



ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection

Présentée et soutenue publiquement le 14 novembre 2002 pour obtenir le grade de
Docteur en Sciences de l'Université d'Avignon et des Pays de Vaucluse

SPÉCIALITÉ : INFORMATIQUE

par

Sylvain Meignier

Composition du jury :

Mme	Catherine Berrut	PR, CLIPS	Présidente du jury
Mme	Régine André-Obrecht	PR, IRIT, Toulouse	Rapporteurs
M	Eric Moulines	PR, ENST, Paris	
M	Frédéric Bimbot	CR, IRISA, Rennes	Examineurs
M	Jean-Claude Junqua	HDR, PSTL, Santa Barbara	
M	Jean-François Bonastre	MC-HDR, LIA, Avignon	Directeurs de thèse
M	Henri Méloni	PR, LIA, Avignon	



Laboratoire Informatique d'Avignon

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury pour leur participation à la soutenance de cette thèse : Madame Catherine Berrut pour avoir présidé le jury, Madame Régine André-Obrecht et Monsieur Eric Moulines, rapporteurs de ce travail, pour avoir consacré du temps à la lecture de ce document, Messieurs Jean-Claude Junqua pour l'intérêt et l'attention qu'il a portés à ces travaux et Monsieur Frédéric Bimbot pour sa participation active au jury.

Je remercie également Monsieur Henri Méloni, co-directeur de cette thèse, pour le suivi et le soutien apportés tout au long de son déroulement.

Ce travail n'aurait vu le jour sans la large contribution de Monsieur Jean-François Bonastre, co-directeur de cette thèse. Je tiens à le remercier vivement autant pour la qualité de son encadrement, que pour sa disponibilité et que pour ses qualités humaines. Merci aussi pour l'ensemble des moments de convivialité passés en dehors du travail.

Je voudrais remercier tous les membres du LIA et de l'IUP qui ont énormément contribué aux conditions de travail très chaleureuses. J'adresse de vives remerciements aux RALeurs, RAPEurs et aux membres d'ELISA qui grâce à leur collaboration ont permis lors de réunions, de pauses cafés, et autres de faire avancer ce travail.

Si j'ai apprécié la collaboration avec Jean-François, c'est aussi parce qu'il a su former une équipe : les RALeurs. Ce travail de thèse ne serait pas ce qu'il est sans la participation active des membres du groupe. Merci à Jean-françois, Corinne et Teva pour leur soutien et leur aide infaillibles.

Je remercie aussi Loïc pour m'avoir supporté pendant ces nombreuses années et avoir contribué à la relecture de nombreux articles. Malgré ses blagues insupportables (sauf certains fous-rires...), je remercie Domi pour sa joie de vivre et sa bonne humeur qui sont venues agrémenter mes deux dernières années de thèse.

Je remercie enfin toute ma famille pour son soutien au cours du parcours aboutissant à cette thèse.

Resumé

L’archivage sous forme numérique des documents offre beaucoup d’avantages. Il permet, en particulier, le transfert rapide et peu coûteux de l’information et le stockage de grandes quantités de documents dans des espaces physiques restreints. Cette accumulation de documents de natures variées crée un besoin d’accès efficace et rapide à l’information.

Cette thèse s’inscrit dans le cadre de l’indexation automatique de documents sonores en locuteurs. L’indexation en locuteurs permet de détecter les locuteurs présents dans un document ou dans une collection de documents et d’annoter leur contenu suivant ces locuteurs. Elle met à la disposition des moteurs de recherche un critère supplémentaire, leur permettant d’accéder aux annotations issues de l’interprétation des documents.

Deux thèmes ont retenu notre attention : la segmentation en locuteurs d’un document sonore et l’appariement en locuteurs d’une collection de documents sonores. La segmentation en locuteurs de documents détermine le nombre de locuteurs d’un document et les interventions de chaque locuteur. L’appariement en locuteurs d’une collection de documents détecte le nombre de locuteurs présents dans la collection et classe les interventions par locuteur. Ce travail est réalisé à partir de documents préalablement segmentés (tâche de segmentation vue précédemment).

Dans le cadre de cette étude, nous proposons une approche au problème de la segmentation en locuteurs par une méthode itérative originale. Celle-ci repose sur une modélisation de la conversation par un modèle de Markov caché qui évolue à chaque détection d’un nouveau locuteur. La méthode proposée a obtenu des performances intéressantes lors des évaluations internationales NIST en segmentation en locuteurs, tout en réduisant d’environ deux tiers les temps de calcul par rapport aux méthodes concurrentes.

L’appariement en locuteurs est une tâche récente et encore peu abordée dans la littérature. Dans ce travail, nous avons choisi une méthode fondée sur la classification hiérarchique ascendante. Nous proposons deux nouvelles mesures permettant de tirer profit des informations apportées par la segmentation des documents, ainsi qu’une nouvelle méthode de détection du nombre de locuteurs. Cette dernière repose sur l’élagage de l’arbre de classification sans nécessiter une coupe préalable de l’arbre. L’appariement de locuteurs a été évalué sur un corpus issu des évaluations NIST.

Mots clés : recherche d’information, document sonore, indexation en locuteurs, segmentation en locuteurs, appariement en locuteurs, reconnaissance du locuteur, méthode statistique, modèle de Markov, classification hiérarchique, maximum de vraisemblance.

Sommaire

I	Position du problème	3
1	Cadre général	5
1.1	L’annotation de documents et les index correspondants	5
1.2	Indexation automatique	5
1.3	L’indexation des documents multimédias	6
1.4	L’indexation en locuteurs de documents sonores	6
1.5	Autres caractéristiques	9
2	Indexation en locuteurs de documents sonores	11
2.1	Les locuteurs	12
2.2	Problèmes rencontrés lors de l’indexation en locuteurs	12
2.2.1	Conditions d’enregistrement	12
2.2.2	Nature de la parole	13
2.2.3	Problème de l’identité des locuteurs	13
2.3	Les documents	13
2.4	Hypothèses pour l’indexation en locuteurs	14
II	Segmentation en locuteurs d’un document sonore	17
3	Principe général et état de l’art	19
3.1	Principe général	20
3.2	Etat de l’art : outils pour la segmentation	21
3.2.1	Paramétrisation	21
3.2.2	Modèles et mesures	22
3.3	Etat de l’art : segmentation en locuteurs	26
3.3.1	Segmentation	26
3.3.2	Détection des ruptures	27
3.3.3	Classification : regroupement des segments	30
3.3.4	Resegmentation	34
3.3.5	Performance de la classification hiérarchique ascendante	35
3.4	Commentaires	35
3.4.1	Structure du modèle classique	35
3.4.2	Détection de ruptures	36
3.4.3	Classification	36

4	Méthode proposée : détection de ruptures et classification	39
4.1	HMM évolutif	40
4.1.1	Principe général	40
4.1.2	Méthode proposée <i>v.s.</i> méthode classique	41
4.2	Structure du modèle	42
4.3	Détails de l'algorithme itératif	42
4.3.1	Initialisation	42
4.3.2	Etape 1 : nouveau modèle de locuteur	43
4.3.3	Etape 2 : Adaptation des modèles de locuteurs et segmentation provisoire	43
4.3.4	Etape 3 : critère d'arrêt	44
4.4	Originalités de la méthode proposée	45
4.5	Exemple	45
4.5.1	Itération 1 : construction du premier locuteur	45
4.5.2	Itération 2 : ajout du deuxième locuteur	47
4.5.3	Itération 3 : ajout du troisième locuteur	49
4.6	Discussion	51
4.6.1	Hierarchisation des modèles	51
4.6.2	Les données servant à l'initialisation d'un locuteur	51
4.6.3	Maximum de vraisemblance	51
4.6.4	Méthode ascendante <i>v.s.</i> méthode descendante	51
5	Méthode proposée : implémentation et évaluation	53
5.1	Méthode d'évaluation	54
5.1.1	Choix liés à l'évaluation	54
5.2	Outils issus de la reconnaissance du locuteur	56
5.2.1	SPRO	56
5.2.2	Plate-forme AMIRAL	56
5.3	Les traitements effectués avant la segmentation	56
5.3.1	Paramétrisation	57
5.3.2	Segmentation initiale	58
5.4	Ajout de locuteur et initialisation d'un locuteur	58
5.4.1	Sélection de données	59
5.4.2	Transition entre les états	60
5.5	Modèle de conversation	61
5.5.1	Modèle de locuteur	62
5.5.2	Probabilités d'émission	63
5.5.3	Critère d'arrêt de la phase d'apprentissage/décodage	64
5.6	Détection du nombre de locuteurs	64
5.6.1	Cas de la segmentation en deux locuteurs	64
5.6.2	Cas de la segmentation en n locuteurs	65
6	Conclusion	69
III	Appariement en locuteurs d'une collection	71
7	Positionnement du problème	73
7.1	Segmentation et appariement en locuteurs	73

SOMMAIRE	VII
7.2 Spécificités de l'appariement en locuteurs	74
7.3 De l'appariement à l'application	74
7.4 Cadre de l'étude : limites	75
8 Etat de l'art	77
8.1 Principe général de la classification hiérarchique ascendante	78
8.2 Aspect combinatoire	79
8.3 Mesure de dissimilarité	79
8.4 Estimation des dissimilarités	80
8.5 Sélection de la partition finale	80
8.5.1 Sélection à une hauteur donnée	81
8.5.2 Elagage du dendrogramme	82
9 Méthode proposée	83
9.1 Principe général	84
9.2 Mesures de dissimilarité proposées	85
9.3 Matrice de dissimilarité	85
9.4 Elagage du dendrogramme	86
10 Evaluation de la méthode	87
10.1 Corpus d'évaluation	88
10.2 Paramétrisation et modèles	88
10.3 Expériences préliminaires	88
10.4 Expériences d'appariement	90
10.4.1 Evaluation de la partition	90
10.4.2 Résultats	90
10.4.3 Commentaire	91
11 Conclusions et perspectives	93
IV Campagnes d'évaluations NIST en reconnaissance du locuteur	95
12 Evaluations NIST	97
12.1 Objectifs des évaluations NIST	97
12.2 Tâches proposées	97
12.2.1 Vérification du locuteur : <i>1-speaker</i>	98
12.2.2 Vérification de locuteur : <i>2-speaker</i>	98
12.2.3 Suivi de locuteur : <i>speaker tracking</i>	99
12.2.4 Segmentation en locuteurs : <i>2-segmentation</i> et <i>n-segmentation</i>	99
12.3 Systèmes proposés aux évaluations NIST	99
13 Evaluation du résultat de la segmentation	101
13.1 Création des segmentations de référence	102
13.2 Critères de qualité d'une segmentation	102
13.3 Evaluation de la détection du nombre de locuteurs	102
13.4 Evaluation des ruptures	103
13.5 Evaluation des segments	104
13.5.1 NIST 2000 et 2001	104

13.5.2 NIST 2002	105
14 Segmentation en locuteurs : <i>n-segmentation</i>	107
14.1 Corpus d'évaluation	108
14.1.1 NIST 2000 et 2001 : <i>CallHome</i>	108
14.1.2 NIST 2002 : <i>Meeting et Broadcast News</i>	108
14.2 Comparaison des corpus	109
14.3 Résultats	109
14.3.1 NIST 2000 et 2001	109
14.3.2 NIST 2002	111
15 Segmentation en locuteurs : <i>2-segmentation</i>	115
15.1 Corpus d'évaluation	116
15.1.1 NIST 2000 et 2001	116
15.1.2 NIST 2002	116
15.2 Résultats	117
15.2.1 NIST 2000 et 2001	117
15.2.2 Résultats NIST 2002	117
15.2.3 Comparaison des corpus	117
16 Conclusion	119
 V Conclusions et perspectives	 121
17 Conclusions et perspectives	123
17.1 Conclusions	123
17.2 Perspectives	125
A Vérification du locuteur, NIST 2000 à 2002	129
B Plate-forme de développement ELISA	131
B.1 Consortium ELISA	131
B.2 Plate-forme pour la tâche de vérification du locuteur	131
B.3 Plateforme pour la tâche de segmentation en locuteurs	132
C Système LIA 2000 et 2001 de segmentation en locuteurs	135
C.1 Système développé par le LIA pour NIST 2001	135
C.2 Système développé par le LIA pour NIST 2001	135
D Descriptif des systèmes proposés lors des évaluations NIST 2002	137
E Bibliographie personnelle	147

Notations

C	une collection de documents.
K	le nombre de documents de la collection C .
L	le nombre de locuteurs de la collection C .
I	le nombre d'interventions dans la collection C .
X_k	le $k^{ième}$ document d'une collection C .
$o_{k,i}$	une observation, une trame ou un vecteur acoustique.
d	la dimension d'un vecteur acoustique.
O_k	une séquence d'observations du document X_k .
T	le nombre d'éléments dans la séquence d'observations O_k .
S_k	la segmentation en locuteur du document X_k .
$s_{k,i}$	le $i^{ième}$ segment de la segmentation S_k .
n_k	le nombre de segments dans S_k .
$\mathcal{X}_{k,i}$	l'étiquette du locuteur i dans le document X_k .
\mathcal{X}_k	l'ensemble des étiquettes des locuteurs dans le document X_k .
L_{X_k}	le nombre de locuteurs dans le document X_k .
$x_{k,i}$	l'intervention (les segments) du locuteur i dans le document X_k .
$\overline{x_{k,i}}$	les segments ne correspondant pas au locuteur i dans le document X_k .
$X_{k,i}$	le modèle des segments du locuteur i dans le document X_k .
$\overline{X_{k,i}}$	les modèles des segments ne correspondant pas au locuteur i dans le document X_k .
\mathcal{C}_i	l'étiquette du locuteur i dans une collection C de documents.
c_i	les segments du locuteur i dans la collection C de documents.
$\overline{c_i}$	les segments ne correspondant pas au locuteur i dans la collection C .
C_i	le modèle des segments c_i du locuteur i dans une collection C de documents.
$\overline{C_i}$	les modèles des segments ne correspondant pas au locuteur i dans une collection C .
$P_{L,I}$	la partition en L classes de locuteurs de la collection de documents contenant I interventions.
$l(x \cdot)$	vraisemblance d'un modèle \cdot pour les observations x .
W	modèle du monde.
w	corpus d'apprentissage du modèle du monde.
$lr(x Z)$	rapport de vraisemblance des modèles Z et W pour les observations x .
d_{CLR}	dissimilarité CLR.
λ	un modèle de Markov caché (HMM).
$E = \{e_i\}$	l'ensemble d'états de λ .
$b_i()$	fonction de densité de probabilité de l'état i du HMM.
$a_{i,j}$	probabilité de transition de l'état i vers l'état j .
$g()$	une gaussienne.

Introduction

Avec l'essor des nouvelles technologies de l'information, la quantité de documents numériques s'accroît considérablement. En plus des documents textuels, le son, la vidéo et l'image prennent une part croissante dans les documents mis à la disposition du public. Beaucoup d'organismes numérisent et archivent ces émissions de radios et de télévisions, permettant ainsi aux auditeurs de consulter les émissions ultérieurement, à partir d'Internet ou à partir de bases de données spécialisées. Manipuler des documents sous forme numérique facilite le transfert d'information, l'archivage des collections de grand volume et la consultation des documents. Cependant, lire, écouter ou regarder l'ensemble de ces documents multimédia à la recherche d'une information précise devient une gageure.

L'accès direct à l'information recherchée, sans parcourir 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.

L'annotation suivant des caractéristiques prédéfinies et la création d'index résultant de cette annotation définissent la tâche d'indexation de documents. Dans le cadre de cette étude, nous nous intéressons à l'indexation des documents sonores. Deux points ont retenu notre attention : la segmentation en locuteurs d'un document sonore et l'appariement en locuteurs d'une collection de documents sonores.

Dans la première partie, une vue générale de l'indexation des documents multimédia est donnée. À partir de cette description, les spécificités des systèmes d'indexation en locuteurs de documents sonores sont précisées. Enfin, les hypothèses déterminant notre travail sont définies et discutées.

La seconde partie s'intéresse à la première tâche de l'indexation en locuteurs : la segmentation. La segmentation en locuteurs d'un document sonore consiste à déterminer le nombre de locuteurs intervenant dans le document et à préciser les instants de début et de fin des paroles prononcées par chaque locuteur. La méthode généralement utilisée repose sur une classification hiérarchique des segments préalablement détectés. Une nouvelle approche est proposée permettant de résoudre les problèmes inhérents à la méthode classique. Elle repose sur la modélisation de la conversation entre les locuteurs par un modèle de Markov caché évolutif qui est complété et modifié à chaque détection d'un nouveau locuteur (à chaque itération du processus). Cette méthode est évaluée dans le cadre des évaluations internationales organisées par l'institut américain NIST (*National Institute of Standard and Technology*).

La troisième partie expose un problème peu abordé dans la littérature : l'appariement en locuteurs. Cette tâche consiste à détecter le nombre de locuteurs intervenant dans la collection et à regrouper les interventions issues des différents documents par locuteur. Elle est appliquée sur une collection de documents sonores préalablement segmentés en locuteurs. Les méthodes décrites dans la littérature s'inspirent généralement des méthodes de classification, employées notamment pour la

segmentation en locuteurs (étape dite de *clustering*). L'approche proposée complète les méthodes classiques de la littérature sur deux points : les mesures de dissimilarité entre les interventions et l'élagage de l'arbre de classification (détection du nombre de locuteurs de la collection). L'idée directrice des travaux est d'utiliser au maximum les informations apportées par la segmentation des fichiers. Les apports proposés sont évalués sur un corpus issu des campagnes d'évaluation NIST et composé d'enregistrements de conversations téléphoniques.

La quatrième partie présente les résultats obtenus à l'aide de notre méthode lors des évaluations internationales organisées par l'institut NIST, durant les années 2000 à 2002. Ces évaluations utilisent des corpus couvrant la plupart des types de documents sonores (conversations téléphoniques, réunions, journaux télévisés).

La dernière partie conclut ce travail de thèse et propose des perspectives sur les deux tâches abordées : la segmentation en locuteurs d'un document sonore et l'appariement en locuteurs d'une collection de documents sonores.

Première partie

Position du problème

Cette partie est une introduction au domaine de l'indexation automatique de documents sonores. Elle présente le cadre général et les principes de l'indexation de documents sonores afin de montrer ses spécificités comparées à l'indexation de documents. Puis, notre sujet d'étude est abordé : l'indexation automatique en locuteurs des documents sonores. Les divers problèmes de l'indexation en locuteurs sont exposés comme la variabilité des conditions d'enregistrement et la variabilité de la voix des locuteurs. Enfin, les hypothèses de cette étude sont définies.

