Exp 4, le système fournit un ensemble de segments homogènes, contenant chacun, un seul locuteur. La prochaine étape consiste à regrouper ces différents segments en groupes homogènes. Chacun d'eux appartient à un locuteur donné. Pour assurer cette tâche de regroupement, nous avons développé un algorithme basé sur le regroupement séquentiel :

- les différents segments homogènes, obtenus lors de la phase de segmentation, sont repérés par leurs instants de début et de fin, ainsi que de leurs numéros dans le document. Les segments de non parole, sont repérés par le numéro "-1".
- application de la mesure de similarité $\mu_{G\beta}$ entre toutes les paires de segments, pour trouver les segments homogènes appartenant au même locuteur, en utilisant un balayage séquentiel de tous les segments (seg).

  si $\mu_{G\beta}[seg(i), seg(j)] \leq$ seuil **alors** seg(i) et seg(j) appartiennent au même locuteur ;

  si $\mu_{G\beta}[seg(i), seg(j)] >$ seuil **alors** seg(i) et seg(j) appartiennent à des locuteurs différents ;
- réorganisation des groupes en mettant ensemble les numéros et les temps de début et de fin des segments appartenant à chaque locuteur. Le nombre final de locuteurs est égal au nombre de groupes trouvés.
- traçage graphique (en fonction du temps) des différents segments homogènes avec identification du locuteur intervenant dans chaque segment.

Le résultat de la première phase de segmentation est donné dans la figure 4.8. Celle-ci représente un ensemble de segments homogènes numérotés de 0 à 96 pour la partie parole du fichier audio étudié (fichier de durée 30 mn). La partie non parole (musique, etc), est indexée par le numéro -1.

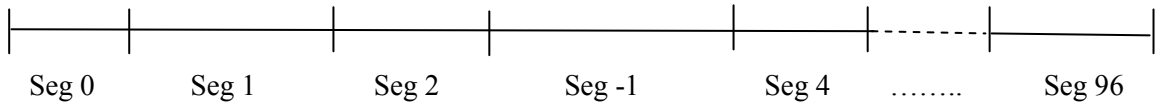

Figure 4.8 : Résultat de la segmentation (ensemble de segments homogènes).

Les différents segments présentés dans cette phase sont illustrés dans le tableau 4.6, qui contient le début et la fin de chaque segment en secondes. Les segments de parole homogènes sont numérotes de 0 à 96 (97 segments).

Tableau 4.7 : Répartition des seg trouvés à l'issue de la segmentation des 3 premières mn de parole.

Répartition des segments																				
Début du segment (s)	0	4	6	7	8	11	14	16	28	33	36	38	47	62	63	108	153	160	163	176
Fin du segment (s)	4	6	7	8	11	14	16	28	33	36	38	47	62	63	108	153	160	163	176	179
Numéro du segment	0	1	2	-1	3	4	5	6	7	-1	8	9	-1	10	11	12	13	14	15	-1

Dans la phase de regroupement des segments homogènes, nous utilisons une mesure de similarité qui permet de détecter les segments homogènes similaires (appartenant au même locuteur) pour les rassembler ensemble. Le résultat du regroupement conduit à 10 groupes trouvés (10 clusters), contre 9 groupes existants réellement dans le flux audio.

Une fois que nous avons regroupé les différents segments par groupes, l'algorithme réaffiche les durées de début et fin de chaque segment portant le nouveau numéro (tab. 4.7).

Tableau 4.8 : Affichage des seg avec les nouveaux numéros et leurs durées pour les 3 mn du début du fichier de parole.

Répartition des segments																				
Début du segment (s)	0	4	6	7	8	11	14	16	28	33	36	38	47	62	63	108	153	160	163	176
Fin du segment (s)	4	6	7	8	11	14	16	28	33	36	38	47	62	63	108	153	160	163	176	179
Numéro du segment	0	0	0	-1	1	0	0	0	0	-1	0	0	-1	0	2	0	0	0	3	-1

La figure 4.9 représente le regroupement final trouvé avec la mesure $\mu_{G\beta}$, où nous pourrons remarquer 10 groupes (10 locuteurs) différents présentés selon leur intervention chronologique dans le signal audio, de durée totale 30 mn. Tandis que, la figure 4.10 représente les 9 locuteurs réels avec leurs temps d'intervention réels représentés sur la même échelle temporelle. La comparaison entre les résultats obtenus et les groupes réels montre un taux de bonne indexation de 75% pour le locuteur principal (le journaliste animateur intervenant dans ce document), ce qui correspond en temps, à une détection correcte de 473 s de parole sur un total de 634 s de parole prononcée réellement par ce journaliste (figures 4.9 et 4.10).