Chapitre 1

Cadre général

1.1 L’annotation de documents et les index correspondants

L’annotation des documents consiste à rechercher les segments de document ou les documents entiers qui contiennent des caractéristiques définies *a priori* [Smeaton 2001]. Les annotations sont mises en forme afin d’obtenir un index mettant en correspondance caractéristiques et segments. L’index facilite l’accès, la recherche et le rangement des documents suivant leurs caractéristiques. Il permet d’accéder directement aux segments correspondant à une caractéristique sans qu’il soit nécessaire de parcourir l’intégralité de la collection ou l’intégralité du document. Seul l’index, de taille réduite par rapport au document, est parcouru. Par exemple, dans un système de recherche documentaire, un utilisateur demande à consulter l’ensemble des documents (textes, sons, vidéos, images) comportant des animaux. Le système de recherche lui propose alors la liste des documents dans lesquels des animaux apparaissent ; ainsi qu’un accès pour consulter (lire, écouter, visionner) seulement les parties de documents dont le contenu se rapporte au sujet de recherche.

1.2 Indexation automatique

Nous définissons l’indexation automatique de documents comme le processus visant à la création d’index [Berrut 1997, Delacourt 2000a, Seck 2001]. Les index sont généralement créés par avance, pour répondre à des besoins spécifiques. Le processus d’indexation (figure 1.1) se divise en quatre phases distinctes :

1. Un ensemble de caractéristiques à rechercher est choisi. Le contenu de cet ensemble dépend de l’application envisagée et plus particulièrement des besoins des utilisateurs.
2. Un processus automatique annote et interprète les documents suivant les caractéristiques prédéfinies.
3. Les annotations sont classées, organisées en fonction des caractéristiques afin d’obtenir un index par document. Un index de la collection est construit à partir des index des documents.
4. La création d’index n’étant pas une fin en soit, il faut définir des outils de recherche et d’accès à l’information adaptés aux besoins des utilisateurs. La performance de ces outils se mesure autant en terme de flexibilité des requêtes que de qualité des réponses apportées. La convivialité et la rapidité d’accès aux documents sont aussi des critères importants à prendre en compte. Cette dernière phase est la frontière entre l’indexation et l’interrogation : les outils de recherche utilisent le résultat de l’indexation pour proposer des documents ou les parties de documents correspondant aux requêtes des utilisateurs.

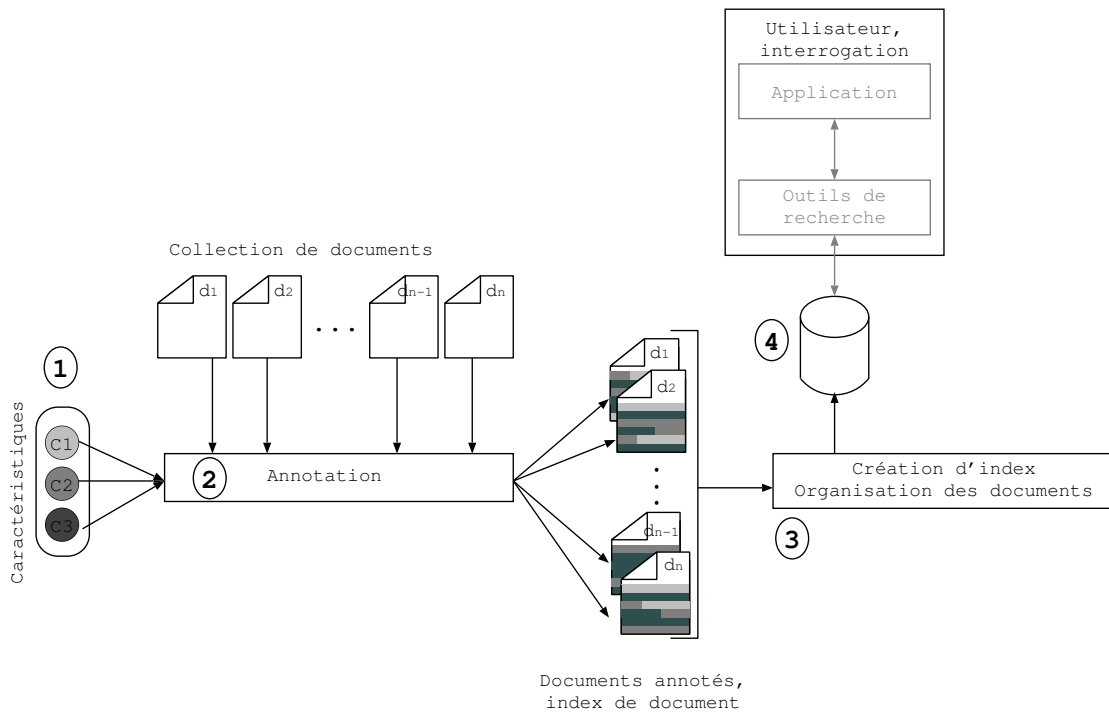


FIG. 1.1 – Principe de l'indexation automatique.

1.3 L'indexation des documents multimédias

L'indexation automatique des documents numériques a débuté par l'indexation des documents textuels [van Rijsbergen 1979, Salton 1983]. Au début des années 1990, avec l'avènement de l'Internet, les premiers moteurs de recherche sont apparus sur "Gopher", l'ancêtre du "World Wide Web" actuel. Dix ans plus tard, la plupart des moteurs de recherche n'indexent encore que les documents textuels et cela de façon rudimentaire. Les requêtes sont exprimées sous la forme d'une liste de mots-clés, alors que l'utilisateur est plus à même de formuler ses besoins sous la forme de questions en langage naturel. Bien que la nature des documents accessibles se soit largement diversifiée, peu de moteurs de recherche, pour ne pas dire aucun, exploitent le contenu des images, les documents sonores, les vidéos ou, plus largement, les documents multimédias. Ces outils restent encore dans le domaine de la recherche. Les articles de [Viswanathan 2000, Bett 2000, Weber 2000] donnent des exemples d'implémentations de systèmes d'indexation et de recherche. Dans les exceptions, la société *Compaq* propose, en libre accès au public, un moteur de recherche (*SpeechBot*) qui indexe les flux audio de radios [Thong 2000].

1.4 L'indexation en locuteurs de documents sonores

Bien que la finalité de l'indexation soit la recherche d'information, l'indexation est aussi abordée dans d'autres domaines de recherche : reconnaissance automatique de la parole (RAP), reconnaissance automatique du locuteur (RAL), traitement du signal, reconnaissance de formes...

La reconnaissance du locuteur s'intéresse, en particulier, à l'indexation en locuteurs des documents sonores. Le principe général présenté au paragraphe 1.2 demeure (figure 1.2). Les quatre

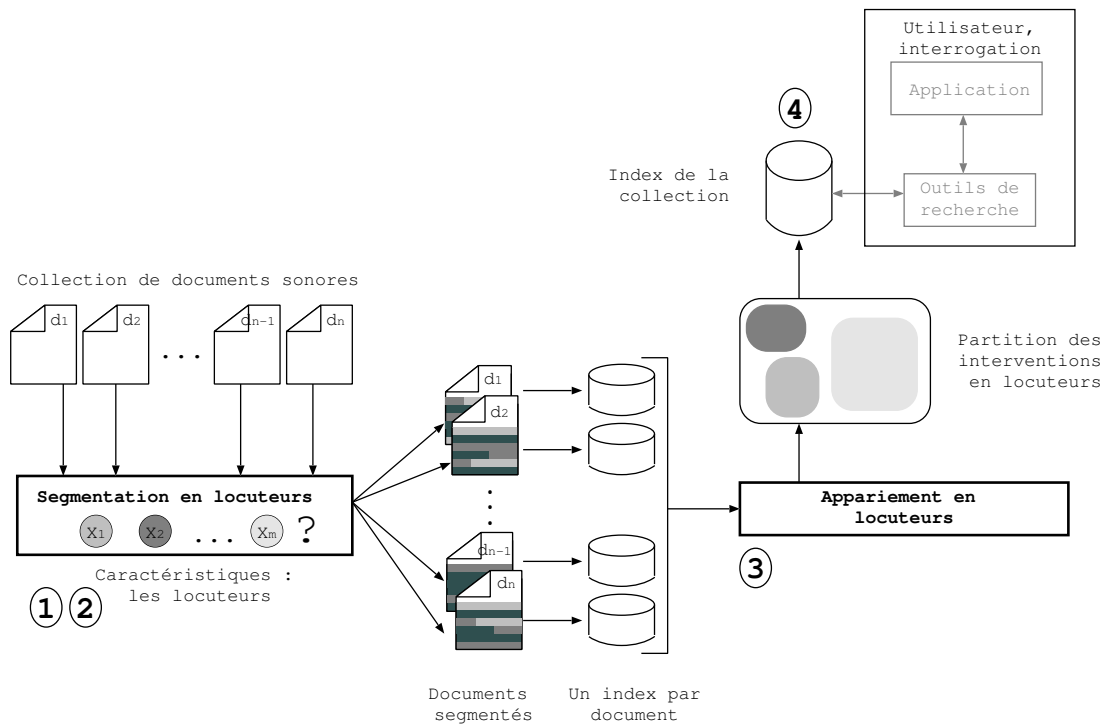


FIG. 1.2 – Principe de l'indexation automatique en locuteurs.

différentes tâches définies pour l'indexation automatique sont présentes de nouveau. Les documents sont annotés suivant une caractéristique : les locuteurs (ou plus exactement, la voix des locuteurs). Cette caractéristique a la particularité d'être un ensemble non prédéfini, *i.e.* les locuteurs sont découverts lors de l'indexation.

L'indexation en locuteurs est composée de quatre tâches :

1. Dans un premier temps les documents sont individuellement annotés suivant les locuteurs. La tâche consiste à :
 - découper le flux audio en segments homogènes,
 - associer chaque segment à un locuteur,
 - trouver le nombre de locuteurs intervenant dans le document.

Pour chaque document est générée une segmentation contenant les locuteurs et précisant l'intervention de chacun d'eux (l'ensemble des segments d'un locuteur). Cette tâche est appelée **segmentation en locuteurs** (figure 1.3).

2. Un index du document est généré à partir de la segmentation en locuteur (et éventuellement à partir de la segmentation suivant d'autres caractéristiques). L'index associe à chaque caractéristique les segments correspondants. Cette tâche consiste uniquement à mettre en forme la segmentation en vue d'une utilisation ultérieure. Plusieurs initiatives sont en cours pour définir un standard d'annotation permettant de décrire le contenu des documents multimédias. A titre d'exemple, nous invitons le lecteur à se reporter aux spécifications du projet *SMIL* [W3C 1998] du *W3C* qui permet la définition du comportement d'une présentation multimédia (incluant des sons, des images, des vidéos...). Ce format, en plus des modules de définition du comportement, contient un module de description des caractéristiques, dont le

locuteur.

3. Une troisième phase recherche les locuteurs intervenant dans la collection. Cette phase est réalisée à partir de l'ensemble des segmentations ou à partir des index des documents. L'objectif est de regrouper en classes les locuteurs intervenant dans plusieurs documents. Une partition en classes est obtenue, dans laquelle chaque classe contient les interventions d'un locuteur. Cette tâche est appelée **appariement en locuteurs** (figure 1.3).
4. Enfin, la partition de la collection est organisée sous la forme d'un index. L'index de la collection met en relation les locuteurs avec leurs documents respectifs. Au travers de l'index de la collection et des index des documents, un système de recherche retrouve les documents et les segments dans chaque document d'un locuteur correspondant aux critères de recherche.

Ce travail de thèse porte sur deux de ces quatre tâches : **la segmentation en locuteurs** et **l'appariement en locuteurs**. D'autres caractéristiques sonores sont susceptibles d'être recherchées dans des documents sonores, cependant l'étude porte uniquement sur les locuteurs.

D'autres caractéristiques pour les documents sonores sont décrites, pour information, au paragraphe suivant (1.5). La partie de mise en forme de l'index en vue de son exploitation n'est pas abordée dans ce travail.

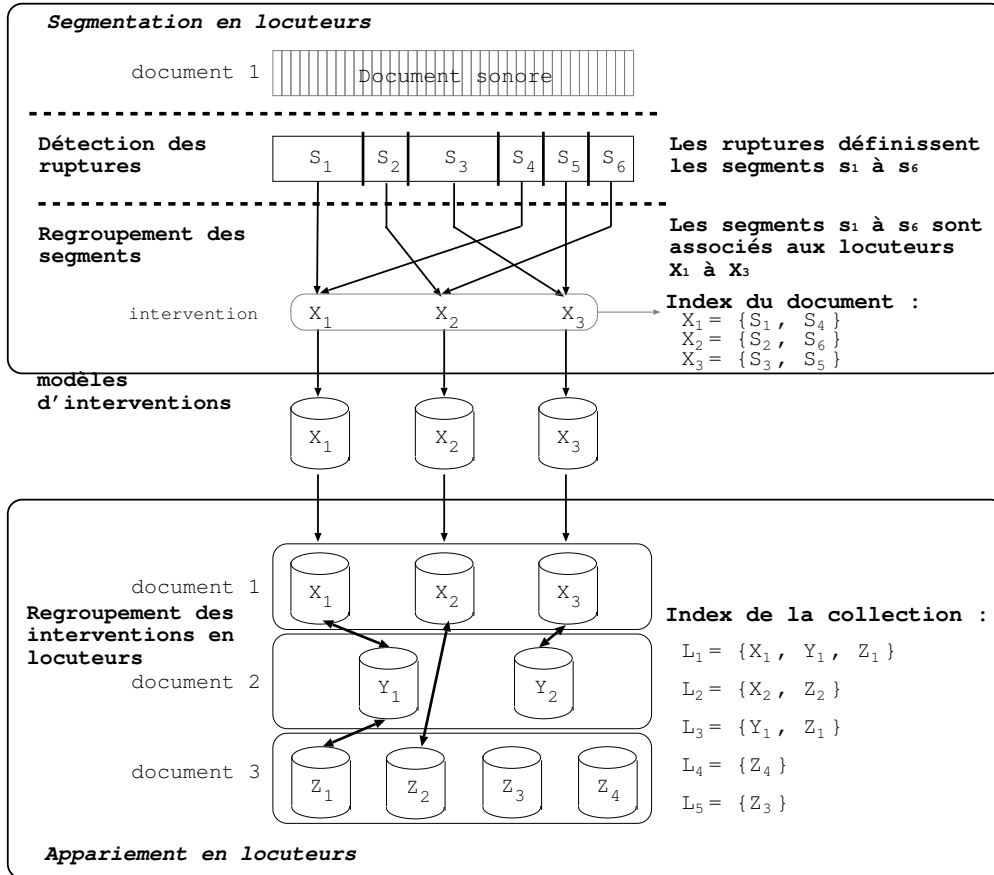


FIG. 1.3 – Indexation en locuteurs : segmentation et appariement.

1.5 Autres caractéristiques

Nous rappelons (*c.f.* 1.2) que les caractéristiques à rechercher sont définies *a priori* suivant les besoins des utilisateurs. Bien que les données bibliographiques (titre, auteur, ...), les données décrivant le document (par exemple la durée des documents, le type d'enregistrement, ...) soient susceptibles de faire l'objet d'une indexation, seules les caractéristiques décrivant le contenu des documents ont retenu notre attention. F. Chen dans [Chen 1997] donne un aperçu de la diversité des caractéristiques en indexation de documents sonores. Nous ne pouvons, bien sûr, être exhaustif dans l'énonciation des caractéristiques possibles.

Les caractéristiques citées ci-dessous sont issues pour la plupart des caractéristiques proposées dans [Chen 1997, NIST 2002c, Jones 2002].

Les caractéristiques acoustiques portent en particulier sur :

- La nature du signal : parole, musique, ou bruit...
- Les sources d'enregistrement et leurs natures : nombre de microphones, type de microphone...
- Le type de canal de transmission : bande étroite (téléphone), bande large (studio),...
- Le locuteur : l'identité, le genre, l'état émotionnel, l'état pathologique.

Les caractéristiques linguistiques principales sont :

- Le découpage du discours en phrases, paragraphes et en unités thématiques.
- La prosodie du discours : la phrase est déclarative, interrogative ou exclamative.
- L'annotation d'entités telles que les noms de personnes, de villes, d'organisations...

Chapitre 2

Indexation en locuteurs de documents sonores

*Dans ce chapitre, les problèmes posés par l'indexation en locuteurs sont discutés pour les deux tâches de notre étude : la **segmentation en locuteurs** et l'**appariement en locuteurs**. Les problèmes sont dus à plusieurs facteurs : la nature de la parole, les conditions d'enregistrement, l'identité des locuteurs... Les types de documents visés par l'indexation en locuteurs sont précisés. Finalement les hypothèses de travail pour les deux tâches sont définies.*

2.1 Les locuteurs

L'indexation en locuteurs recherche les interventions des locuteurs dans des documents sonores (un ou plusieurs). Ce thème de recherche s'inscrit dans le domaine plus large de la caractérisation du locuteur et plus précisément dans le domaine de la reconnaissance automatique du locuteur (RAL). La RAL s'intéresse aux applications qui vérifient, identifient et localisent les locuteurs à partir d'échantillons de leur voix.

Les systèmes de RAL utilisent la variabilité inter-individuelle pour caractériser les locuteurs¹. La variabilité inter-individuelle est multi-forme et provient principalement des différences physiques de l'appareil de production de la parole entre les individus (taille du conduit, amplitude et rapidité de déplacement des articulateurs...). Les informations phonologiques, langagières, prosodiques et psycho-phonétiques sont aussi des paramètres permettant de caractériser les locuteurs [Bonastre 1994]. Les méthodes utilisées en RAL [Bonastre 2000], depuis une dizaine d'années, sont essentiellement des méthodes statistiques et probabilistes. Elles obtiennent, jusqu'à présent, les meilleures performances.

2.2 Problèmes rencontrés lors de l'indexation en locuteurs

Les problèmes rencontrés lors de l'indexation en locuteurs proviennent de la grande variabilité du contenu des documents à traiter. Deux catégories de problèmes en indexation sont à souligner : les problèmes dus aux conditions d'enregistrement d'une part, et les problèmes liés à la nature de la parole d'autre part. Ces deux catégories ne sont pas spécifiques à l'indexation en locuteurs, elles sont aussi présentes en RAL et en reconnaissance automatique de la parole (RAP).

2.2.1 Conditions d'enregistrement

Les conditions d'enregistrement influencent le contenu des documents sonores. Les informations liées au canal de transmission et au matériel d'enregistrement sont des éléments venant perturber (déformer et dégrader) le signal. Suivant le canal de transmission utilisé, des pertes d'information plus ou moins importantes sont mesurées. Les enregistrements téléphoniques, par exemple, sont davantage dégradés que les enregistrements en studio. Avec l'arrivée massive de la téléphonie mobile de nouveaux phénomènes comme les pertes d'information et la compression des données apportent de nouvelles difficultés.

Le matériel d'enregistrement (en particulier les microphones) a des spécificités techniques propres qui influencent le signal. Il a été montré que ces conditions d'enregistrement ont une incidence sur les performances des systèmes de RAL [Van Vuuren 1996].

Les documents peuvent contenir un enregistrement provenant de plusieurs sources. La nature de ces sources, dans cette étude, est généralement de la parole, mais aussi de la musique ou des bruits liés à l'environnement. Toutes les sources sont susceptibles d'émettre en simultané. Par exemple, au début des journaux télévisés ou radiophoniques, le présentateur annonce les titres tandis que le générique de l'émission est encore diffusé.

Note : Suivant les documents à traiter les problèmes soulevés au paragraphe précédent deviennent une aide à la segmentation : par exemple dans le cadre des conversations téléphoniques, généralement, chaque locuteur utilise un combiné différent. Les spécificités des microphones et des canaux de transmission facilitent la segmentation en locuteurs.

¹Dans certaines situations, le canal d'enregistrement est une aide à la détection des locuteurs, *c.f.* 2.2.1.

2.2.2 Nature de la parole

Bien que le signal de parole varie en fonction de l'individu (ce qui permet de différencier les individus entre eux, *c.f.* § 2.1), les variations intra-locuteur rendent problématique la tâche de reconnaissance :

- Les expériences montrent qu'il existe des variations dans le signal pour un même locuteur prononçant plusieurs fois le même énoncé, *i.e.* le locuteur est incapable de reproduire à l'identique un même énoncé.
- Des variations intra-locuteur à court terme sont aussi présentes. Elles sont principalement liées à l'état pathologique et émotionnel (fatigue, rhume, stress...) [Homayounpour 1995, Scherer 1998, Karlsson 1998, Banziger 2000].
- Des variations à long terme interviennent également, elles proviennent en particulier du vieillissement de l'individu.

Bien entendu, d'autres problèmes peuvent intervenir, par exemple les locuteurs qui changent volontairement leur voix ou qui imitent un autre locuteur.

La segmentation en locuteurs n'est pas, ou peu, influencée par les variations intra-locuteur à long terme. Les documents traités sont généralement trop courts pour que ces variations aient une influence sur l'indexation. Par contre, lors de l'appariement en locuteurs, la variation intra-locuteur à court terme et à long terme sont des phénomènes à prendre en compte. Les documents ont pu être enregistrés à plusieurs jours (mois ou année) d'intervalle.

La parole spontanée, en opposition à la parole préparée ou lue, est un facteur qui rend plus difficile la détection des locuteurs. Ce type de parole favorise les recouvrements de voix entre locuteurs (prise de la parole d'un locuteur alors que son interlocuteur n'a pas fini de s'exprimer) et montre une plus grande variabilité.

2.2.3 Problème de l'identité des locuteurs

L'indexation en locuteurs recherche les entités de type locuteur sans déterminer leur identité. Un locuteur indexé est représenté soit par une étiquette (un mot défini arbitrairement), soit par des exemples de sa voix (ou par un modèle de sa voix). La mise en correspondance entre un échantillon de voix et une identité de locuteur est une tâche relevant de la vérification ou d'identification automatique du locuteur (VAL ou IAL). Cette recherche de l'identité peut être faite *a posteriori* à partir de données complémentaires.

2.3 Les documents

Les documents susceptibles d'être traités sont de natures variées. Parmi ces documents, nous nous intéressons plus particulièrement aux conversations téléphoniques, aux enregistrements de journaux télévisés ou radiophoniques, aux films et aux enregistrements de réunions. Ces trois types de documents couvrent la majorité des problèmes évoqués au paragraphe 2.2. Les conditions d'enregistrement (canal, matériel, environnement, diversité des sources...), les facteurs liés à la nature de la parole (parole préparée ou spontanée) sont très variés dans ces types de corpus.

Les corpus utilisés lors des expériences présentées dans cette thèse sont issus des campagnes d'évaluation NIST en reconnaissance du locuteur pour les tâches de segmentation en locuteurs [NIST 2002d]. Les évaluations NIST en segmentation permettent de comparer les résultats obtenus par les différents participants. Tous les participants utilisent les mêmes corpus, les expériences sont

effectuées suivant un protocole standard et les résultats sont évalués à partir des mêmes méthodes. Les comparaisons sont d'autant plus faciles. Les corpus proposés sont des conversations préparées ou spontanées en bande large ou étroite entre deux ou plusieurs locuteurs.

2.4 Hypothèses pour l'indexation en locuteurs

Dans les paragraphes précédents, l'indexation en locuteurs de documents sonores a été positionnée dans le cadre général de l'indexation, puis les problèmes spécifiques liés aux documents sonores et aux locuteurs ont été relevés. A partir de ces éléments, nous définissons les hypothèses de travail suivantes.

- Les tâches d'indexation en locuteurs sont réalisées en aveugle : **aucune connaissance en dehors du document n'est disponible**. De cette condition découlent les hypothèses suivantes :

- **Aucun échantillon de voix des locuteurs n'est disponible :**

Nous supposons que les locuteurs ne sont pas connus du système d'indexation (par avance). Aucun enregistrement de la voix des locuteurs n'est disponible pour apprendre les caractéristiques du locuteur. Les locuteurs devront être détectés à partir des enregistrements à indexer.

- **La langue des locuteurs est inconnue :**

Les documents, comme les collections, sont susceptibles de contenir des locuteurs parlant différentes langues. Par exemple, les bases de données *CallHome* et *CallFriend* contiennent des documents en anglais, allemand, mandarin, japonais, arabe et espagnol [LDC 1998]. En outre dans un même document, un locuteur peut s'exprimer dans une langue différente des autres intervenants. De même les journaux contiennent des interventions de personnes s'exprimant dans une autre langue.

- **Le nombre de locuteurs n'est pas connu :**

Le système d'indexation doit déterminer le nombre de locuteurs pour chaque document (respectivement pour la collection). Cette proposition est bien entendue discutable pour les conversations téléphoniques, qui généralement ne contiennent que deux locuteurs.

- **Aucune indexation sur d'autres caractéristiques n'est disponible :**

Ni une indexation Homme/Femme, ni une indexation Parole/Bruit/Silence, ni la transcription du discours ne sont disponibles *a priori* lors de l'indexation en locuteurs. Dans le cadre de notre travail, nous avons choisi explicitement de ne pas rechercher ces caractéristiques.

- L'objectif est d'indexer des documents enregistrés dans des conditions réelles². Nous supposons que :

- **L'environnement d'enregistrement n'est pas contrôlé :**

Les enregistrements sont réalisés en environnement réel. Après écoute de certains enregistrements, nous avons noté en fond sonore différents éléments comme des pleurs d'enfants, des portes qui claquent, de la musique, des dés roulant sur une table...

- **Les informations sur le type d'enregistrement sont fournies :**

A part les journaux télévisés qui intègrent des conversations de qualité studio avec des conversations de qualité téléphonique, les sources d'enregistrements et les canaux de transmission sont de même nature. Le nombre de sources d'enregistrement n'est pas une donnée disponible, mais ce nombre est au mieux égal au nombre de sources émettrices. La fréquence d'échantillonnage des enregistrements est en accord avec le canal de transmission utilisé :

²Les caractéristiques des conditions d'enregistrement ont été décrites au paragraphe 2.2.1.

8kHz pour le téléphone, 16kHz pour les conditions "studio". Pour les journaux télévisés ou radiophoniques, les présentateurs peuvent interviewer par téléphone des reporters qui sont en extérieur. L'enregistrement contient alors une alternance de segments bande large (studio) et bande étroite (téléphone).

- Les méthodes employées pour l'indexation en locuteurs imposent généralement les deux hypothèses suivantes [Delacourt 2000a, Moraru 2001] :

- **Les documents contiennent de la parole :**

Nous supposons que les documents contiennent majoritairement de la parole. Cette contrainte semble raisonnable pour les tâches visées prioritairement (indexation de conversations téléphoniques, de journaux télévisés et de réunions).

- **Les locuteurs ne se coupent pas la parole :**

Cette hypothèse est irréaliste pour des conversations spontanées, mais généralement les méthodes utilisées jusqu'ici ne sont pas capables de détecter si plusieurs locuteurs parlent simultanément.

Deuxième partie

Segmentation en locuteurs d'un document sonore

La recherche des locuteurs intervenant au sein d'une conversation constitue une tâche essentielle pour l'indexation par le contenu de documents sonores. Dans ce chapitre, le principe général de la segmentation automatique en locuteurs est étudié avant de décrire les outils issus de la reconnaissance du locuteur et de la parole. Puis, différentes méthodes de segmentation présentées jusqu'ici sont examinées. Enfin, la méthode de segmentation proposée dans ce travail est détaillée.

Chapitre 3

Principe général et état de l'art

Le principe général de la segmentation en locuteurs a été défini par H. Gish dans ses travaux avec la société BBN. H. Gish a proposé l'architecture type d'un système [Siu 1991, Siu 1992]. Dans ce chapitre, les techniques fondamentales utilisées en segmentation sont tout d'abord introduites : la paramétrisation acoustique et les outils statistiques comme les modèles multi-gaussiens et les modèles de Markov cachés. La majorité des systèmes de segmentation repose sur une structure proche de celle définie par H. Gish. Les modules principaux de l'architecture sont présentés : la détection des ruptures permettant d'obtenir une segmentation initiale, la classification hiérarchique des segments initiaux pour obtenir les modèles de locuteur et l'étape optionnelle de resegmentation.

3.1 Principe général

Dans le domaine du traitement automatique de la parole, la segmentation en locuteurs consiste à proposer pour un flux audio un descriptif définissant le nombre de locuteurs et l'intervention de chaque locuteur.

La plupart des travaux en segmentation d'un document sonore utilisent des méthodes statistiques et probabilistes pour modéliser les différentes caractéristiques à rechercher. Les travaux présentés dans cette thèse s'inscrivent dans ce même cadre, statistique et probabiliste.

Les travaux fondamentaux définissant la segmentation en locuteurs ont été réalisés par la société *BBN* sous la direction de H. Gish [Siu 1991, Siu 1992]. L'application consiste à indexer les transmissions d'une tour de contrôle de trafic aérien. Les conversations entre les contrôleurs et les pilotes sont enregistrées, puis segmentées par locuteur. Les enregistrements peuvent contenir plusieurs dialogues entre un contrôleur et un pilote. Après la segmentation en locuteurs, les dialogues contrôleur/pilote sont reconstruits. H. Gish propose dans ce travail une architecture reprise et complétée par la suite dans la majorité des systèmes de segmentation. Le processus se décompose en trois phases distinctes : la paramétrisation, la détection des ruptures et la classification.

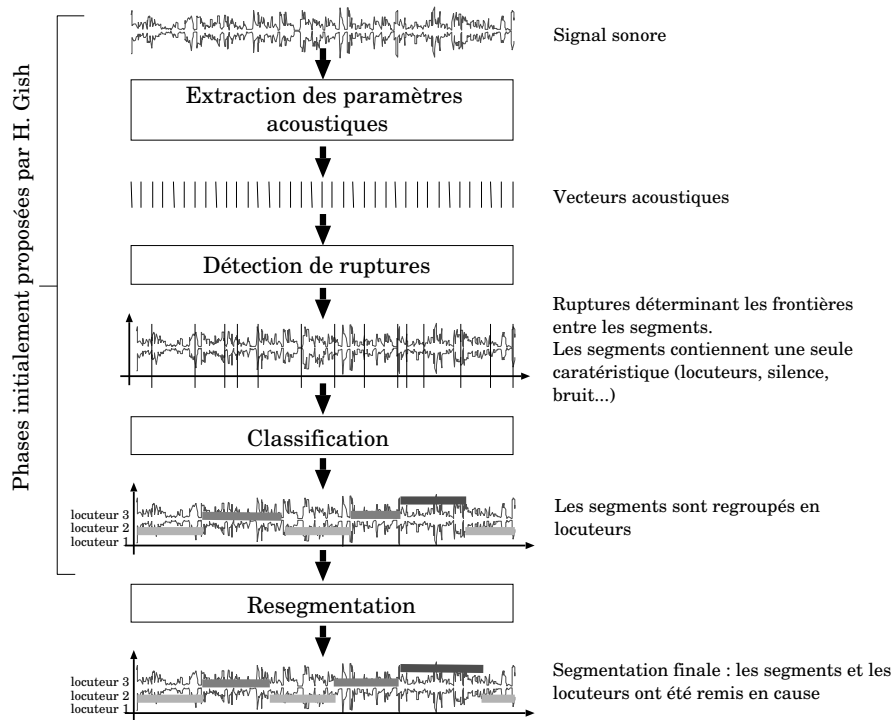


FIG. 3.1 – Principe général de la segmentation en locuteurs : le processus se décompose en quatre phases.

De nombreux chercheurs comme Douglas Reynolds du *MIT Lincoln Laboratory*, André Adami de *OGI* et Daniel Moraru du *CLIPS* ont complété l'architecture avec une phase finale de resegmentation du signal [Reynolds 2000, Adami 2002, Moraru 2002].

La figure 3.1 présente une vue générale de l'enchaînement des différentes phases de segmentation. Les éléments de cette architecture jouent les rôles suivants :

- La paramétrisation convertit le document sonore (l'enregistrement) en paramètres acoustiques.

- La détection des ruptures coupe le document en segments de petites tailles. Les segments créés contiennent un seul locuteur.
- La classification groupe les segments locuteur par locuteur jusqu'à découvrir le nombre de classes correspondant au nombre de locuteurs intervenant dans le document. Chaque classe contient à la fin de la classification l'ensemble des segments prononcés par un locuteur. La classe définit alors l'intervention du locuteur dans le document.
- Enfin, la resegmentation affine les frontières des segments. L'enregistrement est de nouveau découpé en segments en fonction des données contenues dans les classes.

Dans cette architecture, les hypothèses de travail définies au paragraphe 2.4 sont respectées. Aucune information *a priori* sur les locuteurs potentiels n'est utilisée. En particulier, le nombre de locuteurs est inconnu et aucun échantillon de voix des locuteurs n'est disponible.

3.2 Etat de l'art : outils pour la segmentation

3.2.1 Paramétrisation

Le processus de segmentation en locuteurs commence par l'étape de paramétrisation acoustique. Le signal n'est pas directement exploité. Il contient des informations redondantes ou non pertinentes pour notre étude qui doivent être supprimées (*c.f.* § 2.2). 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.

Bien que quelques résultats soient présentés pour différentes paramétrisations au paragraphe 5.3.1, elles ne font pas l'objet d'une étude complète. Nous invitons le lecteur à se reporter aux études des différentes paramétrisations couramment utilisées en reconnaissance du locuteur et de la parole : [Reynolds 1994, Homayounpour 1994, Charlet 1997]. Une étude de différentes paramétrisations pour la segmentation en catégories sonores¹ est proposée dans [Li 2001].

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 cepstraux sont retenus :

- Les vecteurs acoustiques issus d'une analyse en banc de filtre à échelle linéaire appelés LFCC (*Linear Frequency Cepstral Coefficients*).
- Les coefficients obtenus à partir d'une échelle de Mel² (MFCC, *Mel Frequency Cepstral Coefficients*).

Ces deux types de paramétrisation ont aussi l'avantage de produire des coefficients peu corrélés [Soong 1988]. Dans la suite du document, un vecteur acoustique est indifféremment appelé observation ou trame.

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 sont problématiques : pour des documents contenant différents canaux de transmission (conversation téléphonique), le système risque de reconnaître autant le canal de transmission que le locuteur.

¹Silence, locuteur, multi-locuteur, musique, bruit, locuteur+musique, multi-locuteur+musique.

²L'échelle de Mel est une échelle proche de la perception humaine.

Laboratoire	Type de tâche	Nombre de coefficients statiques	Référence
LIA	VAL	16 LFCC (+ delta, sans C_0)	[Fredouille 2000b]
LIA	segmentation	20 LFCC (+ E)	[Moraru 2003]
MIT	VAL	19MFCC + Delta (sans C_0)	[Reynolds 2000]
MIT	segmentation	24MFCC (C_0 ?)	[Reynolds 2000]
OGI	VAL	13MFCC + Delta + Delta Delta (sans C_0)	[Kajarekar 2002]
OGI	segmentation	24 LSP (C_0 ?)	[Adami 2002]
CLIPS	segmentation	16 MFCC + E	[Moraru 2003]

TAB. 3.1 – Comparaison du nombre de coefficients statiques pour les tâches de VAL et de segmentation : le type de coefficient et les coefficients complémentaires (différentiels et énergie) sont précisés. Ces paramètres sont définis pour des conversations téléphoniques. Delta : dérivées premières ; Delta Delta : dérivées secondes ; E : énergie ; C_0 : premier coefficient.

Des méthodes de compensation sont appliquées sur les paramètres pour atténuer les distorsions engendrées par le canal de transmission [Pelecanos 2001, Hermansky 1994]. 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.

Coefficients différentiels et énergie

Les vecteurs acoustiques sont complétés, si nécessaire, par les dérivées premières et secondes des vecteurs [Furui 1981]. 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", "Delta Delta". L'énergie du signal (notée E) et les dérivées de l'énergie sont aussi des coefficients utilisés.

Le tableau 3.1 reporte les types de coefficients de trois systèmes de reconnaissance du locuteur et de quatre systèmes de segmentation. En segmentation, les coefficients différentiels ne sont pas utilisés dans les systèmes présentés, l'énergie ou le premier coefficient cepstral (C_0) leur est préféré.

Bilan

Bien que la plupart des systèmes d'indexation en locuteurs utilisent les coefficients cepstraux, les méthodes de calculs des coefficients, les types de filtres, le nombre de coefficients, les coefficients complémentaires et les méthodes de compensation n'ont pas encore fait l'objet d'une étude aussi complète que dans les domaines de la RAL et de la RAP.

3.2.2 Modèles et mesures

La modélisation des locuteurs repose sur les études faites en reconnaissance automatique du locuteur (RAL) pour la tâche de vérification automatique du locuteur (VAL). L'approche statistique et probabiliste avec des modèles multi-gaussiens (*Gaussian Mixture Model*, GMM) proposée dans [Reynolds 1992, Reynolds 1994, Schroeder 2000] domine le domaine de la RAL depuis une dizaine d'années³. Pour attester ce fait, lors des évaluations NIST 2002 en VAL, seulement deux laboratoires sur les vingt et un participants ont proposé une modélisation des locuteurs différente des GMM (les systèmes de ces deux laboratoires reposent sur une approche connexionniste).

D'autres approches ont été proposées dans la littérature, nous citons à titre d'exemple l'approche prédictive [Grenier 1980], l'approche connexionniste [Oglesby 1990], les méthodes statistiques du

³Pour les tâches de vérification du locuteur indépendantes du texte où le système dispose d'environ deux minutes de parole pour l'apprentissage d'un modèle de locuteur et de 15 à 45 secondes pour les enregistrements de tests.

second ordre [Gish 1986, Bimbot 1995], la quantification vectorielle [Soong 1992], la programmation dynamique [Furui 1981] ou les modèles de Markov cachés [Poritz 1982, Rosenberg 1990]. Une classification de ces stratégies est proposée dans [Bonastre 2000].

Si les GMM sont devenus les modèles de référence en RAL, les modèles de Markov cachés (*Hidden Markov Model*, HMM) occupent la même place en reconnaissance de la parole. Les HMM ont été introduit en RAP depuis les années 1975 par [Backer 1975, Jelinek 1976].