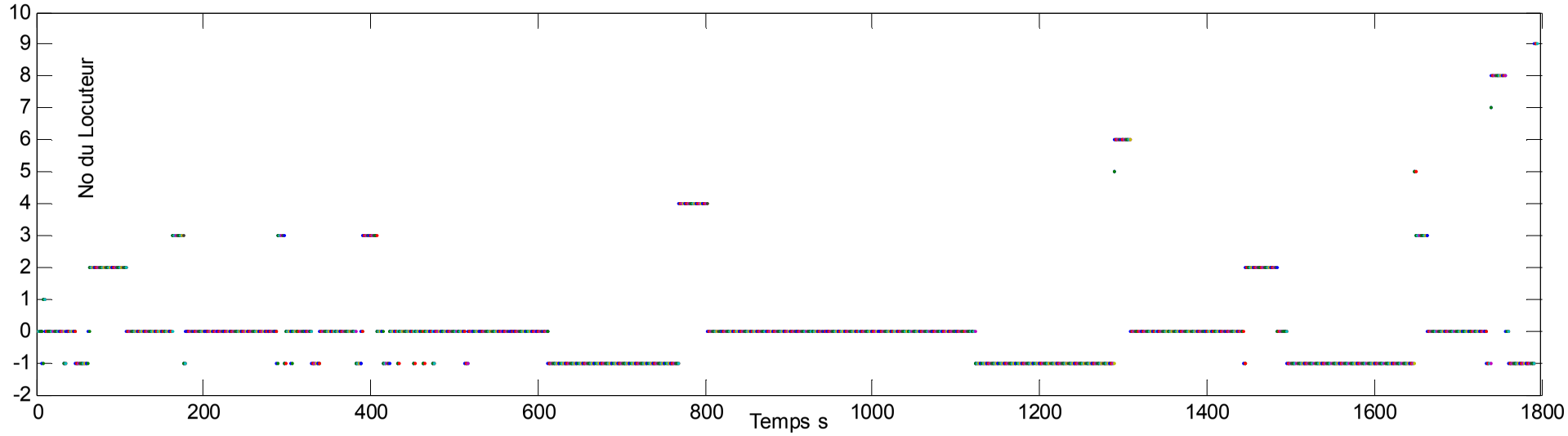

Figure 4.9 : Regroupement par la distance $\mu_{G\beta}$ avec 10 groupes du document de Broadcast-News.

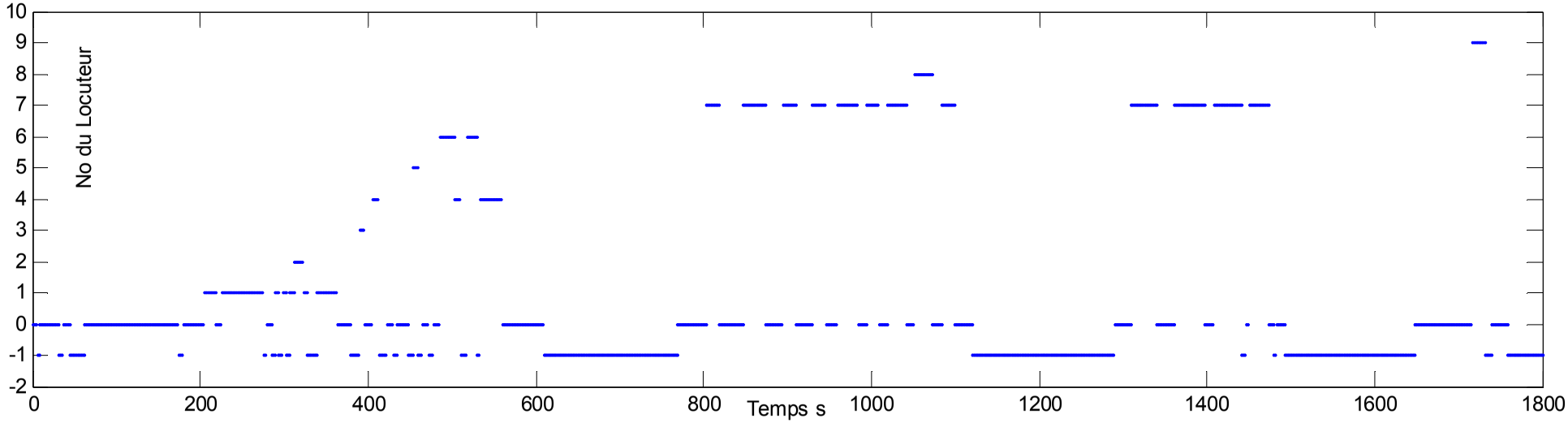

Figure 4.10 : Indexation de référence du document de 30 minutes de Broadcast-News.

4.4. Discussion sur les Expériences d'Indexation Sans Connaissances des Locuteurs

Pour effectuer la tâche de segmentation sans connaissances a priori du locuteur, nous avons proposé une nouvelle approche pour la détection des changements de locuteur, qui consiste à déterminer les moments de parole correspondant au début et fin des interventions des locuteurs dans le document audio, sans se soucier de l'identité de celui qui parle.

Notre méthode de segmentation est basée sur un système de discrimination de locuteurs, appliqué le long du fichier audio à traiter, afin de retrouver les instants dans lesquels un locuteur donné, a cédé la parole à un autre locuteur. Pour accomplir la tâche de segmentation, nous avons effectué quatre séries d'expériences : dans l'expérience N° 2, nous avons testé différentes caractéristiques réduites et comparé ces dernières à une nouvelle caractéristique relative "DRSC" que nous avons développée pour optimiser l'apprentissage des classifieurs MLP et SVM. Cette nouvelle caractéristique s'avère intéressante pour la caractérisation de locuteurs, comparativement aux autres caractéristiques. Dans l'expérience N° 3, nous nous sommes intéressés à la discrimination des locuteurs et nous avons testé comparativement les différents classifieurs proposés (μ_{Gc}, MLP et SVM). De plus, nous avons proposé différentes architectures pour fusionner ces classifieurs, dans le but d'améliorer davantage la qualité de discrimination. Dans l'expérience N° 4, nous avons appliqué ces classifieurs discriminatifs, ainsi que leur fusion, en segmentation des documents audio (ceci a été fait sur un fichier de parole extrait de "HUB-4 Broadcast-News 1996"). Ces expériences vont nous permettre de tirer deux conclusions importantes, dans :

- l'Expérience N° 2 : les résultats confirment que le classifieur utilisant la caractéristique DRSC donne la meilleure performance pour caractériser un locuteur avec un EER de 7.20%, comparativement aux autres caractéristiques. Cette nouvelle caractéristique relative du locuteur, utilisée comme entrée des classifieurs MLP et SVM, a optimisé leur apprentissage, diminué la taille de l'ensemble d'apprentissage et a amélioré la précision de discrimination et de segmentation ;

- les Expériences N° 3 et 4 : le classifieur SVM donne les meilleures performances comparativement aux autres classifieurs simples. Cependant, la fusion parallèle entre les trois classifieurs (μ_{Gc}, MLP et SVM) donne des résultats de discrimination et de segmentation encore meilleurs comparativement aux classifieurs individuels.