Modèle multi-gaussien

• Définition

Une séquence de vecteurs acoustiques, correspondant dans notre cas à un locuteur, est modélisée par un GMM. Un GMM est une somme pondérée de gaussiennes multidimensionnelles. Chaque gaussienne du modèle est caractérisée par un vecteur moyen et une matrice de covariance.

Un GMM à n composantes est défini par :

- Un ensemble de poids $p = \{p_k\}_n$ dont la somme est égale à 1 (avec $p_k \in]0, 1]$).
- Un ensemble de vecteurs de moyenne $\mu = \{\mu_k\}_n$ et un ensemble de matrice de covariance $\Sigma = \{\Sigma_k\}_n$ ou chaque couple (μ_k, Σ_k) définit une distribution gaussienne multidimensionnelle de dimension d^4 égale à la dimension d'une observation.

Le GMM à n composantes du locuteur X_i est noté par $X_i(p, \mu, \Sigma)$ avec p l'ensemble des poids, μ l'ensemble des vecteurs de moyenne et Σ l'ensemble des matrices de covariance.

• Estimation d'un modèle

Un GMM est estimé à partir d'une séquence d'observations O , appelée données d'apprentissage, au moyen de l'algorithme EM-ML optimisant le critère du maximum de vraisemblance (*Expectation Maximisation - Maximum Likelihood*, [Dempster 1977]). La qualité de l'estimation du modèle est liée à la quantité de données d'apprentissage et au nombre de composantes du modèle.

Une autre approche est généralement utilisée quand la quantité de données d'apprentissage n'est pas suffisante pour estimer correctement le modèle par l'algorithme EM-ML. Elle consiste à adapter un modèle GMM (noté W) connu *a priori* à une séquence d'observations pour obtenir le modèle de locuteur. Le modèle W est généralement un modèle appris, au moyen de l'algorithme EM-ML, sur un grand nombre de locuteurs. Ce modèle est généralement nommé modèle du monde [Carey 1992], il a pour fonction de modéliser une population générique de locuteurs.

L'estimation du modèle X_i est obtenue par les méthodes d'adaptation MAP (*Maximum A Posteriori*, [Gauvain 1994]) ou MLLR (*Maximum Likelihood Linear Regression*, [Leggetter 1995]). Les méthodes d'adaptation ont pour fonction de dériver le modèle de locuteur du modèle générique W .

• Mesure

La probabilité qu'une séquence d'observations O soit émise par le modèle de locuteur X_i est approchée par l'estimation du maximum de vraisemblance, notée $l(o_t|X_i)$. Elle est donnée par la somme pondérée des vraisemblances des gaussiennes du modèle X_i :

$$l(o_t|X_i) = \sum_{k=0}^{k=n} p_k g_k(o_t) \quad (3.1)$$

avec

$$g_k(o_t) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(o_t - \nu_k)^T (\Sigma_k)^{-1} (o_t - \nu_k)\right)$$

⁴La matrice de covariance est diagonale dans la plupart des systèmes de vérification.

La vraisemblance entre le modèle X_i et la séquence d'observations $O = \{o_1, \dots, o_t, \dots, o_T\}$ correspond à la moyenne géométrique des vraisemblances, cette moyenne est notée :

$$l(O|X_i) = \sqrt[t]{\prod_{t=1}^T l(o_t|X_i)} \quad (3.2)$$

• **Test d'hypothèses**

Dans un processus de vérification du locuteur, une séquence d'observations O prononcées par un locuteur est comparée au modèle de locuteur X_i correspondant à l'identité revendiquée \mathcal{X}_i . Le processus s'appuie sur un test d'hypothèses formé par [Liu 1996, Furui 1997, Gravier 1998, Fredouille 2000a] :

- H_0 : la séquence d'observation O a été émise par le modèle X_i
 - H_1 : la séquence d'observation O a été émise par le modèle \overline{X}_i
- avec la règle de décision :

$$\begin{cases} \frac{P(H_0)}{P(H_1)} \geq \alpha \implies \text{acceptation} \\ \frac{P(H_0)}{P(H_1)} < \alpha \implies \text{rejet} \end{cases} \quad (3.3)$$

Le modèle \overline{X}_i représente tous les locuteurs différents de X_i . Quand la population de locuteurs est inconnue, \overline{X}_i peut être approché par le modèle générique de locuteurs W [Carey 1992, Reynolds 1997]. $P(H_0)$ et $P(H_1)$ sont approchées par les vraisemblances $l(O|X_i)$ et $l(O|W)$. La règle de décision de l'équation 3.3 devient :

$$\begin{cases} lr(O|X_i) = \frac{l(O|X_i)}{l(O|W)} \geq \alpha \implies \text{acceptation} \\ lr(O|X_i) = \frac{l(O|X_i)}{l(O|W)} < \alpha \implies \text{rejet} \end{cases} \quad (3.4)$$

$lr(O|X_i)$ est appelé rapport de vraisemblance. Le paramètre α est le seuil de décision à fixer en fonction de l'application visée. Il représente le risque que l'utilisateur est disposé à accepter [Saporta 1990, Bonastre 2000].

Modèle de Markov caché

Les modèles de Markov cachés (*Hidden Markov Model*, HMM) sont abondamment documentés dans la littérature. Pour plus de détails, le lecteur se reportera au tutoriel de [Rabiner 1989] ou aux thèses de [Lefevre 2000] et [Barras 1996]. Dans ce paragraphe seulement la définition d'un modèle de Markov caché et la probabilité d'une séquence d'observations sont données.

• **Définition**

Un HMM est un automate probabiliste d'états finis, qui change d'état à chaque unité de temps (figure 3.2). A chaque état est associée une fonction de densité de probabilités d'émission et chaque transition entre états porte une probabilité. A l'instant t , l'automate est dans l'état i , un vecteur o_t est engendré avec la probabilité $b_i(o_t)$ avec b_i la fonction de densité de probabilités d'émission associée à l'état i . b_i est fréquemment un GMM⁵. La probabilité $b_i(o_t)$ est alors approchée par la vraisemblance du modèle pour l'observation o_t . La transition d'un état i à un état j prend la probabilité a_{ij} . Le modèle de Markov caché λ est défini par (E, A, B) :

⁵Remarque : un HMM à un état est équivalent à un GMM.

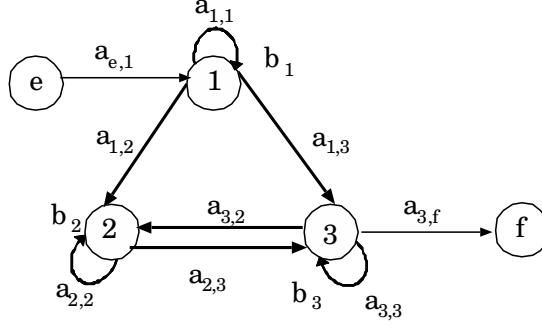


FIG. 3.2 – Exemple de modèle de Markov à 3 états.

- $E = \{1, \dots, i, \dots, N\}$ un ensemble d'états. L'état i est associé au modèle X_i .
- $A = \{a_{i,j}\}$ un ensemble de probabilités de transition entre les états. La somme des probabilités sortant d'un état est égale à 1 (équation 3.5).

$$\sum_{j \in E} a_{i,j} = 1 \quad (3.5)$$

- $B = \{b_i\}$. A chaque état i est associée une fonction de densité de probabilité d'émission b_i . La probabilité du modèle X_i pour l'observation o_t est notée $b_i(o_t)$.

Deux états non-émetteurs (e, f) sont ajoutés au HMM, aucune fonction de densité de probabilité ne leur est associée. Ces états e et f permettent d'interconnecter des HMM entre eux et d'imposer, si nécessaire, des contraintes sur les choix du premier et du dernier état émetteur. Dans la figure 3.2, l'automate commence dans l'état e pour se rendre obligatoirement dans l'état 1 et le processus finit dans l'état f en venant de l'état 3. En général ces états ne sont pas représentés dans les schémas et ils n'interviennent pas dans le compte du nombre d'états.

• Probabilité d'une séquence d'observations

La probabilité, $P(O, Q|\lambda)$, qu'une séquence d'observations $O = \{o_1, o_2, o_3, \dots, o_t, \dots, o_T\}$ passant par la séquence d'états Q ait été émise par le HMM λ , est le produit des probabilités d'émission et des probabilités de transition suivant la séquence d'états Q (équation 3.6).

$$\begin{aligned} Q &= \{q_0, q_1, q_2, q_3, \dots, q_t, \dots, q_T, q_{T+1}\} = e, 1, 2, 1, \dots, 2, \dots, 3, f \\ &\text{avec} \\ P(O, Q|\lambda) &= a_{e,1} b_1(o_1) a_{1,2} b_2(o_2) a_{2,1} b_1(o_3) \dots b_2(o_t) \dots b_3(o_T) a_{3,f} \\ &\text{qui peut s'exprimer sous la forme} \\ P(O, Q|\lambda) &= a_{q_0 q_1} \prod_{t=1}^{t=T} b_{q_t}(o_t) a_{q_t q_{t+1}} \end{aligned} \quad (3.6)$$

• Estimation de la séquence d'états

Généralement seule la séquence d'observations O est connue, la séquence d'états Q à travers le modèle est cachée. Pour obtenir cette séquence, une des solutions est de calculer la séquence de probabilités maximale $\hat{P}(O|\lambda)$ par l'algorithme de décodage de Viterbi [Viterbi 1967]. Cet algorithme permet d'obtenir l'alignement (*i.e.* la séquence d'états cachés) d'une séquence d'observations sur un HMM. L'algorithme de décodage de Viterbi résout l'équation 3.7.

$$\hat{P}(O|\lambda) = \max_Q \left(a_{q_0 q_1} \prod_{t=1}^{t=T} b_{q_t}(o_t) a_{q_t q_{t+1}} \right) \quad (3.7)$$

Cet algorithme optimise le critère du Maximum de Vraisemblance. Parmi l'ensemble des séquences d'états possibles, l'algorithme Viterbi sélectionne la séquence d'états de vraisemblance maximum.

3.3 Etat de l'art : segmentation en locuteurs

Les méthodes de segmentation en locuteurs décrites dans la littérature se décomposent en quatre étapes : la paramétrisation, la détection de ruptures, la classification et la resegmentation (*c.f.* § 3.1). Un survol des techniques pour la paramétrisation a été donné au paragraphe 3.2.1. La composition d'une segmentation est définie avant de décrire les trois dernières étapes du système classique de segmentation.

3.3.1 Segmentation

Une segmentation est constituée d'un ensemble de segments. Chaque segment est caractérisé par ses instants de début et de fin ainsi que par un libellé de locuteur.

Pour un document X représenté par les vecteurs acoustiques $\{o_1, \dots, o_T\}$, une segmentation en n segments est définie par $S = \{s_i\}_n$ avec pour chaque segment $s_i = (d_i, f_i, e_i)$, d_i l'instant de début, f_i l'instant de fin et e_i le libellé du locuteur associé. s_i vérifie les conditions suivantes :

$$\begin{cases} d_i > 0 \\ d_i > f_i \\ f_i < T \\ e_i \in \mathcal{X}, \text{ l'ensemble des locuteurs du document.} \end{cases} \quad (3.8)$$

Cette méthode impose aux segments d'avoir des instants de début et de fin en accord avec la durée de l'enregistrement. Cependant, aucune contrainte n'est définie sur le positionnement des segments les uns par rapport aux autres. Les segments peuvent se chevaucher dans le cas où les locuteurs s'exprimeraient en simultanée (figure 3.3).

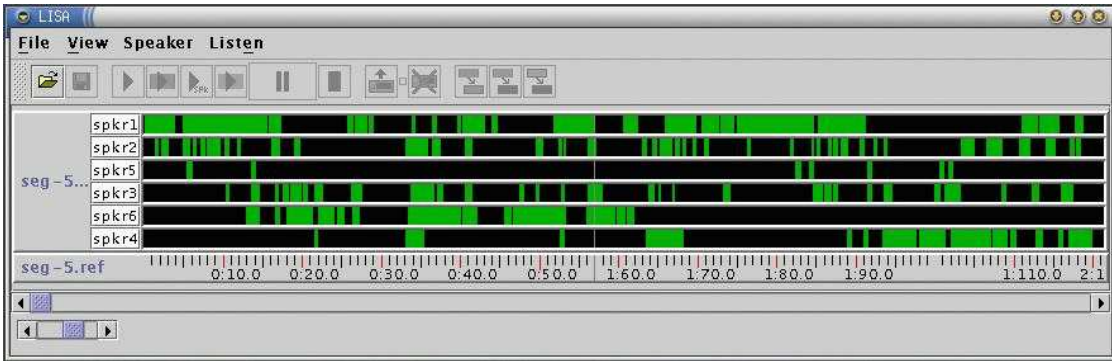


FIG. 3.3 – Exemple de segmentation en 5 locuteurs : les locuteurs s'expriment en même temps, les zones claires correspondent à des segments, le fichier est extrait de la base Meeting NIST.

En supposant que les locuteurs ne parlent pas en même temps, une autre méthode pour définir une segmentation est d'associer à chaque observation une étiquette de locuteur. Pour la séquence d'observations $O = \{o_1, \dots, o_t, \dots, o_T\}$, la segmentation S est définie par :

$$\begin{cases} S = \{s_1, \dots, s_i, \dots, s_T\} = \{e_1, \dots, e_i, \dots, e_T\} \\ \text{avec} \\ e_i \in \mathcal{X} \end{cases} \quad (3.9)$$

Cette dernière représentation est utilisée dans la suite de ce document. Dans les hypothèses énoncées au paragraphe 2.4, il est noté que les méthodes présentées supposent que les locuteurs ne se coupent pas la parole.

3.3.2 Détection des ruptures

Généralement, la détection des ruptures est intégrée dans des processus plus complexes comme la reconnaissance du locuteur, la reconnaissance de la parole ou dans notre cas la segmentation. Le document à traiter est segmenté en zones homogènes les plus longues possibles contenant un seul type de caractéristiques acoustiques (ici un locuteur). La méthode la plus couramment utilisée détecte les instants dans le signal où il existe un changement de caractéristique acoustique (figure 3.4). Les ruptures déterminent alors les frontières des segments contenant une seule caractéristique. Seuls les changements instantanés ou d'une très courte durée sont détectés [Seck 2001].

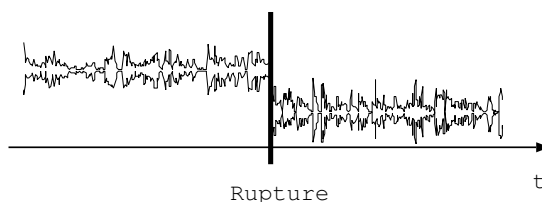


FIG. 3.4 – Exemple de rupture.

Les trois approches communément utilisées reposent sur :

- La détection des silences.
- La détection de changements de caractéristique acoustique (sans identifier la nature des changements).
- L'identification de la nature des segments (par exemple : parole, silence, musique, ...).

Ces méthodes sont bien documentées dans les thèses de Perrine Delacourt et de Mouhamadou Seck [Delacourt 2000a, Seck 2001]. Thomas Kemp, dans l'article [Kemp 2000], propose une évaluation de ces trois différentes méthodes. Le principe général de chacune des méthodes est décrit dans les paragraphes suivants.

Détection des silences

La détection des ruptures sur les silences suppose qu'il existe des silences entre les interventions des locuteurs (figure 3.5). Les silences sont généralement caractérisés par un faible niveau d'énergie. Les zones de faible énergie suffisamment longues sont alors considérées comme étant des silences. La détection des silences est liée à deux paramètres :

- Un seuil de décision (à partir de quelle énergie décider que le signal est du silence).

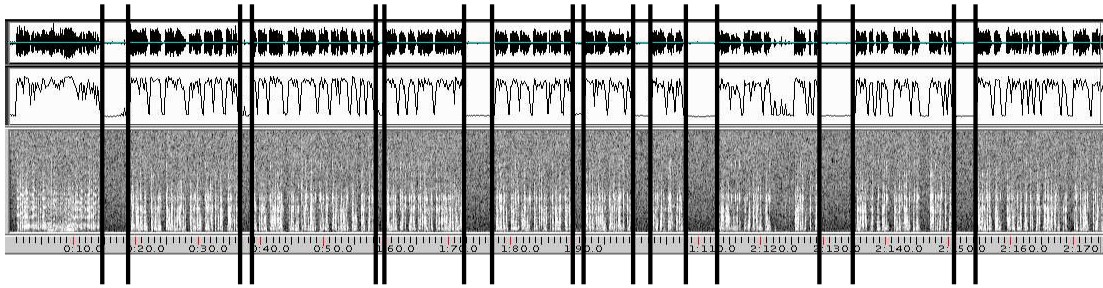


FIG. 3.5 – Exemple de rupture sur les silences : les ruptures délimitent les frontières entre la parole et les silences.

- Une durée minimale des segments de silences (à partir de quelle durée de silence considérer qu'il existe des interventions de part et d'autre du silence détecté).

Le premier paramètre, le seuil, dépend des conditions d'enregistrement et en particulier du bruit de fond. Il peut être très différent d'un enregistrement à l'autre et reste difficile à déterminer automatiquement. Le deuxième paramètre, la durée, présente également une difficulté. Fixer la durée à une valeur trop importante ne permet pas de détecter les interventions de locuteurs, alors qu'une durée trop courte entraîne généralement la détection de tous les silences même ceux présents entre les phonèmes.

Il semble raisonnable de penser que cette méthode n'est pas adaptée aux enregistrements comme les conversations téléphoniques ou les réunions. Dans ces types de discours, les locuteurs se coupent souvent la parole. Les histogrammes de la figure 3.6 montrent que les corpus de parole spontanée (*Meeting*, *Switchboard 2000*, *Switchboard 2002* et *CallHome*) contiennent peu de silence entre les segments (barres pleines), alors que l'histogramme du corpus de parole préparée (*Broadcast News*) montre qu'il existe toujours des silences entre les segments.

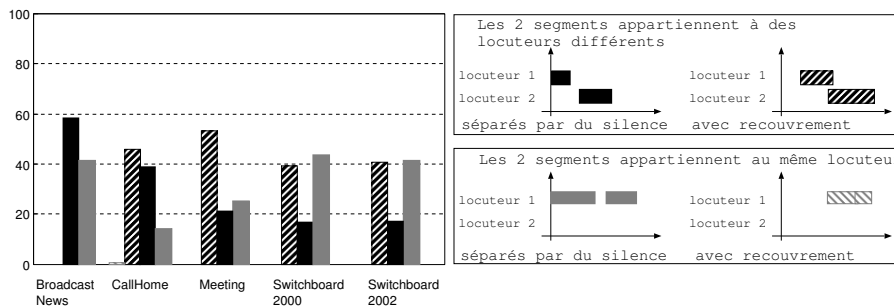


FIG. 3.6 – Histogramme des types de décalages entre les segments : 2 segments consécutifs sont séparés par du silence ou deux segments consécutifs se recouvrent. Deux cas sont considérés : le cas où les segments appartiennent au même locuteur ou à deux locuteurs différents. Corpus de parole spontanée : *Meeting*, *Switchboard 2000*, *Switchboard 2002* et *CallHome* ; corpus de parole préparée : *Broadcast News*.

Douglas Reynolds propose de palier le problème du choix des paramètres et de la variabilité des corpus en découpant, après la détection des ruptures, les segments en sous segments de taille fixe d'une seconde [Reynolds 2000]. La détection des silences surestime le paramètre portant sur la durée minimale des silences, ainsi seuls les longs silences sont détectés. Puis le découpage des segments en sous segments courts de taille fixe garantit qu'une majorité des segments sont homogènes.

Détection des changements de caractéristique acoustique :

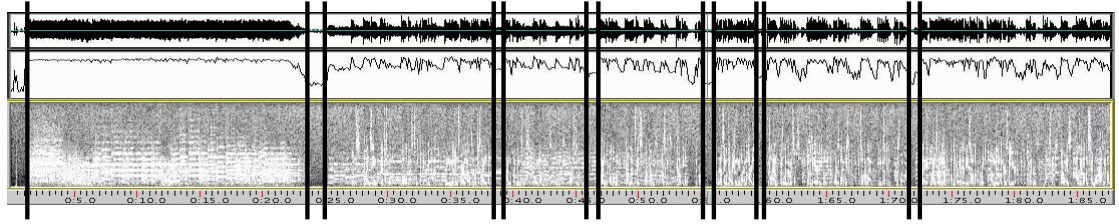


FIG. 3.7 – Exemple de détection des changements de caractéristique acoustique : les frontières délimitent des changements entre de la parole, de la musique, du silence...

La détection des changements de caractéristique acoustique est une extension de la détection des silences. Au lieu de chercher les ruptures entre les zones de silence et de non-silence, les changements de caractéristique acoustique sont détectés (figure 3.7). La méthode ne détermine pas les caractéristiques présentes dans le signal mais seulement leurs changements. Par exemple, les changements de locuteurs ou les changements entre de la parole et de la musique sont recherchés.

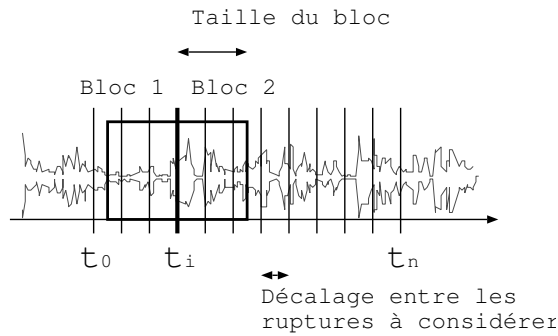


FIG. 3.8 – Méthode de détection des changements de caractéristique acoustique : une similarité est calculée entre les blocs 1 et 2 pour tous les instants t . La décision de la présence d'une rupture est prise par rapport à un seuil.

La méthode généralement proposée calcule une mesure de similarité à un instant t entre deux blocs de signal consécutifs [Delacourt 2000a]. La décision de la présence d'une rupture à l'instant t est prise en comparant la valeur de la mesure à un seuil. Plusieurs problèmes découlent de cette approche :

- le choix de la mesure,
- la durée de décalage de la rupture à l'instant t ,
- la taille des blocs de part et d'autre de la rupture,
- et le seuil de décision (à partir de quelle mesure considérer qu'une rupture dans le signal est présente).

Une des mesures les plus utilisées est le rapport de vraisemblance généralisé (*Generalized Likelihood Ratio*, GLR) ou son extension au critère BIC (*Bayésien Information Criterion* [Chen 1998, Delacourt 2000b]).

Comme dans le cas de la détection des silences, il est nécessaire de fixer un seuil dépendant en partie du type de signaux et du discours à traiter. La taille des deux blocs d'analyse est aussi un

paramètre à fixer. Cette taille est contrainte par la quantité de données nécessaire pour estimer la mesure de similarité et par la taille des segments. La parole spontanée (conversation téléphonique, réunion) a généralement une durée moyenne de segments plus courte que la parole préparée (interview).

Détection des ruptures par identification de la nature des segments

Une autre approche du problème est d'identifier la nature des zones de signal. Les ruptures sont alors présentes entre les zones de signal consécutives de natures différentes.

La méthode la plus utilisée s'appuie sur le calcul de vraisemblance à partir de GMM ou sur un modèle de Markov caché pour détecter les changements de caractéristiques [Gauvain 1999, Hain 1998, Koolwaaij 2000]. Par exemple, un modèle de Markov ergodique (où tous les états sont interconnectés) est construit. Chaque état correspond à une caractéristique sonore particulière (silence, parole, musique en général). Les transitions modélisent les changements de caractéristiques entre deux instants. L'algorithme de Viterbi permet d'obtenir l'étiquetage suivant les caractéristiques définies de chacun des vecteurs acoustiques du signal.

Cette approche nécessite un corpus d'apprentissage pour construire le modèle de chacune des classes acoustiques. Mais aucun seuil n'est nécessaire, contrairement aux deux précédentes méthodes.

Cette approche est adaptée à la segmentation du signal suivant des caractéristiques connues d'avance, par contre elle n'est pas adaptée à la détection des ruptures entre les locuteurs. Par hypothèse, les locuteurs ne sont pas connus du système.

Discussion et bilan

L'analyse des résultats des campagnes NIST pour la tâche de segmentation en locuteurs sur des conversations téléphoniques montre que la méthode de détection de ruptures la plus performante reste la détection sur les silences. Des stratégies plus complexes à mettre en œuvre, comme la détection des changements par la mesure BIC, n'améliorent pas les performances de la segmentation en locuteurs [Reynolds 2000].

Par contre, Thomas Kemp dans [Kemp 2000] teste les trois méthodes sur un corpus de journaux télévisés allemands. Il n'arrive pas à la même conclusion. Les meilleures performances⁶, en termes de détection de ruptures, sont obtenues avec la mesure BIC. Les segments de ces journaux télévisés ont une durée moyenne de 25 secondes. Cette durée moyenne est très largement supérieure aux durées des segments des corpus NIST de parole spontanée. Les mesures de similarité BIC peuvent être calculées à partir de fenêtres de grande taille.

De l'ensemble de ces travaux, il se dégage que le choix de la méthode et de ses paramètres est fortement lié à la nature du corpus à traiter. Pour la parole spontanée, une détection sur les silences est suffisante. Pour la parole préparée, la détection des changements de caractéristiques par la méthode BIC semble la plus appropriée.

3.3.3 Classification : regroupement des segments

Suite à la détection de ruptures, un ensemble initial de segments $S = \{s_1, \dots, s_i, \dots, s_n\}$ est disponible. Chaque segment s_i contient un seul locuteur. Le regroupement cherche la partition P en classes des segments telle que chaque classe contient les segments d'un locuteur. Nous rappelons

⁶Alors que, les conclusions de D. Reynolds sont données après application de la classification et de la resegmentation.

que - par hypothèse - le nombre de locuteurs n'est pas connu et qu'aucun échantillon de la voix des locuteurs n'est disponible. La classification est qualifiée de "non-supervisée".

La classification non-supervisée d'objets (*i.e.* les segments) n'est pas un problème spécifique au domaine de la parole. Une branche des mathématiques étudie ces problèmes. Les méthodes appliquées à notre problème de regroupement en locuteurs sont généralement issues de ce domaine des mathématiques.

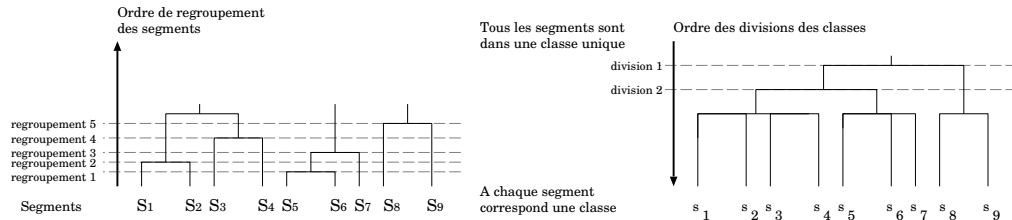


FIG. 3.9 – *Classification hiérarchique : approche descendante v.s. approche ascendante.*

La classification hiérarchique est la méthode généralement proposée dans la littérature [Siu 1992, Wilcox 1994, Johnson 1998, Delacourt 2000a, Reynolds 2000]. Dans cette famille, deux approches itératives sont utilisées : l'approche ascendante et l'approche descendante.

- L'approche descendante commence par placer tous les objets dans une seule classe. Cette classe unique est divisée en n sous-classes. Puis les sous-classes sont encore divisées jusqu'à obtenir des classes ne contenant qu'un seul objet.
- La classification ascendante commence avec une classe par objet. A chaque itération, n classes sont fusionnées jusqu'à obtenir une classe contenant tous les objets.

Dans les deux approches, le résultat d'une classification est présenté sous la forme d'un arbre appelé dendrogramme qui illustre les regroupements/divisions faits à chaque itération (figure 3.9). Il est aussi nécessaire de définir une règle de regroupement/division et un critère de sélection de la partition finale.

Classification hiérarchique descendante

• Méthode

Une approche descendante a été proposée par Sue E. Johnson dans [Johnson 1998, Johnson 1999]. La méthode initialise la classification en plaçant l'ensemble des observations dans une classe (la racine de l'arbre). A chaque itération, les classes sans fils (les feuilles de l'arbre) sont divisées en quatre sous-classes. Puis les données sont réaffectées dans les feuilles tant que la vraisemblance des données augmente. Une phase de combinaison et de fusion est appliquée à la suite de chaque division. Les classes recombinaées ou fusionnées ne sont plus divisées aux itérations suivantes.

• Evaluation

La méthode a été utilisée dans le cadre de la reconnaissance automatique de la parole sur le corpus *Hub4-1997* dans le cadre des évaluations NIST [LDC 2002]. L'objectif était de construire des classes définissant des locuteurs pour l'adaptation MLLR des modèles de parole. Les performances de la méthode ont été mesurées selon deux critères :

- Le nombre de segments détectés.
- Le nombre moyen de locuteurs par segment détecté.

• Résultat

La méthode proposée est comparée à la référence (déterminée manuellement) et à un autre système automatique développé par le CMU⁷. Les résultats obtenus suivant les deux critères sont :

- La méthode descendante engendre un nombre de segments plus proche de la référence que la méthode ascendante du CMU (Référence : 634, classification descendante : 749, CMU : 769).
- Le nombre moyen de locuteurs par segment est aussi plus proche de la référence (Référence : 1, classification descendante : 1,173, CMU : 1,239).

• Commentaires

La méthode d'évaluation utilisée ne correspond pas au critère utilisé au évaluation NIST pour la tâche de segmentation en locuteurs (*c.f.* § 13), toutefois le gain après adaptation des modèles de parole aux segments est de 0,9% en taux d'erreur de reconnaissance de la parole en absolu par rapport à la méthode de segmentation CMU.

Classification hiérarchique ascendante

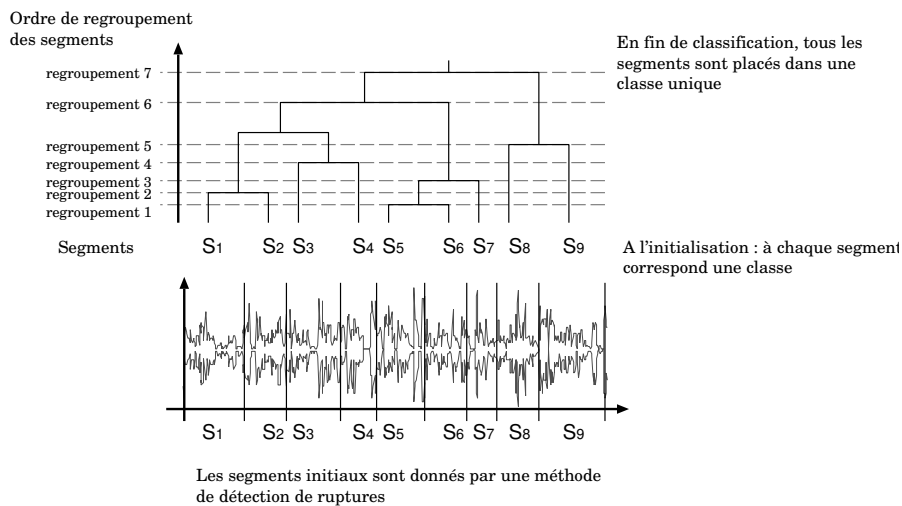


FIG. 3.10 – Classification hiérarchique ascendante : regroupement en classes.

La classification hiérarchique ascendante est la principale méthode proposée dans la littérature. Pour les aspects théoriques de la classification hiérarchique ascendante, le lecteur pourra se référer à [Everitt 1993] par exemple. Une implémentation de cette technique est décrite dans [Struyf 1997].

• Principe général

La classification initiale est composée d'un ensemble de classes, où chaque classe contient un seul objet (ici un segment déterminé par une des méthodes de détection des ruptures). A chaque étape, l'algorithme groupe les deux classes les plus proches selon une mesure de dissimilarité calculée pour tous les couples de classes. L'algorithme s'arrête quand toutes les classes ont été regroupées en une seule classe. Un dendrogramme qui illustre les associations successives est présenté dans la figure 3.10. La partition finale est construite en sélectionnant un ensemble de nœuds (définissant des classes) parmi tous les nœuds du dendrogramme.

Une méthode de classification hiérarchique ascendante est définie par :

⁷Le système CMU (*Carnegie Mellon University*) est utilisé par NIST dans les évaluations en reconnaissance de la parole [Siegler 1997], il repose sur une technique de classification hiérarchique ascendante utilisant une mesure de Kullback-Leibler.

- Une mesure entre les classes pour choisir les classes à fusionner, l'ensemble de ces mesures forme une matrice.
- Une règle d'agglomération des classes. Après la fusion de deux classes, il est nécessaire de réévaluer les mesures entre cette nouvelle classe et les autres.
- Une méthode d'élagage du dendrogramme pour sélectionner la partition finale.

• **Mesures de dissimilarité et matrice de dissimilarité**

Une mesure de dissimilarité d exprime la non-proximité entre deux classes (u, v) contenant des segments. Les mesures de dissimilarité $d(u, v)$ vérifient les conditions suivantes [Saporta 1990] :

$$\begin{cases} d(u, v) = d(v, u) \\ d(u, v) \geq 0 \\ d(u, u) = 0 \end{cases} \quad (3.10)$$

Différentes mesures de dissimilarité sont proposées dans la littérature. Ces mesures sont généralement issues de modèles gaussiens et reposent sur la mesure de vraisemblance présentée au paragraphe 3.2.2.

Les mesures les plus courantes sont les suivantes :

- Le rapport de vraisemblance généralisé (*Generalized Likelihood Ratio*, GLR ; [Siu 1992, Solomonoff 1998]).
- Le critère d'information bayésien (*Bayesian Information Criterion*, BIC ; [Chen 1998]).
- Le rapport de vraisemblance croisé (*Cross Likelihood Ratio*, CLR ; [Reynolds 1998]).
- L'entropie croisée (*Cross entropy*, EC ; [Solomonoff 1998]).

Les deux premières méthodes montrent les coûts de calcul les plus importants. Elles nécessitent de calculer les modèles associés aux classes u , v et uv (la classe fusionnant les données de u et v). Mais elles mènent à des classifications de bonne qualité. Les deux dernières méthodes sont moins contraignantes, elles n'utilisent que des modèles des classes u et v ; mais elles sont moins performantes.

Une matrice de dissimilarité de dimension n est produite à partir d'une de ces mesures. Cette matrice est composée des mesures entre toutes les paires de classes (u, v) . Les dissimilarités étant symétriques, la matrice des dissimilarités est exprimée sous la forme d'une matrice carrée triangulaire supérieure. La diagonale de la matrice n'est pas utilisée : une classe ne peut pas être fusionnée avec elle-même.

A chaque étape de l'algorithme, la paire de classes (parmi les classes disponibles) montrant la dissimilarité la plus petite est fusionnée en une nouvelle classe (équation 3.11). La dimension de la matrice est réduite de 1. Toutes les paires de dissimilarités entre la nouvelle classe et les autres classes sont réestimées.

$$(u, v) = \underset{x, y \in \{C\}}{\text{ArgMin}} d(x, y) \quad (3.11)$$

avec

$$x \neq y ; u, v, x, y \in \{C\}, \text{ l'ensemble des classes disponibles}$$

$$d(x, y), \text{ une mesure de dissimilarité}$$

• **Réestimation des dissimilarités pour les nouvelles classes**

Après chaque fusion, les dissimilarités entre la nouvelle classe et les autres classes sont réévaluées.

Deux stratégies sont envisageables :

- Soit le modèle de la nouvelle classe uv est appris, puis les dissimilarités $d(uv, \cdot)$ entre la classe uv et les autres sont calculées.

- Soit une méthode d'agglomération permet d'obtenir les dissimilarités $d(uv, \cdot)$ à partir des dissimilarités $d(u, \cdot)$ et $d(v, \cdot)$.

Les méthodes d'agglomération sont abondantes dans la littérature [Solomonoff 1998, Jain 1999, Everitt 1993, Johnson 1997]. Elles sont très efficaces en terme de coût de calcul, mais ne garantissent de bons résultats que si les mesures vérifient la propriété de l'inégalité triangulaire (dans ce cas la mesure de dissimilarité est une distance).

- **Sélection de la partition finale**

Comme vu précédemment, à la fin de l'algorithme de classification hiérarchique, un dendrogramme est construit dans lequel chaque nœud correspond à la fusion de deux classes. La partition est définie en sélectionnant les nœuds à conserver (la partition finale doit contenir tous les segments). Plusieurs techniques existent dans la littérature [Solomonoff 1998, Everitt 1993] pour générer la partition. Ces techniques (figure 3.11) consistent à couper le dendrogramme à une hauteur donnée ou à sélectionner un ensemble de classes à différentes hauteurs (élagage). Dans les deux cas (coupe et élagage), un critère de sélection est nécessaire.

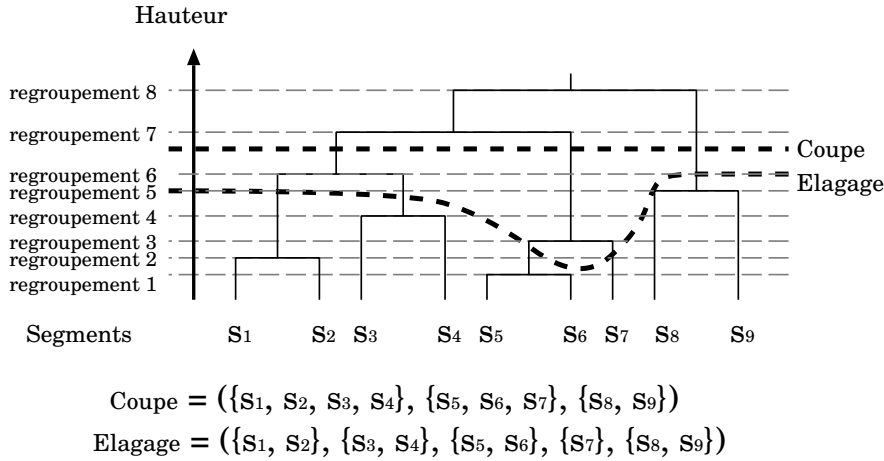


FIG. 3.11 – Classification hiérarchique : sélection de la partition finale par coupe horizontale ou par élagage.