Quant à la tâche de regroupement (expérience N° 5), la technique de regroupement séquentiel utilisant la $\mu_{G\beta}$ a permis de regrouper les différents segments trouvés lors de la phase de segmentation (97 segments initiaux) en seulement 10 groupes différents, représentant chacun un locuteur. Toutefois, le nombre réel de ces locuteurs était 9, donc la mesure a détecté un groupe en plus. Quant à l'homogénéité de ces groupes, la mesure utilisée nous a fourni un taux de bon suivi d'environ 75%. Selon ces résultats, nous pouvons conclure que cette méthode est assez intéressante.

4.5. Conclusion

Nous avons exposé dans ce chapitre, nos résultats expérimentaux correspondant aux deux types d'indexation étudiées (avec et sans connaissances a priori des locuteurs). Ces résultats sont issus de quatre séries d'expériences :

- dans la première expérience, nous avons évalué notre nouvel algorithme ISI d'indexation entrelacée sur la base de données HUB-4. Cet algorithme est développé pour assurer la tâche d'indexation avec connaissances des modèles des locuteurs ;
- dans la deuxième expérience, des tests de comparaison effectués sur deux BDs (microphonique et téléphonique) ont permis de montrer l'amélioration des résultats des classifieurs MLP et SVM, en utilisant comme entrée la nouvelle caractéristique DRSC, que nous avons développée ;
- la troisième expérience effectue plusieurs tests de discrimination qui ont été menés sur 4 BDs (2 BDs microphoniques et 2 autres téléphoniques). Ces derniers ont permis de faire une comparaison entre les différents classifieurs implémentés (μ_{Gc}, MLP et SVM). En outre, les résultats montrent que les différentes architectures de fusion proposées ont apporté une nette amélioration des résultats ;
- pour la quatrième expérience, nous avons effectué une segmentation aveugle d'un signal de parole multilocuteurs de durée 30 mn, extrait de HUB-4, en utilisant les différents classifieurs avec une fusion parallèle de ces derniers. Nous avons constaté que la fusion améliore toujours la précision des résultats ;
- dans la cinquième expérience, un regroupement séquentiel, utilisant la mesure $\mu_{G\beta}$ symétrique, rassemble les petits segments homogènes similaires, issus de la segmentation, en groupes (clusters) correspondant, chacun, à un seul locuteur, tout en contenant l'intervention globale de ce locuteur dans le flux audio.

Les résultats de cette partie expérimentale montrent que les systèmes développés sont intéressants en indexation des documents audio.

Conclusions Générales et Perspectives

Ce travail de thèse s'intéresse à l'annotation de documents audio multilocuteurs et plus particulièrement aux émissions radio et télé-diffusées, en utilisant comme clé de recherche (index), les voix des locuteurs participant à la conversation. Il s'agit de l'indexation en classes de locuteurs, dans le but d'obtenir un archivage hiérarchique des interventions audio en fonction des différents locuteurs.

La tâche d'indexation fait appel à deux disciplines différentes, la première s'intéresse à découper le flux audio en segments homogènes les plus longs possibles et ne contenant qu'un locuteur à la fois : c'est la segmentation en locuteurs, tandis que la deuxième tâche consiste à identifier les différents segments ou bien à les regrouper en classes de locuteurs : c'est l'étiquetage ou le regroupement en locuteurs.

Nous rappelons qu'il existe deux types d'indexation, en l'occurrence : l'indexation avec et sans connaissances a priori des locuteurs. La première est adaptée aux applications de suivi du locuteur, où le système d'indexation dispose du modèle des locuteurs à suivre (exemple : poursuite du locuteur par caméra) : ces derniers sont appelés "cibles". Par contre, dans l'indexation sans connaissances des locuteurs, le système ignore complètement l'identité ainsi que le nombre des locuteurs présents dans le fichier audio. Cette dernière tâche est souvent plus difficile à traiter que la première, du fait que le système n'a aucune information sur les locuteurs à indexer. C'est pourquoi, une minutieuse segmentation dite "aveugle", basée sur la détection des points de rupture, est appliquée pour trouver les instants correspondant aux changements de locuteur.

Durant notre travail de recherche, nous avons abordé les deux types d'indexation. Nous avons ainsi développé, pour chaque type, de nouvelles techniques qui ont fait l'objet de plusieurs expérimentations sur des émissions radio et télé-diffusés réelles (tels que HUB-4 Broadcast News).

Dans l'indexation avec connaissances des locuteurs, nous avons proposé et développé une nouvelle technique, que nous avons appelée Indexation Entrelacée de la Parole (abréviation en Anglais : ISI). Cette technique, basée sur les mesures statistiques du second ordre (SOSM), a permis de diminuer les erreurs de confusion et d'améliorer la résolution de segmentation. Les résultats d'évaluation, dans HUB-4 Broadcast News, ont montré l'efficacité de l'approche ISI, associée au classifieur statistique, dans l'élimination des zones de confusion et dans la qualité des résultats obtenus. En plus de ces avantages, l'algorithme ISI a la qualité d'être simple à implémenter, rapide et non coûteux en temps de calcul.

Dans l'indexation sans connaissances des locuteurs, nous avons proposé d'effectuer cette tâche en proposant un système d'indexation basé sur un système de discrimination des locuteurs. Pour cela, nous avons élaboré trois classifieurs, qui sont : le classifieur mono-gaussien symétrique « μ_{Gc} », le perceptron multi couches « MLP » et les Machines à Vecteurs de Support « SVM », ainsi que des techniques de fusion que nous avons développées pour fusionner les scores des différents classifieurs, et ceci dans le but d'améliorer les résultats d'indexation. Ces techniques sont basées sur une fusion des scores en utilisant une somme pondérée des sorties des différents classifieurs. Ces techniques sont connues sous le nom de "fusion parallèle des scores". Toutefois, nous avons proposée une autre architecture pour fusionner le classifieur statistique μ_{Gc} et le classifieur neuronal MLP ; il s'agit de la "fusion sérielle" ou la "fusion hybride".

De plus, nous avons développé une nouvelle caractéristique réduite pour modéliser les locuteurs et qui est introduite comme entrée pour les classifieurs MLP et SVM. Nous l'avons nommée "Caractéristique Relative du Locuteur" (abréviation en Anglais : RSC). Cette nouvelle caractérisation a amélioré les performances d'indexation des deux classifieurs connexionnistes, elle a aussi optimisé et accéléré leur convergence, sans pour autant modifier leurs architectures. Pour l'expérience de discrimination du locuteur, les meilleures performances sont fournies par le classifieur à large bande SVM (avec un EER = 8.63%). De plus, les techniques de fusion qui ont été appliquées entre les différents classifieurs ont encore amélioré la précision de discrimination, prouvant une fois encore l'intérêt de la fusion. Pour la tâche de segmentation, les SVMs paraissent les plus précis, mais la fusion parallèle entre les trois classifieurs apporte encore une amélioration nette des résultats, ce qui nous encourage à utiliser la fusion de plusieurs classifieurs en segmentation. Cette dernière fournit un ensemble de segments homogènes, qui demandent un regroupement en locuteurs, où chacun des groupes contient l'intervention globale d'un seul locuteur dans le flux audio.

La tâche de regroupement est réalisée en utilisant la technique séquentielle basée sur la mesure $\mu_{G\beta}$. Le choix de cette technique vient du fait que le regroupement séquentiel prend en considération le "voisinage" entre les segments. Quant à la mesure symétrique ($\mu_{G\beta}$), nous avons opté pour l'utilisation de cette mesure dans le calcul de la distance de similarité entre les segments homogènes, car celle-ci permet la discrimination de segments de parole de durées différentes. La technique de regroupement, ainsi établie, nous a permis de regrouper les différents segments homogènes trouvés par la segmentation en 10 groupes de locuteurs différents dans la base de données étudiée, alors que le nombre réel des locuteurs était de 9. La mesure a détecté un groupe en plus. Concernant l'homogénéité de ces groupes (appartenance des segments du groupe au même locuteur), cette technique a fourni un taux de bon suivi d'environ 75%.

Les résultats globaux de cette étude montrent que les systèmes développés sont intéressants en indexation des documents audio, comme peuvent le confirmer les chiffres suivants.

Ainsi, nous avons obtenu un taux de réussite de :

- 92.3% pour l'indexation avec connaissances a priori des locuteurs ;
- 93.23% et 96.72% respectivement, pour la discrimination du locuteur, sur des signaux radiodiffusés et sur des signaux téléphoniques ;
- 78.76% pour la segmentation sans connaissances a priori des locuteurs ;
- 75% pour le clustering (regroupement) sans connaissances a priori des locuteurs.

Comme perspectives, nous suggérons une étude complémentaire sur les signaux audio utilisant des technologies de communication modernes, tels que la technologie MP3 (*Motion Picture Experts Group Audio Layer 3*), la technologie VoIP (*Voice over Internet Protocol*) ou la technologie GSM (*Global System for Mobile Communications*). En réalité, ce qui nous intéresse le plus dans ce domaine de recherche, c'est surtout l'indexation dans des conditions difficiles et réelles qu'on rencontre couramment dans la vie courante.

Références Bibliographiques

🕮 RÉFÉRENCES BIBLIOGRAPHIQUES 🕮

[1] J. Wolf, Efficient Acoustic Parameters for Speaker Recognition, JASA, Volume 51, pp. 2045-2056, 1972.

[2] B. Jacob, J. Mariéthoz, G. Gravier, & F. Bimbot, Robustesse de la vérification du locuteur par un mot de passe personnalisé, XXIII^{èmes} Journées d'Etudes sur la Parole (JEP), pp. 357-360, Aussois (France), 2000.

[3] J. Kharroubi & G. Chollet, Utilisation de mots de passe personnalisés pour la vérification du locuteur, XXIII^{èmes} Journées d'Etudes sur la Parole (JEP), pp. 331-334, Aussois (France), 2000.

[4] T. Matsui & S. Furui, Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 125-128, Adélaide (Australie), 1994.

[5] J. Lindberg & H. Melin, Text-prompted versus sound prompted passwords in speaker verification system, European Conference on Speech Communication and Technology (Eurospeech), pp. 22-25, Rhôdes (Grèce), Septembre 1997.

[6] S. Johnson, Who spoke when ? Automatic segmentation and clustering for determining speaker turns, European Conference on Speech Communication and Technology (Eurospeech), Volume 5, pp. 2211-2214, Budapest (Hongrie), Septembre 1999.

[7] P. Delacourt, La segmentation et le regroupement par locuteurs pour l'indexation de documents audio, Thèse de Doctorat, Institut Eurecom, Nice (France), 2000.