3.3.4 Resegmentation

A partir de la partition obtenue à la fin de la classification hiérarchique, un modèle de locuteur est appris pour chacune des classes. Les frontières des segments définies lors de la détection des changements de locuteurs, sont remises en cause [Reynolds 2000, Adami 2002]. Chaque observation est attribuée au modèle de locuteur le plus vraisemblable. Les vraisemblances entre deux trames consécutives peuvent avoir une grande variabilité alors que ces trames ont une forte probabilité d'appartenir au même locuteur. Pour corriger ce problème, la vraisemblance d'une trame pour un modèle est lissée en utilisant la moyenne des vraisemblances, calculée à l'aide d'une fenêtre glissante⁸. Si l'algorithme peut être logiquement itéré, D. Reynolds constate une dégradation de la qualité de la segmentation lorsque l'algorithme est appliqué itérativement.

Si la phase de resegmentation est optionnelle, elle semble apporter un gain de performance significatif. Ce point est confirmé par les résultats obtenus par le CLIPS [Moraru 2002], où la

⁸A. Adami [Adami 2002] emploie une fenêtre de Hamming.

Laboratoire	Taux d'erreur	Pureté	Nb de loc. détecté	Nb de loc.	Corpus	Référence
CLIPS	11%	-	2	fixé à 2	cellulaire	[Moraru 2002]
Eurecom	-	81,1%	~10	non fixé	filaire	[Delacourt 2000c]
MIT	10%	-	2	fixé à 2	filaire	[Reynolds 2000]
OGI	7%	-	2	fixé à 2	filaire	[Adami 2002]

TAB. 3.2 – Comparaison des performances de systèmes de segmentation utilisant une classification hiérarchique ascendante. Taux d'erreur : erreur d'affectation définie par NIST en 2000/2001 pour les systèmes CLIPS, OGI et MIT. Pureté : qualité de chacune des classes. Nb de loc. détectés : nombre de locuteurs détectés par le système. Nb de loc. : nombre de locuteurs définis par le système. Corpus : conversations téléphoniques cellulaires ou filaires. Les systèmes sont évalués sur des corpus proches : SwitchBoard cellulaire ou filaire.

phase de resegmentation apporte un gain de 2%, en absolu, sur l'erreur d'affectation⁹.

3.3.5 Performance de la classification hiérarchique ascendante

Le tableau 3.2 indique les performances obtenues par différents systèmes de segmentation sur des conversations téléphoniques à deux locuteurs (cellulaires ou filaires). Les systèmes CLIPS, MIT et OGI sont évalués au moyen des outils NIST 2000/2001. Ces trois systèmes obtiennent des performances proches de 10% d'erreur. Ce score est bon au vu de la difficulté de la tâche. Une segmentation triviale composée d'un seul segment couvrant l'ensemble du document obtient un taux d'erreur d'environ 30%.

Le système Eurecom est évalué par deux critères : le nombre de locuteurs détectés et la pureté des classes¹⁰. Ces mesures ne sont pas comparables avec la mesure proposée par NIST, toutefois le système Eurecom est le seul système présenté ici qui détermine automatiquement le nombre de locuteurs présents dans les documents.

3.4 Commentaires

3.4.1 Structure du modèle classique

L'approche classique propose une structure linéaire pour la segmentation en locuteurs des documents sonores. Chacun des principaux modules (détection des ruptures, classification et resegmentation) ne communique les nouvelles informations que dans un sens :

1. La détection des ruptures propose une segmentation initiale du signal en segments homogènes à la classification.
2. La segmentation engendrée par la classification initialise les modèles de locuteurs de la segmentation.

Cette structure rend impossible l'utilisation des informations apprises dans un module de haut niveau par les modules de plus bas niveaux. Par exemples :

- Les modèles de locuteurs appris lors de la classification pourraient être utiles lors de la détection des ruptures.

⁹Ce résultat est obtenu à partir d'un sous-ensemble d'enregistrements issus des évaluations NIST 2001 en reconnaissance du locuteur pour la tâche de segmentation [NIST 2001].

¹⁰Qui correspond au rapport entre la durée des segments du locuteur majoritaire sur la durée des segments de la classe.

- La remise en cause des segments lors de la resegmentation pourrait aider la classification des segments et la détection des ruptures.
- Les modèles de locuteur calculés dans la phase de resegmentation peuvent aussi être utiles dans les deux phases précédentes.

3.4.2 Détection de ruptures

Les méthodes de détection de ruptures essaient de concilier deux critères antinomiques :

- Les segments doivent être longs pour apprendre des modèles robustes dans la phase de classification.
- Les segments doivent contenir un seul locuteur : la classification suppose que les segments initiaux sont mono-locuteur.

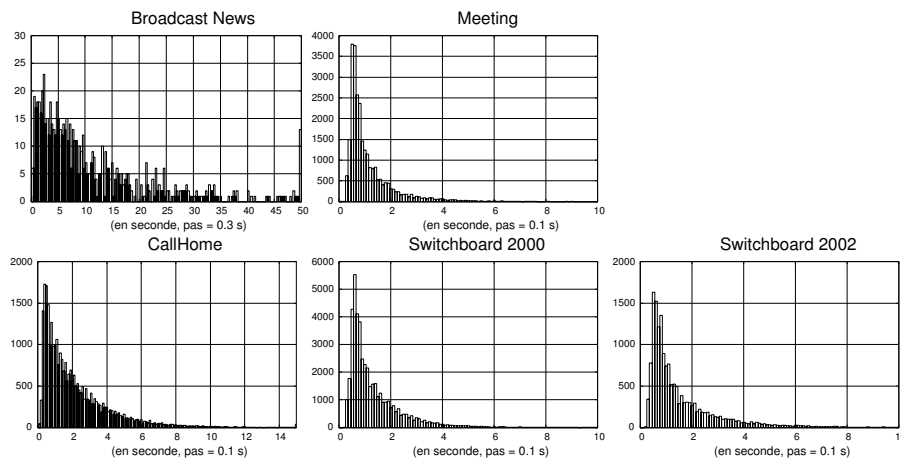


FIG. 3.12 – Histogramme des durées des segments pour différents corpus.

Les paramètres, en particulier les seuils de décision, dépendent du type de documents. Pour les conversations spontanées, la durée des segments est généralement inférieure à 3 secondes, alors que pour les conversations préparées, la durée des segments est plus élevée et variable. La figure 3.12 montre la répartition de la durée segmentale pour quatre corpus de parole spontanée (*Meeting*, *Switchboard 2000*, *Switchboard 2002* et *CallHome*) et un corpus de conversations préparées (*Broadcast News*).

3.4.3 Classification

Comme vue au paragraphe 3.3.3, la méthode généralement utilisée pour regrouper les segments est la classification hiérarchique ascendante. Cette méthode a été utilisée avec succès dans beaucoup de domaines. Elle a l'avantage d'utiliser un cadre théorique robuste, si les mesures sont des distances.

Cependant, elle présente une faiblesse majeure concernant la fusion des classes. Pendant la construction du dendrogramme, les noeuds contiennent *a priori* les segments d'un même locuteur. Ces noeuds ne sont jamais remis en cause lors la création d'un nouveau noeud (création par fusion de deux noeuds) alors que les dissimilarités ne garantissent pas l'inégalité triangulaire.

Les dissimilarités sont réévaluées après chaque création de noeud. Le coût de calcul des nou-

velles dissimilarités n'est pas négligeable¹¹ (*c.f.* § 3.3.3). Pour restreindre le temps de calcul à un niveau acceptable, les modèles des classes utilisés (les modèles de locuteurs) sont nettement moins performants que les modèles utilisés en RAL. Généralement, seuls les poids des modèles sont adaptés (un coefficient par composante) au lieu d'adapter les moyennes (n coefficients par composante, n étant la taille du vecteur acoustique). Une autre solution employée consiste à réduire fortement le nombre de composantes des modèles¹².

¹¹Les dissimilarités sont évaluées à chaque itération pour palier le problème d'absence de la propriété de l'inégalité triangulaire.

¹²Le CLIPS utilise 32 composantes alors qu'un système classique de RAL utilise de 256 à 2048 composantes.

Chapitre 4

Méthode proposée : détection de ruptures et classification

Cette partie présente une méthode fondée sur un HMM évolutif (E-HMM) où les locuteurs sont détectés et ajoutés un à un. L'originalité principale de la méthode proposée porte sur l'exploitation des informations (locuteurs détectés et la segmentation provisoire) dès qu'elles sont disponibles en intégrant dans un même processus la détection de ruptures et la classification. L'approche proposée est comparée à l'approche classique du chapitre précédent. Puis, le processus de segmentation est décrit ; un exemple de segmentation illustre la méthode. Enfin, les détails de la méthode sont donnés et évalués sur les corpus utilisés lors des évaluations NIST pour la tâche de segmentation en locuteurs.

4.1 HMM évolutif

Pour pallier les pertes d'informations engendrées par la circulation ascendante des informations de la méthode classique, nous proposons une méthode reposant sur un modèle de conversation utilisant un HMM évolutif. Dans cette approche, toutes les informations disponibles sont exploitées à chaque étape et remises en cause à l'étape suivante.

4.1.1 Principe général

La méthode présentée modélise le problème de la segmentation en locuteurs à l'aide d'un modèle de Markov caché (HMM). Les états de ce modèle représentent les locuteurs du document ; les transitions entre ces états modélisent les changements de locuteurs. Les locuteurs, *i.e.* les états du HMM, sont ajoutés un à un à chaque itération du processus (figure 4.1).

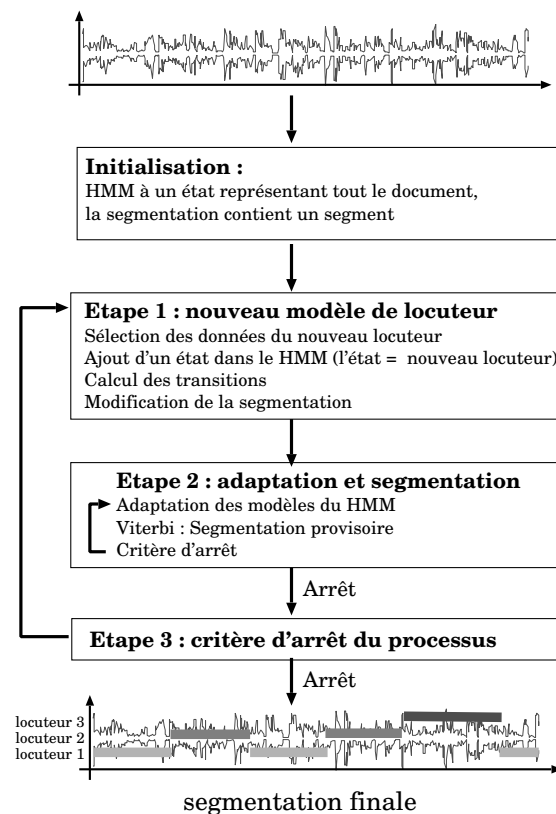


FIG. 4.1 – Méthode proposée.

La méthode se décompose en trois étapes précédées d'une initialisation du processus :

Initialisation :

- L'HMM contient un seul état représentant l'ensemble des locuteurs du document.
- La segmentation contient un seul segment couvrant tout le document.

Etape 1 : un nouveau modèle de locuteur est créé

- Des données sont sélectionnées pour initialiser un nouveau modèle de locuteur.
- La segmentation provisoire et l'HMM sont modifiés pour prendre en considération le nouveau locuteur.

Etape 2 : adaptation des modèles de locuteurs et segmentation intermédiaire

- Une étape itérative adapte les modèles du HMM et propose une segmentation provisoire.
- Cette étape s'arrête quand le maximum de vraisemblance au niveau de la segmentation provisoire est atteint.

Etape 3 : critère d'arrêt du processus

- Le processus s'arrête quand tous les locuteurs ont été détectés, sinon il reprend à l'étape 1.

Dans le modèle de segmentation proposé, la segmentation provisoire est remise en cause à chaque étape du processus. Cette opération s'effectue à deux niveaux :

- Lors de la phase d'adaptation des modèles de locuteurs (étape 2) : un segment précédemment attribué à un locuteur peut être affecté à un autre locuteur ; par ailleurs un segment peut être divisé en plusieurs sous-segments (*c.f.* l'exemple au § 4.5.2).
- A la fin du processus (étape 3) : l'arrêt du processus a pour conséquence de supprimer le dernier locuteur ajouté.

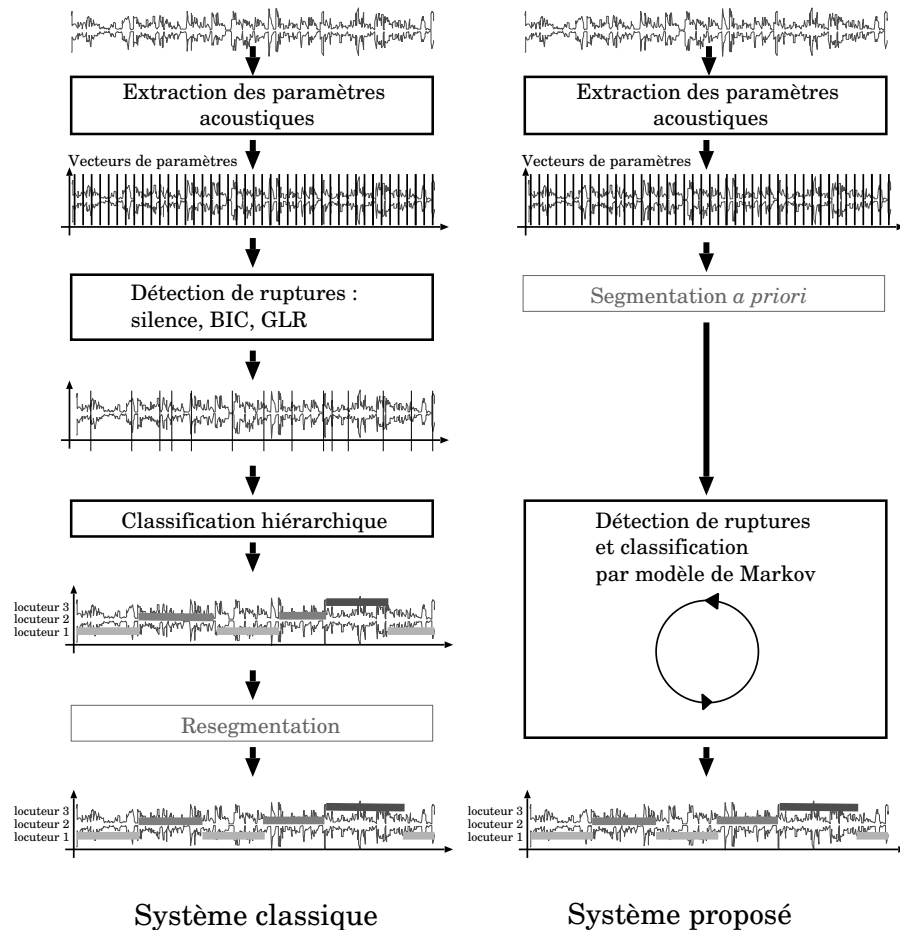
4.1.2 Méthode proposée *v.s.* méthode classique

FIG. 4.2 – Comparaison entre les modules d'un système classique où l'information est propagée de module en module et la méthode proposée où l'information est remise en cause à chaque itération. En grisé : étapes optionnelles.

Le travail présenté dans ce document est destiné à pallier les défauts des méthodes classiques exposées au paragraphe 3.4. La détection des ruptures, la classification des segments et la resegmentation sont faites en simultanées (figure 4.2). **Ainsi les informations sont disponibles dans un même processus, aussi bien pour trouver les ruptures que pour détecter les locuteurs ; nous cherchons à bénéficier à tout moment des informations déjà découvertes en cours de segmentation.**

La méthode proposée s'apparente à une méthode de classification descendante : **la classe initiale contenant tous les segments est coupée en classes de locuteurs lors des itérations successives.** La stratégie de classification retenue est différente de la méthode classique qui utilise une technique ascendante (des classes sont fusionnées jusqu'à obtenir les locuteurs).

4.2 Structure du modèle

Le signal à indexer est constitué d'une séquence d'observations $O = (o_1, o_2, \dots, o_T)$.

La conversation entre les locuteurs est représentée par un modèle de Markov caché ergodique. Chaque état modélise un locuteur et les transitions représentent les changements entre les états (les locuteurs).

Le modèle de Markov caché λ est défini par (E, A, B) :

- $E = \{1, 2, \dots, N\}$ un ensemble d'états. L'état i est associé au modèle X_i du locuteur \mathcal{X}_i .
- $A = \{a_{i,j}\}$ un ensemble de probabilités de transition entre les états.
- $B = \{b_i\}$ un ensemble de probabilités d'émission. A chaque état i est associée une fonction de densité de probabilités d'émission b_i . On note $b_i(o_t)$ la probabilité du modèle X_i pour l'observation o_t .

L'algorithme de Viterbi à partir du modèle de conversation et du document permet d'obtenir la séquence d'états optimale. Chacun des états modélisant un locuteur, la séquence d'états correspond à la segmentation optimale en fonction des informations disponibles (le HMM et les modèles de locuteurs attachés aux états).

4.3 Détails de l'algorithme itératif

La construction du modèle de segmentation est réalisée par un processus itératif, qui détecte et ajoute à chaque itération i un locuteur \mathcal{X}_i . Ce processus est réalisé en trois étapes, précédées d'une initialisation (figures 4.3 et 4.4).

Note : lors de la description du processus, le numéro de l'itération est porté en exposant.

4.3.1 Initialisation

Itération 1 : ajout du locuteur X1

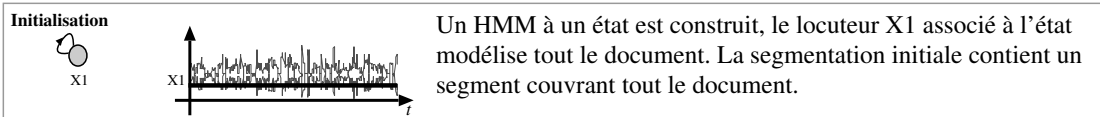


FIG. 4.3 – Exemple de segmentation en locuteurs d'un document sonore : initialisation du modèle de conversation.