[8] M. Przybocki & A. Martin, Two-channel telephone data for speaker detection and speaker tracking, European Conference on Speech Communication and Technology (Eurospeech), Volume 5, pp. 2215-2218, Budapest (Hongrie), Septembre 1999.

[9] A. Martin & M. Przybocki, The NIST 1999, speaker recognition evaluation - an overview, Digital Signal Processing, a review journal - Special issue on NIST 1999 speaker recognition workshop, Volume 10, pp. 1-3, 2000.

[10] C. Fredouille, Approche statistique pour la reconnaissance automatique du locuteur, Thèse de Doctorat, Université d'Avignon, Institut d'Informatique d'Avignon, (France), 2000.

[11] H. Sayoud & S. Ouamour, Speaker Tracking in Multimedia Talk, RIAO'04, pp. 819-825, Avignon (France), 26-28 April 2004.
http://www.riao.org/fr/programme.html. ISBN 2 905450-09-6.

[12] J. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, & C. Wellekens, A speaker tracking system based on speaker turn detection for NIST evaluations. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1177–1180, Istanbul (Turquie), 2000.

[13] A. Rosenberg, I. Magrin-Chagnolleau, S. Parthasarathy, & Q. Huang, Speaker detection in broadcast speech databases, International Conference on Spoken Language Processing (ICSLP), Volume 4, pp. 1339-1342, Sydney (Australia), 1998.

[14] J. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, & C. Wellekens, Différentes stratégies pour le suivi du locuteur, Reconnaissance des Formes et Intelligence Artificielle (RFIA), pp. 123-129, Paris (France), 2000.

[15] K. Sonmez, L. Heck, & M. Weintraub, Speaker tracking and detection with multiple speakers, European Conference on Speech Communication and Technology (Eurospeech), pp. 2219-2222, Budapest (Hongrie), Septembre 1999.

[16] S. Meignier, J. Bonastre, C. Fredouille, & T. Merlin, Evolutive HMM for speaker tracking system, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1177-1180, Istanbul (Turquie), 2000.

[17] I. Magrin-Chagnolleau, A.E. Rosenberg, & S. Parthasarathy, Detection of Target Speakers in Audio Databases, Proceedings of ICASSP'99, Volume 2, pp. 821-824, Phoenix, Arizona (USA), March 1999.

[18] P.C. Woodland, M.J.F. Gales, D. Pye, & S.J. Young, The Development of the 1996 HTK broadcast news transcription system, In : DARPA Speech Recognition Workshop, pp. 97-99, 1997.

[19] J.L. Gauvain, L. Lamel, & G. Adda, Partitioning and transcription of broadcast news data, International Conference on Spoken Language Processing ICSLP'98, Volume 4, pp. 1335-1338, Sydney (Australia), 1998.

[20] H. Sayoud, S. Ouamour, & B. Boudraa, Audio documents Indexing and speaker filtering, TALN, pp. 141-148, Nancy (France), 24-27 juin 2002.

[21] H. Sayoud, S. Ouamour & M. Guerti, Recherche des Points de Rupture dans les Discours Multi-Locuteurs - Application en Indexation par Locuteurs -, SETIT 2005, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, paper number 189 (Tunisia), March 27-31 2005.

[22] S. Ouamour, H. Sayoud, & M. Boudraa, Application of Statistical Measures for the detection of speaker transitions, AMAM'03, Nice (France), 10-13 Février 2003. http://acm.emath.fr/amam/.

[23] S. Ouamour, M. Guerti, & H. Sayoud, Statistical Discrimination and Identification of Some Acoustic Sounds, 3rd IEEE-GCC Conference, Manama (Bahrain), 19-22 March 2006. Article sur CD. www.ieeegcc.org/program/DigitalSignalProcessing-II.doc.

[24] H. Sayoud, S. Ouamour & M. Guerti, Tentative de Discrimination des Sons Parlés et Non-Parlés, SETIT, 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, paper number 190 (Tunisia), March 27-31, 2005.

[25] S. Meignier, Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection, Thèse de doctorat, Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse, Avignon (France), 2002.

[26] M. Nishida & Y. Ariki, Real time speaker indexing based on subspace method: applications to TV news articles and debate, International Conference on Spoken Language Processing, Volume 4, pp. 1347-1350, 1998.

[27] M. Nishida & Y. Ariki, Speaker indexing for news articles debates and drama in broadcasted TV programs, IEEE International Conference on Multimedia Computing and Systems, pp. 466-471, 1999.

[28] S. Tsekeridou & I. Pitas, Audio-visual content analysis for content-based video indexing, IEEE International Conference on Multimedia Computing and Systems, pp. 667-672, 1999.

[29] S. Ouamour, H. Sayoud, & M. Boudraa, Application of Statistical Measures in Speaker Identification, AMAM'03, Nice (France), 10-13 Février 2003. http://acm.emath.fr/amam/.

[30] S. Ouamour & H. Sayoud, Automatic Speaker Recognition Using Statistical Measures, IASSE, pp. 100-103, Nice (France), 1-3 July 2004. ISBN 1-880843-52-X.

[31] D. A. Reynolds & P. Torres-Carrasquillo, The MIT Lincoln Laboratories RT-04F diarization systems : Applications to broadcast audio and telephone conversations, Rich Transcription Workshop (RTW' 04), Palisades, NY, Fall 2004.

[32] S. Ouamour, M. Guerti & H. Sayoud, PENS: A Confidence Parameter Estimating the Number of Speakers, Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, TRWEL08, pp. 151-164, Athens (Greece), 25- 27 August 2008.

[33] J. Pelecanos & S. Sridharan, Feature warping for robust speaker verification, a Speaker Odyssey, The Speaker Recognition Workshop, pp. 213–218, Crete (Greece), 2001.

[34] H. Hermansky & N. Morgan, RASTA processing of speech, IEEE Transactions on Speech and Audio Processing, tome 2, pp. 578–589, 1994.

[35] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Transactions on Acoustics, Speech, and Signal Processing, tome 29(2), pp. 254–272, 1981.

[36] D. Moraru, S. Meignier, L. Besacier, & J.-F. Bonastre, Combining experts for automatic speaker segmentation, Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP), Hong Kong (China), 2003.

[37] D. A. Reynolds, R. B. Dunm, & J. J.Laughlin, The lincoln speaker recognition system NIST EVAL2000, Proceedings of International Conference on Spoken Language Processing (ICSLP), pp. 470-473, Beijing (China), 2000.

[38] A. Adami, S. S. Kajarekar, & H. Hermansky, A new speaker change detection method for two-speaker segmentation, Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002), tome IV, pp. 3908–3911, 2002.

[39] R. Sinha, S. E. Tranter, J. J. F. Gales, & P. C. Woodland, The cambridge university march 2005 speaker diarization system, European Conference on Speech Communication and Technology (Interspeech), pp. 2437–2440, Lisbon (Portugal), 2005.

[40] X. Zhu, C. Barras, L. Lamel, & J.-L Gauvain, Speaker diarization : from broadcast news to lectures, NIST 2006 Spring Rich Transcrition Evaluation Workshop, Washington DC (USA), 2006.

[41] T. Kristjansson, S. Deligne, & P. Olsen, Voicing features for robust speech detection, Proc. International Conference on Speech and Language Processing, pp. 369-372, Lisbon (Portugal), 2005.

[42] W.-H. Tsai, S.-S. Cheng, Y.-H. Chao, & H.-M. Wang, Clustering speech utterances by speaker using eigenvoice-motivated vector space models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 725-728, Philadelphia (USA), 2005.

[43] D. Sturim, D. Reynolds, E. Singer, & J.P. Campbell, Speaker indexing in large audio databases using anchor models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City (USA), 2001.

[44] M. Collet, D. Charlet, & F. Bimbot, A correlation metric for speaker tracking using anchor models, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 713-716, Philadelphia (USA), 2005.

[45] J. M. Pardo, X. Anguera, & C. Wooters, Speaker diarization for multiple distant microphone meetings : Mixing acoustic features and inter-channel time differences, Proc. International Conference on Speech and Language Processing, pp. 2194-2197, 2006.

[46] L. Lu & H.-J. Zhang, Speaker change detection & tracking in real-time news broadcasting analysis, ACM International Conference on Multimedia, pp. 602–610, 2002.

[47] S. S. Chen, M. J. F. Gales, R. A. Gopinath, D. Kanvesky, & P. Olsen, Automatic transcription of broadcast news, Speech Communication 37, pp. 69–87, 2002.

[48] L. Perez-Freire & C. Garcia-Mateo, A multimedia approach for audio segmentation in TV broadcast news, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 369–372, Montreal (Canada), 2004.

[49] G. Schwarz, A sequential student test, The Annals of Statistics Volume 42 (3), pp. 1003–1009, 1971.

[50] J. F. Lopez & D. P. W. Ellis, Using acoustic condition clustering to improve acoustic change detection on broadcast news, Proc. International Conference on Speech and Language Processing, Volume 4, pp. 568-571, Beijing (China), 2000.

[51] A. Vandecatseye, J.-P. Martens, & al. The cost278 pan-european broadcast news database, LREC'04, pp. 873–876, Lisbon (Portugal), 2004.

[52] S. Shaobing Chen & P. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the bayesian information criterion, Proceedings DARPA Broadcast News Transcription and Understanding Workshop, Virginia (USA), 1998.

[53] S. sian Cheng, & H. min Wang, A sequential metric-based audio segmentation method via the bayesian information criterion, Eurospeech'03, pp. 945-948, Geneva (Switzerland), 2003.

[54] M. Cettolo & M. Vescovi, Efficient audio segmentation algorithms based on the BIC, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 537-540, 2003.

[55] A. Tritschler & R. Gopinath, Improved speaker segmentation and segments clustering using the bayesian information criterion, Eurospeech'99, Budapest (Hungary), pp. 679-682, 5-9 Septembre, 1999.

[56] P. Delacourt, D. Kryze, & C. J. Wellekens, Detection of speaker changes in an audio document, Eurospeech'99, pp.1195-1198, Budapest (Hungary), 1999.

[57] S. sian Cheng & H. min Wang, METRIC-SEQDAC : A hybrid approach for audio segmentation, Proc. International Conference on Speech and Language Processing, pp. 1617-1620, Jeju (S. Korea), 2004.

[58] A. M. Xavier, Robust Speaker Diarization for meetings, PhD Thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya Barcelona (Espagne), October 2006.

[59] A. S. Willsky & H. L. Jones, A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems, IEEE Transactions on Automatic Control AC-21(1), pp. 108–112, 1976.

[60] U. Appel & A. Brandt, Adaptive sequential segmentation of piecewise stationary time series. Inf. Sci, Volume 29 (1), pp. 27–56, 1982.

[61] R. Gangadharaiah, B. Narayanaswamy, & N. Balakrishnan, A novel method for twospeaker segmentation, Proc. International Conference on Speech and Language Processing, pp. 2337-2340, Jeju (S. Korea), 2004.

[62] H. Gish, M.-H. Siu, & R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 873–876, Toronto (Canada), 1991.

[63] T. Kemp, M. Schmidt, M. Westphal, & A. Waibel, Strategies for automatic segmentation of audio data, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1423–1426, Istanbul (Turkey), 2000.