L'HMM contient un seul état représentant l'ensemble des locuteurs du document

La première itération ($i = 1$) du processus initialise la segmentation et le HMM. Le modèle de Markov $\lambda^1 = (E^1, A^1, B^1)$ est composé d'un seul état ($E^1 = \{1\}$). Cet état est associé au modèle de locuteur X_1^1 représentant l'ensemble du document. Le modèle de locuteur X_1^1 est appris sur la totalité des observations O . Le locuteur X_1^1 (X_1^i par la suite) représente tous les locuteurs qui n'ont pas encore été détectés.

La segmentation contient un seul segment couvrant tout le document

Une première segmentation triviale $S^1 = (s_1^1, \dots, s_T^1) = (\mathcal{X}_1, \dots, \mathcal{X}_1)$ est générée (figure 4.3). A chaque observation o_i correspond une étiquette $s_i^1 = \mathcal{X}_1$. Toutes les observations appartiennent au locuteur \mathcal{X}_1 .

La segmentation S^1 sera remise en cause à l'itération suivante.

Note : Les différentes étapes, les itérations $i > 1$, sont précisées dans les paragraphes de 4.3.2 à 4.3.4. La figure 4.4 illustre plus particulièrement le fonctionnement du processus pour les itérations $i = 2$ et $i = 3$.

4.3.2 Etape 1 : nouveau modèle de locuteur

Des données sont sélectionnées pour initialiser un nouveau modèle de locuteur

A chaque itération, un nouveau locuteur est initialisé à partir de données extraites du locuteur \mathcal{X}_1 . La stratégie de classification est descendante, la classe du locuteur \mathcal{X}_1 est divisée en deux classes à chaque itération. La méthode de sélection de la séquence d'observations est discutée au paragraphe 5.4.2.

Par conséquent, le modèle X_i^i est construit à partir d'une séquence de longueur fixe d'observations $(o_r, o_{r+1}, \dots, o_{r+t})$ affectée au locuteur \mathcal{X}_1 .

La segmentation provisoire et l'HMM sont modifiés pour prendre en considération le nouveau locuteur

Les probabilités de transition sont calculées suivant les règles définies dans le paragraphe 5.4.2. Le nouveau modèle de Markov $\lambda^i = (E^i, A^i, B^{i-1})$ est ainsi construit.

La segmentation S^i est générée : la séquence $(o_r, o_{r+1}, \dots, o_{r+t})$ est affectée au locuteur \mathcal{X}_i .

$$\begin{cases} s_j^i = s_j^{i-1} \forall j \notin \{r, \dots, r+t\} \\ s_r^i = s_{r+1}^i = \dots = s_{r+t}^i = i \end{cases} \quad (4.1)$$

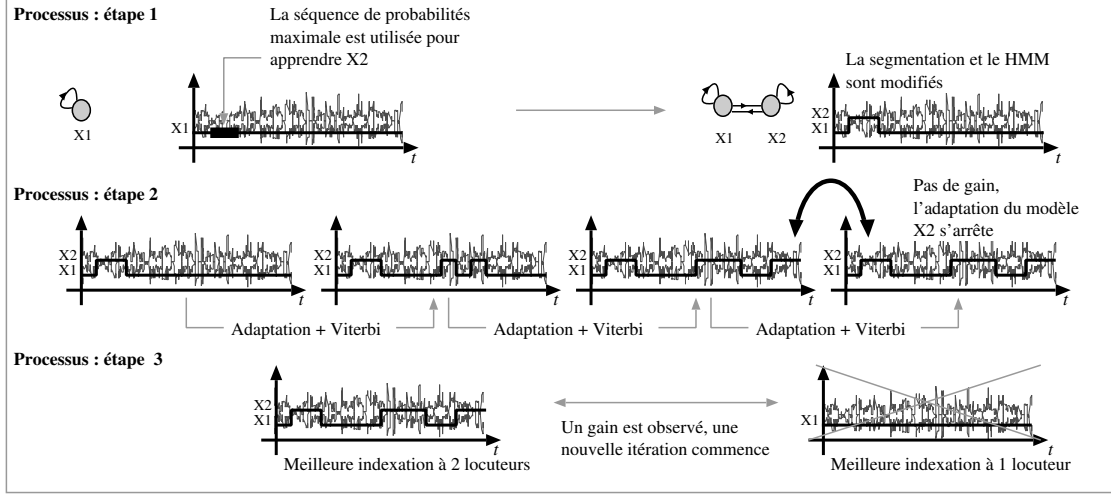
Un nouvel état i est ajouté dans l'ensemble E^{i-1} . L'ensemble des états est alors $E^i = E^{i-1} \cup \{i\}$.

4.3.3 Etape 2 : Adaptation des modèles de locuteurs et segmentation provisoire

Dans cette phase, elle-même itérative, le processus adapte les paramètres du HMM λ^i :

- a : Les modèles de locuteurs sont adaptés aux segments qui leurs sont attribués. Pour chaque $k \in \{1, \dots, i\}$, le modèle de locuteur X_k^i est adapté en fonction des observations qui lui sont affectées dans la segmentation S^i .
- b : L'ensemble des probabilités d'émission B^i est calculé.

Itération 2 : ajout du locuteur X2



Itération 3 : ajout du locuteur X3

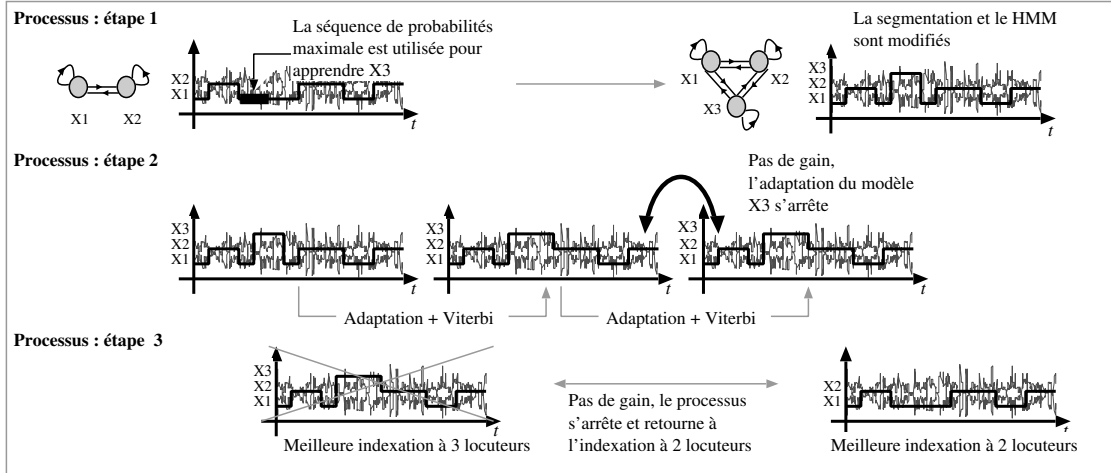


FIG. 4.4 – Exemple de segmentation en locuteurs d'un document sonore : ajout de locuteurs.

c : L'algorithme de Viterbi permet de calculer une nouvelle version de S^i pour obtenir l'alignement optimal par rapport au HMM actuel. La probabilité de la segmentation est :

$$P(O, S^i | A^i, B^i) = \prod_{t=1}^{t=T} b_{s_t^i}^i(o_t) a_{s_{t-1}^i s_t^i}^i \quad (4.2)$$

d : Si un gain est observé entre deux itérations de l'étape 3, le processus reprend en a.

La segmentation à i locuteurs la plus vraisemblable est obtenue en fin d'adaptation. L'adaptation du modèle s'arrête quand aucun gain en probabilité n'est observé sur la séquence d'états : le maximum de vraisemblance du modèle de conversation a été atteint. Ce critère d'arrêt est discuté au paragraphe 5.5.3

4.3.4 Etape 3 : critère d'arrêt

Pour deux itérations consécutives, les segmentations à $i - 1$ et à i locuteurs sont comparées. Si la segmentation à i locuteurs est plus probable que la segmentation à $i - 1$ locuteurs alors le

processus reprend à l'étape 1 pour obtenir une segmentation à $i + 1$ locuteurs. Les critères d'arrêt sont étudiés au paragraphe 5.6.

4.4 Originalités de la méthode proposée

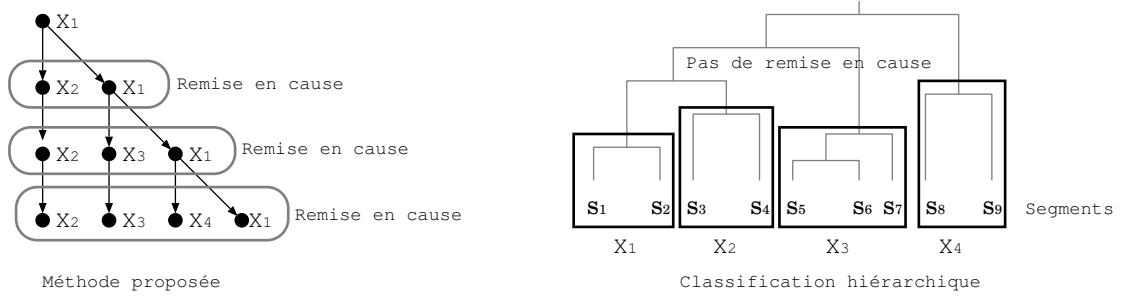


FIG. 4.5 – Classification : différences entre l'approche proposée et de la méthode classique.

L'utilisation d'un modèle de Markov caché pour la segmentation en locuteurs n'est pas en soi une méthode nouvelle. [Wilcox 1994, Gauvain 2001] ont proposé un système où un modèle de conversation reposant sur un HMM est utilisé dans la phase de resegmentation. L'originalité de la méthode porte sur trois points :

1. La méthode de classification descendante divise **la classe initiale** représentant tous les locuteurs en deux sous-classes **à chaque itération** (c.f. § 4.3.3). La figure 4.5 illustre cette différence.
2. Le critère du maximum de vraisemblance est respecté dans les quatre étapes du processus. En particulier, l'apprentissage des modèles, le décodage et critères d'arrêt vérifient ce critère.
3. Les paramètres du modèle de conversation et la segmentation sont remis en cause à chaque itération.

4.5 Exemple

La méthode est illustrée dans les paragraphes suivants au moyen d'un exemple de segmentation réelle d'un document contenant deux locuteurs. Le fichier utilisé provient du corpus d'évaluation décrit au paragraphe 5.1, il obtient un taux d'erreur d'affectation de 6,71% avec la méthode d'évaluation NIST 2000 (c.f. § 13.5.1).

4.5.1 Itération 1 : construction du premier locuteur

Le processus construit un modèle de Markov caché contenant un seul état associé au locuteur libellé X_1 (figure 4.6. Les zones grisées correspondent aux segments.). La segmentation est composée d'un seul segment couvrant la totalité du document. L'étiquette de l'état 1 est associée à ce segment. Dans cette étape d'initialisation, tous les locuteurs intervenant dans le document sont modélisés par l'état X_1 . A partir des données attribuées à X_1 (tout le document ici), le modèle X_1 est construit. Enfin, les probabilités de transition et d'émission du HMM sont calculées ainsi que la probabilité du chemin.

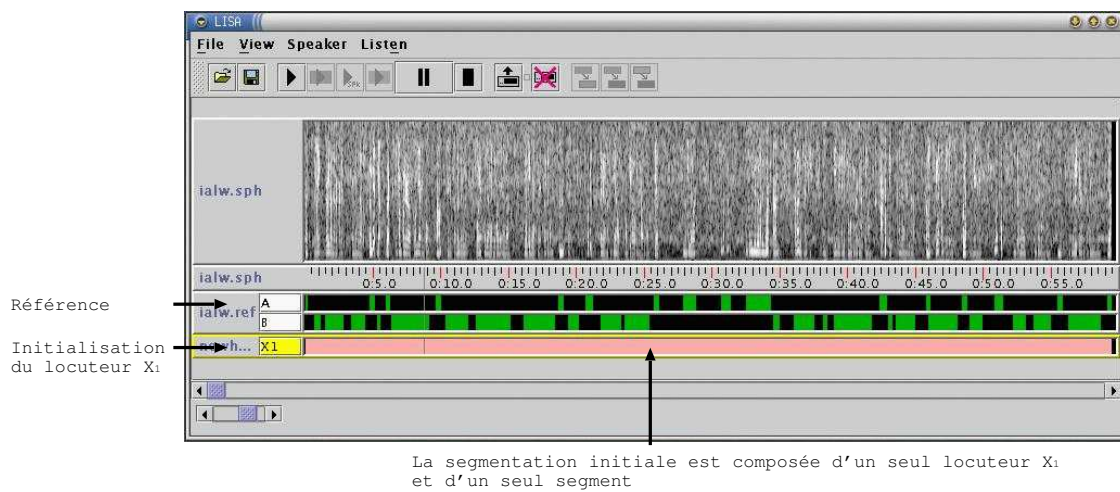


FIG. 4.6 – Exemple de segmentation en locuteurs d'un document sonore : itération 1, initialisation. Les zones grisées correspondent aux segments.

4.5.2 Itération 2 : ajout du deuxième locuteur

L'objectif est d'extraire des données attribuées à X_1 pour construire un nouveau locuteur. Un sous-ensemble de blocs consécutifs est sélectionné, à partir duquel le modèle du deuxième locuteur X_2 est construit. Les blocs sont choisis parmi les blocs X_1 .

Le système ajoute un état 2, qui représente le nouveau locuteur. A partir des blocs choisis, un nouveau segment est construit et attribué à X_2 . Le libellé de ce segment est modifié pour refléter son appartenance au locuteur X_2 (figure 4.7).

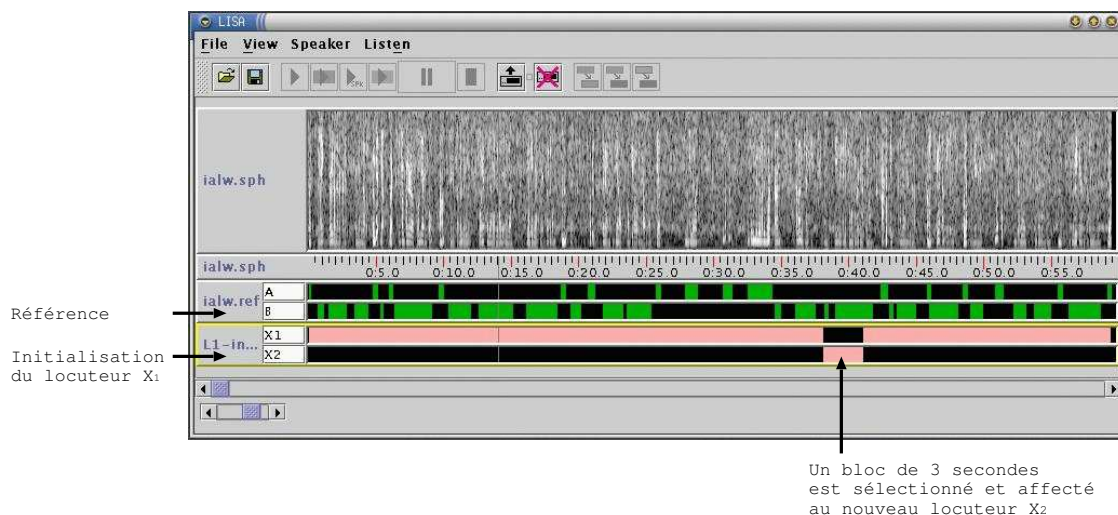


FIG. 4.7 – Exemple de segmentation en locuteurs d'un document sonore : initialisation du locuteur X_2 . Les zones grisées correspondent aux segments.

Les paramètres du HMM sont adaptés en fonction de la nouvelle segmentation, *i.e.* les modèles X_1 et X_2 sont adaptés en fonction de leurs données respectives. Les probabilités de transition et d'émission sont calculées. Puis l'algorithme de Viterbi propose la segmentation la plus vraisemblable par rapport au HMM. L'algorithme fournit également la probabilité du chemin de décodage (*i.e.* la probabilité de la segmentation).

L'adaptation des modèles de locuteurs et le décodage de Viterbi sont appliqués tant que la vraisemblance de la solution (obtenue par Viterbi) augmente entre deux itérations. Dans l'exemple présenté, cinq itérations sont nécessaires avant d'atteindre le maximum de vraisemblance. Aussi bien le schéma d'adaptation que le décodage par Viterbi garantissent une convergence au sens du maximum de vraisemblance. Le processus itératif s'arrête lorsque le maximum de vraisemblance du chemin de décodage est atteint (figure 4.8).

Les modèles de locuteurs X_1 et X_2 sont adaptés à chaque itération de l'étape 3. Cet apprentissage implique de calculer de nouvelles probabilités d'émission.

Enfin, le critère d'arrêt détermine si le nombre de locuteurs détectés est optimal. Dans l'exemple, la segmentation à deux locuteurs est plus probable que la segmentation à un locuteur, le processus reprend à l'étape 1 pour ajouter un nouveau locuteur.

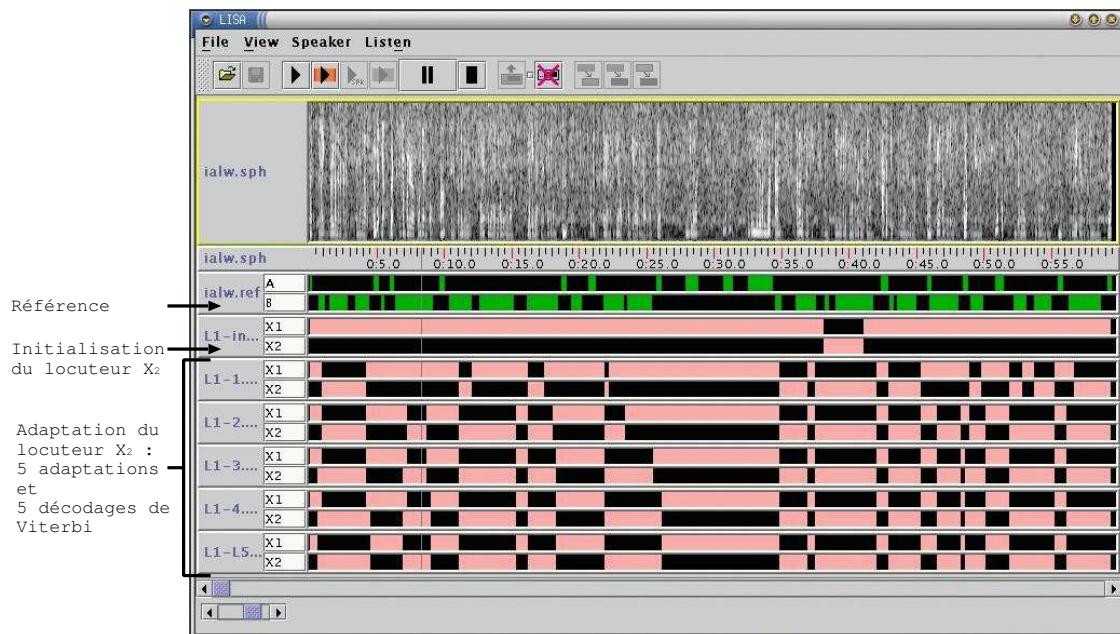


FIG. 4.8 – Exemple de segmentation en locuteurs d'un document sonore : adaptation du locuteur X_2 . Les zones grisées correspondent aux segments.