[64] M. A. Siegler, U. Jain, B. Raj, & R. M. Stern, Automatic segmentation, classification and clustering of broadcast news audio, DARPA Speech Recognition Workshop, Chantilly (USA), pp. 97–99, 1997.

[65] J. Hung, H. Wang, & L. Lee, Automatic metric based speech segmentation for broadcast news via principal component analysis, Proc. International Conference on Speech and Language Processing, Volume 4, pp. 121-124, Beijing (China), 2000.

[66] J. P. Campbell, Speaker recognition : a tutorial. Proceedings of the IEEE, pp. 1437–1462, 1997.

[67] H.G. Kim, D. Ertelt, & T. Sikora, Hybrid speaker-based segmentation system using model-level clustering, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 745- 748, Philadelphia (USA), 2005.

[68] L. Lu & H.-J. Zhang, Real-time unsupervised speaker change detection. ICPR'02, Volume 2, pp. 358-361, Quebec City (Canada), 2002.

[69] M. Nishida & T. Kawahara, Unsupervised speaker indexing using speaker model selection based on bayesian information criterion, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, pp. 172–175, Hong Kong (China), 2003.

[70] L. Lu, S. Z. Li, & H.-J. Zhang, Content-based audio segmentation using support vector machines, ACM Multimedia Conference, pp. 203–211, Ottawa, Ontario (Canada), 2001.

[71] S. Nakagawa & H. Suzuki, A new speech recognition method based on VQ-distortion and hmm, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 676–679, Minneapolis (USA), 1993.

[72] K. Mori & S. Nakagawa, Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 1, pp. 413–416, Salt Lake City (USA), 2001.

[73] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, & J. Martinez, Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 521-524, Toulouse (France), 2006.

[74] R. Duda & P. Hart, Pattern classification and Scene analysis, John Wiley & Sons, 1973.

[75] H. Jin, F. Kubala, & R. Schwartz, Automatic speaker clustering. DARPA Speech Recognition workshop, pp. 108–111, Chantilly (USA), 1997.

[76] B. Zhou & J. H. Hansen, Unsupervised audio stream segmentation and clustering via the bayesian information criterion, Proc. International Conference on Speech and Language Processing, Volume 3, pp. 714–717, Beijing (China), 2000.

[77] M.-H. Siu, G. Yu, & H. Gish, An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 189–192, San Francisco (USA), 1992.

[78] Q. Jin, K. Laskowski, T. Schultz, & A. Waibel, Speaker segmentation and clustering in meetings. NIST 2004 Spring Rich Transcription Evaluation Workshop, pp. 597-600, Montreal (Canada), 2004.

[79] S. S. Chen & P. Gopalakrishnan, Clustering via the bayesian information criterion with applications in speech recognition, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume 2, pp. 645–648, Seattle (USA), 1998.

[80] S. Johnson & P. Woodland, Speaker clustering using direct maximization of the MLLR adapted likelihood, Proc. International Conference on Speech and Language Processing, Volume 5, pp. 1775–1779, 1998.

[81] S. Tranter & D. Reynolds, Speaker diarization for broadcast news, ODYSSEY'04, Toledo (Spain), 2004.

[82] T. Hain, S. Johnson, A. Turek, P. Woodland, & S. J. Young, Segment generation and clustering in the HTK broadcast news transcription system, DARPA Broadcast News Transcription and Understanding Workshop, pp. 133–137, 1998.

[83] S. Tranter, Two-way cluster voting to improve speaker diarization performance, Proc.IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 753-756, Montreal (Canada), 2005.

[84] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, & I. Magrin-Chagnolleau, The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Montreal (Canada), 2004.

[85] C. Fredouille, D. Moraru, S. Meignier, L. Besacier, & J.-F. Bonastre, The NIST 2004 spring rich transcription evaluation : Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation, NIST Rich Transcription Evaluation Workshop, Montreal (Canada), Spring 2004.

[86] A. Tritschler & R. Gopinath, Improved speaker segmentation and segments clustering using the bayesian information criterion, Eurospeech'99, pp. 679–682, Budapest (Hungary), 5-9 September 1999.

[87] L. Canseco-Rodriguez, L. Lamel, & J.-L. Gauvain, Speaker Diarization from Speech Transcripts, Proc. International Conference on Speech and Language Processing, pp. 1272-1275, Jeju Island, (S. Korea), 2004.

[88] L. Canseco, L.Lamel, & J.-L. Gauvain, A comparative study using manual and automatic transcriptions for diarization, IEEE Automatic Speech Recognition and Understanding Workshop, San Juan, Puerto Rico, 2005.

[89] G. Lathoud, I. McCowan, & J. Odobez, Unsupervised location-based segmentation of multi-party speech, ICASSP-NIST Meeting Recognition Workshop, Montreal (Canada), May 2004.

[90] J. M. Pardo, X. Anguera, & C. Wooters, Speaker diarization for multi-microphone meetings using only between-channel differences, MLMI, pp. 257-264, Bethesda, MD (USA), 2006.

[91] G. Lathoud, I. McCowan, & J. Odobez, Unsupervised location-based segmentation of multi-party speech, ICASSP-NIST Meeting Recognition Workshop, Montreal (Canada), 2004.

[92] D. Ellis & J. C. Liu, Speaker turn detection based on between-channels differences, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2004.

[93] J. Ajmera, Robust Audio Segmentation, Thèse de Doctorat, Ecole Polytechnique Fédérale de Lausanne, 2004.

[94] F. Bimbot, I. Magrin-Chagnolleau, & L. Mathan, Second-order statistical measures for text-independent speaker identification, Speech Communication, Volume 17, pp. 177-192, Août 1995.

[95] E. Davalo & P. Naim, Des Réseaux de neurones, 2ème édition EYROLLES, Paris (France), 1993.

[96] H. Sayoud, Reconnaissance Automatique du Locuteur -Approche Connexionniste-, Thèse de Doctorat, Université des Sciences et de la Technologie Houari Boumedienne USTHB, Institut d'Electronique, Alger (Algérie), 2003.

[97] Y. Bennani, Approches Connexionnistes Pour La Reconnaissance Automatique du Locuteur : Modélisation et Identification, Thèse de Doctorat, Université de Paris Sud (France), 1992.

[98] J. Master, Pratical Neural Network Recipes, Acadimic press 1992.

[99] A. Freeman & D. Skapura, Neural Networks, Algorithms, Applications and programming techniques, édition Addison-Wesley, 1992.

[100] B. Kröse & P. V. Der-Smagt, An introduction to neural networks, 8ème édition, 1996. http://neuron.tuke.sk/math.chtf.stuba.sk/pub/vlado/NN_books_texts/Krose_Smagt_neuro-intro.pdf

[101] J. Hérault, & C. Jutten, Réseaux Neuronaux Et Traitement Du Signal, Editions Hermès, Paris (France), 1994.

[102] Encyclopédie Wikipédia : http://fr.wikipedia.org/wiki/Machine_%C3%A0_vecteurs_de_support.

[103] W. Vincent, Speaker Verification using Support Vector Machines, PhD thesis, Department of Computer Science, University of Sheffield, Sheffield (United Kingdom), June 2003.

[104] F. Bimbot and al, A Further Investigation on AR-Vector Models for Text-Independent Speaker Identification, ICASSP Volume 1, pp. 401-404, Atlanta (United States), May 1996.

[105] H. Sayoud & S.Ouamour, Optimal Spectral Resolution in Speaker Identification -Application in noisy environment-, 1st Biosecure Residential Workshop, pp. 69-69, Paris (France), 1-26 August 2005. www.tsi.enst.fr/biosecure/posters/sayoud_reso_P1.pdf

[106] H. Sayoud & S.Ouamour, Looking for the Best Spectral Resolution in Automatic Speaker Recognition, 3rd IEEE-GCC 2006 Conference, Manama (Bahrain), 19-22 March 2006.
www.ieeegcc.org/program/DigitalSignalProcessing-II.doc.

[107] B.V. Dasarathy, Decision Fusion, IEEE Computer Society Press, Los Alamitos, CA, 1994.

[108] P. Verlinde, Contribution à la vérification multi-modale de l'identité en utilisant la fusion de décisions, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, MA, Bruxelles (Belgique), September 17th 1999.

[109] A. K. Jain, A. Ross, and S. Prabhakar, An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology Journal, Volume 14 (1), 4-20 January, 2004.

[110] Y. Stylianou, Y. Pantazis, F. Calderero, P. Larroy, F. Severin, S. Schimke, R. Bonal, F. Matta, and A. Valsamakis. GMM- Based Multimodal Biometric Verification. Final Project Report 1, Enterface'05, Mons (Belgium), July 18 - August 12, 2005.

[111] M.Yamaguchi, M. Yamashita, & S. Matsunaga, Spectral cross-correlation features for audio indexing of broadcast news and meetings, Proc. International Conference on Speech and Language Processing, pp. 613-616, 2005.

[112] Y. Bennani & P. Gallinari, Neural Networks for discrimination and modelization of speakers, Speech Communication, Volume 17 (1-2), pp. 159-175, 1995.

[113] J. Kittler, Multiple classifier systems in decision-level fusion of multimodal biometric experts, 1st BioSecure residential workshop, Paris (France), 1- 26 August 2005.

[114] S. Ouamour, M. Guerti, & H. Sayoud, ISI A New Method for Automatic Speaker Tracking and Detection, 3rd IEEE-GCC 2006 Conference, Manama (Bahrain), 19-22 March 2006. Article sur CD.
www.ieeegcc.org/program/DigitalSignalProcessing-II.doc.

[115] S. Ouamour, M. Guerti, & H. Sayoud, Speaker based segmentation on broadcast news - on the use of ISI technique-, Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, pp. 197-200, Athens (Greece), 28- 30 August 2006.

[116] S. Ouamour, M. Guerti, & H. Sayoud, A New Relativistic Vision in Speaker Discrimination, Canadian Acoustics Journal, Volume 36, Number 4, pp. 24-34, December 2008.

[117] J. Bonastre & 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'Acoustique, pp. 357-360, Marseille (France), 14-18 April 1997.

[118] S. Ouamour, M. Guerti, & H. Sayoud , "ISI" Une Nouvelle Technique pour la Segmentation Automatique par Locuteurs, 4th International Conference : Sciences of Electronic, Technologies of Information and Telecommunications SETIT, (Tunisia) March 25-29, 2007. Article sur CD.

[119] S. Ouamour, M. Guerti & H. Sayoud, Suivi Automatique du Locuteur Utilisant une Segmentation Equidistante, SETIT 2005, 3rd International Conference : Sciences of Electronic, Technologies of Information and Telecommunications SETIT, paper number 195, (Tunisia), March 27-31 2005. Article sur CD.

[120] P. Nguyen, SWAMP : An isometric frontend for speaker clustering, NIST 2003 Rich Transcription Workshop, Boston (USA), 2003.

[121] S. Ouamour, M. Guerti & H. Sayoud, Authentification Discriminative du Locuteur Basée sur une Fusion Statistique–Connexionniste, 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications SETIT, (Tunisia), March 25-29, 2007.

[122] A. Freeman & D. Skapura, Neural Networks, Algorithms, Applications and programming techniques, édition Addison-Wesley, 1992.