4.5.3 Itération 3 : ajout du troisième locuteur

L'itération 3 se déroule comme l'itération 2. Le modèle du locuteur \mathcal{X}_3 est construit à partir d'un sous-ensemble de blocs consécutifs libellé \mathcal{X}_1 .

L'état représentant le locuteur \mathcal{X}_3 est ajouté au HMM. La segmentation est modifiée pour prendre en compte le sous-ensemble de blocs consécutifs (figure 4.9). Les transitions du HMM sont adaptées pour intégrer le nouvel état.

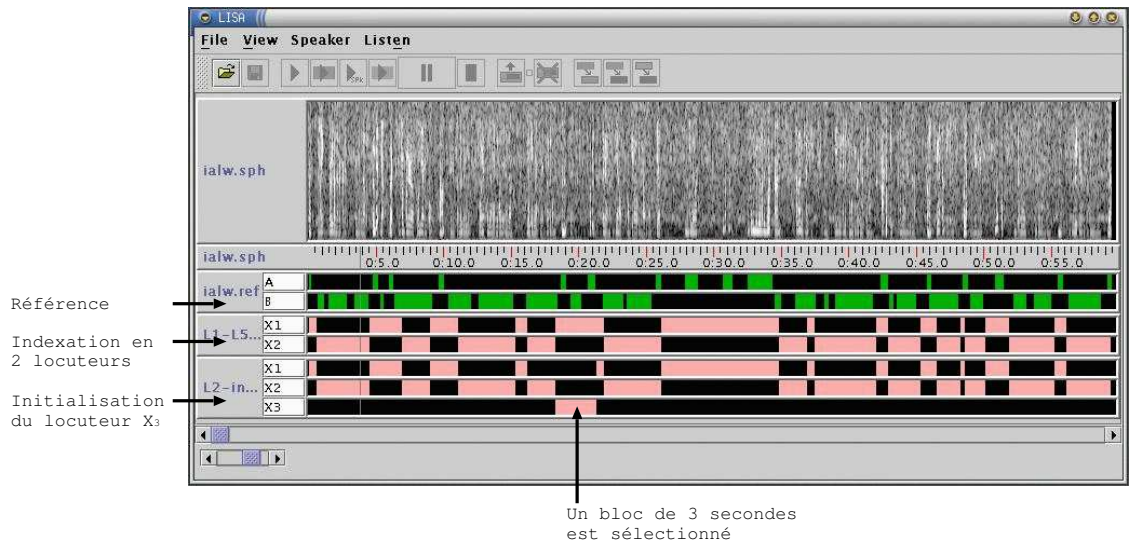


FIG. 4.9 – Exemple de segmentation en locuteurs d'un document sonore : initialisation du locuteur \mathcal{X}_3 . Les zones grisées correspondent aux segments.

L'adaptation des trois modèles de locuteurs et le décodage par Viterbi sont effectués itérativement (figure 4.10).

Le critère d'arrêt pour l'ajout de locuteur est étudié. Pour le locuteur \mathcal{X}_3 , aucun gain n'est observé. Le processus d'ajout s'arrête. La segmentation en deux locuteurs représente la segmentation finale (figure 4.11).

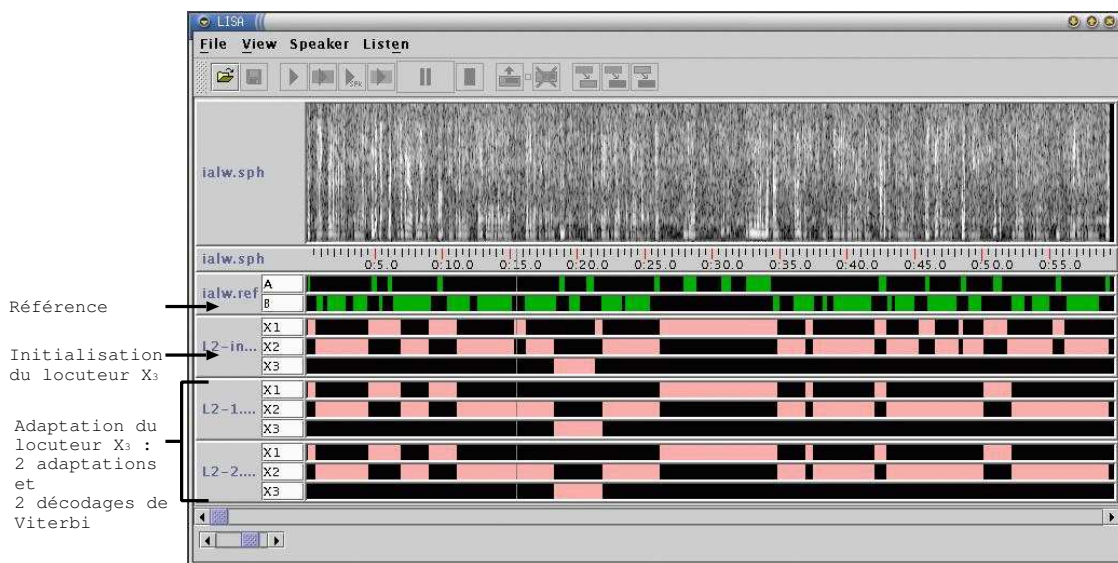
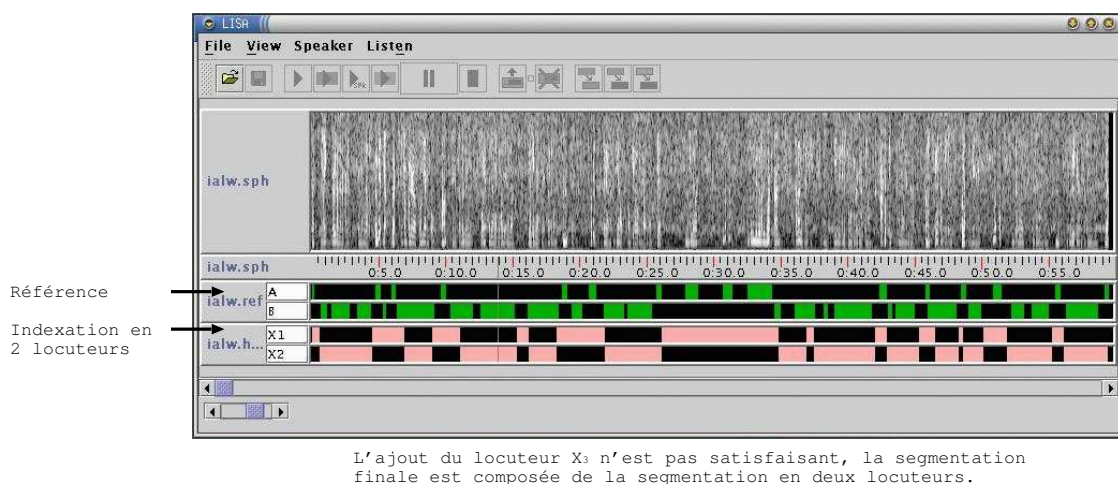


FIG. 4.10 – Exemple de segmentation en locuteurs d'un document sonore : adaptation du locuteur \mathcal{X}_3 . Les zones grisées correspondent aux segments.



L'ajout du locuteur X_3 n'est pas satisfaisant, la segmentation finale est composée de la segmentation en deux locuteurs.

FIG. 4.11 – Exemple de segmentation en locuteurs d'un document sonore : la segmentation en deux locuteurs est conservée, le processus s'arrête. Les zones grisées correspondent aux segments.

4.6 Discussion

4.6.1 Hiérarchisation des modèles

La méthode de sélection d'un nouveau locuteur suppose que les segments libellés \mathcal{X}_1 contiennent les locuteurs non détectés. A la première itération, cette hypothèse est vérifiée : X_1 modélise la totalité de l'enregistrement. Aux itérations suivantes, cette hypothèse reste valide si les locuteurs \mathcal{X}_i ($i \neq 1$) n'ont pas attiré les segments de locuteurs qui n'ont pas encore été détectés.

4.6.2 Les données servant à l'initialisation d'un locuteur

Un nouveau locuteur est initialisé à partir d'un sous-segment extrait des données libellées \mathcal{X}_1 . Pour qu'un locuteur ne capte pas les données d'un autre locuteur (voire de plusieurs), la méthode devrait garantir que le sous-segment ne contient que les données d'un seul locuteur.

4.6.3 Maximum de vraisemblance

Dans la classification descendante, l'ajout d'une classe entraîne une augmentation de la vraisemblance pour les modèles associés à la classe. "Splitting a node will always result in an increase in likelihood" écrit Sue Johnson dans [Johnson 1998]. L'ajout d'un locuteur dans l'HMM entraîne, après adaptation des modèles de locuteurs, une augmentation de la vraisemblance des émissions. Le modèle de conversation avec les probabilités d'émission les plus élevées est le HMM contenant autant d'états que d'observations (figure 4.12).

Bien que l'évolution des probabilités de transition compensent en partie ce problème (les probabilités de transition diminuent quand le nombre d'états augmente), il est donc probable que le critère d'arrêt de la segmentation ne pourra pas se fonder uniquement sur la probabilité de la segmentation, bien que la probabilité de transition diminue quand le nombre d'états augmente.

4.6.4 Méthode ascendante *v.s.* méthode descendante

La méthode proposée est une stratégie de classification descendante : à chaque itération un nouvel état est inséré dans l'HMM. La méthode peut être adaptée à une stratégie ascendante : le modèle initial est composé d'un état par observation (figure 4.12). Puis dans la phase itérative d'adaptation et de décodage, les observations sont affectées à d'autres états jusqu'à atteindre le maximum de vraisemblance. Les états associés à aucune observation sont supprimés. Cependant, cette stratégie est difficile à mettre en œuvre (d'après le paragraphe 4.6.3), la somme des probabilités d'émission est maximum pour un modèle à T états !

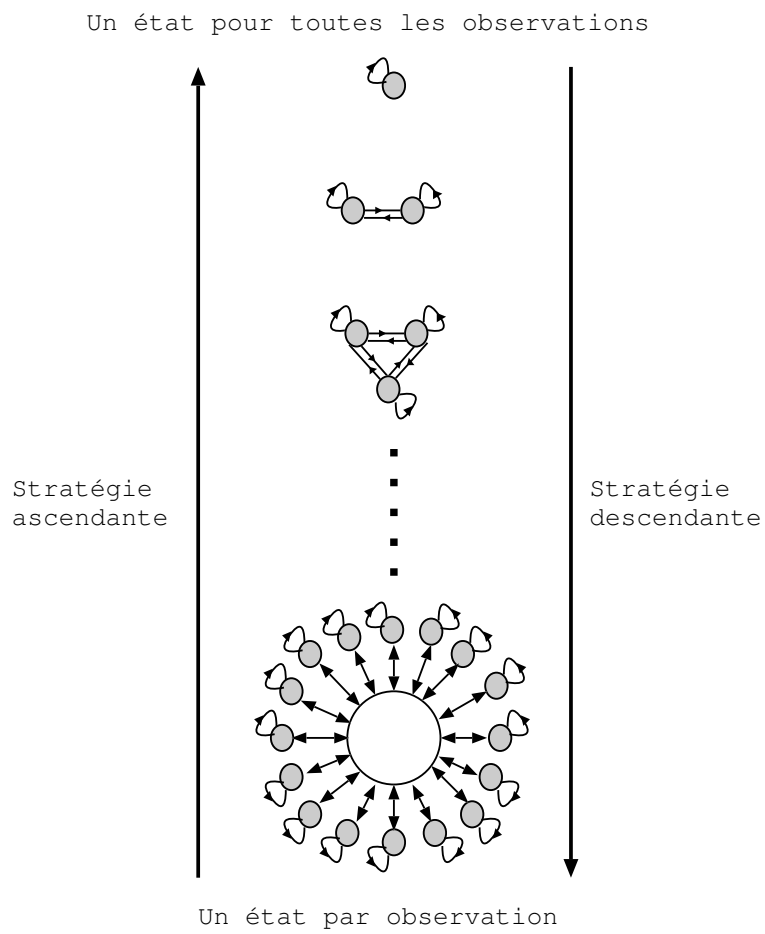


FIG. 4.12 – HMM évolutif utilisant une stratégie ascendante ou descendante pour la sélection des locuteurs.

Chapitre 5

Méthode proposée : implémentation et évaluation

Dans ce chapitre, des points particuliers de la méthode proposée sont discutés et étayés par des expériences menées sur un corpus de développement issu des évaluations de reconnaissance du locuteur NIST 2000 pour la tâche de segmentation. L'influence des paramètres acoustiques est mesurée et commentée. Différentes méthodes d'apprentissage sont évaluées. Les critères d'arrêt de l'adaptation des modèles et de l'ajout de locuteur sont étudiés. Comme dans la méthode classique, la détection du nombre de locuteurs reste l'enjeu majeur de la segmentation en locuteurs. Un autre point important dans la méthode proposée est la sélection du segment initial pour un nouveau locuteur.

Dans le chapitre précédent, les grandes lignes de la méthode de segmentation ont été exposées. Dans ce chapitre, les trois étapes constituant la méthode sont détaillées et évaluées à partir d'un corpus de développement. En particulier, quatre points ont retenu notre attention (figure 5.1) :

- Les traitements à effectuer avant de segmenter un document : la paramétrisation et la pré-segmentation.
- La détection et l'initialisation d'un nouveau locuteur (étape 1 dans la figure).
- Le modèle de conversation reposant sur un HMM : les modèles de locuteurs, les probabilités d'émission et de transition et l'arrêt de l'apprentissage/décodage d'un HMM (étape 2 dans la figure).
- La détection du nombre de locuteurs (étape 3 dans la figure).

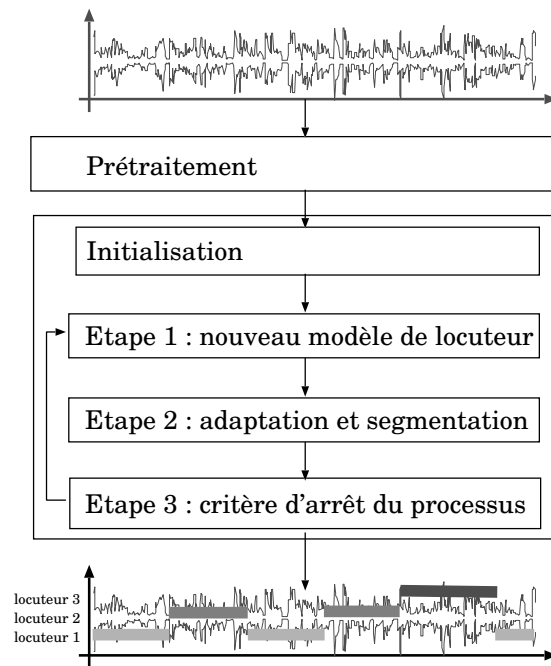


FIG. 5.1 – Méthode d'évaluation : rappel des étapes de la méthode de segmentation proposée.

Avant d'exposer ces quatre points, la méthode d'évaluation est précisée et des informations sont données sur les outils issus de la vérification du locuteurs.

5.1 Méthode d'évaluation

5.1.1 Choix liés à l'évaluation

La tâche

Les trois premiers objectifs (les prétraitements, la détection d'un locuteur et le modèle de conversation) sont évalués pour un nombre de locuteurs fixé *a priori* à deux, ainsi la méthode s'affranchit, dans un premier temps, du problème de la détection du nombre de locuteurs.

Le dernier objectif, la détection du nombre de locuteurs, est discuté pour des documents à deux locuteurs et à n locuteurs au paragraphe 5.6.

Le corpus

Le corpus de développement contient 100 enregistrements *Switchboard II phase II* et 62 enregistrements *CallHome* de langue anglaise [LDC 1998]. Ces enregistrements sont extraits du corpus de segmentation de la campagne d'évaluations NIST 2000 et 2001. Les enregistrements sont des conversations téléphoniques spontanées, majoritairement à deux locuteurs (154 enregistrements contiennent deux locuteurs, 8 enregistrements contiennent 3 locuteurs).

La méthode d'évaluation

La méthode d'évaluation des segmentations correspond à une erreur d'affectation entre la segmentation hypothèse (le résultat donné par le système) et la segmentation de référence fournie par NIST. Cette méthode, présentée au paragraphe 13.5.1, a été développée par NIST pour les évaluations NIST 2000 et 2001.

Le système de référence

Le tableau 5.1 définit notre système de référence qui obtient un taux d'erreur d'affectation de 10,79% (10,79% des trames ne sont pas affectées au bon locuteur). Dans la suite du chapitre, tous les paramètres dont les valeurs ne sont pas précisées reprennent les valeurs correspondantes du tableau.

Paramétrisation	LFCC CMS Limitation de bande (300Hz-3400Hz) Taille de la fenêtre pour le calcul d'un coefficient Pas de décalage la fenêtre Nombre de trames par bloc pour la pré-segmentation	20 + Energie non appliqué non appliqué 20ms 10ms 30
Locuteur	Type initialisation d'un nouveau modèle Nombre de blocs pour initialiser un modèle Initialisation d'un nouveau locuteur à partir de	sur un ensemble de blocs 10 \mathcal{X}_1
Nombre de locuteurs	Minimum Maximum	1 2 (ou 10)
Apprentissage des locuteurs	Type MAP coefficient du monde MAP coefficient du locuteur Initialisation par Initialisation à chaque apprentissage	adaptation MAP 0,4 0,6 le modèle du monde oui
Transitions	De l'état i vers l'état i ($a_{i,i}$) De l'état i vers l'état j ($a_{i,j}$)	0,6 0,4 $\frac{(NB\ de\ locuteurs) - 1}{calculées\ à\ chaque\ ajout\ d'état}$
Arrêt de la convergence	Ajout locuteur Adaptation locuteurs	probabilité du chemin + nb de segments probabilité du chemin

TAB. 5.1 – Les paramètres du système de référence.

5.2 Outils issus de la reconnaissance du locuteur

5.2.1 SPRO

La paramétrisation acoustique est calculée au moyen du module SPRO développé par Guillaume Gravier (IRISA). Ces outils sont inclus dans la plate-forme commune de développement du consortium ELISA¹ [Magrin-Chagnolleau 2001, Gravier 1999].

5.2.2 Plate-forme AMIRAL

Les modèles de locuteurs employés et les probabilités d'émission sont calculés par le système de reconnaissance du locuteur AMIRAL, développé au LIA [Bonastre 2000, Fredouille 2000a, Fredouille 2000b, Besacier 2000]. Le système AMIRAL est évalué dans le cadre des évaluations NIST depuis 1998. L'annexe D présente les résultats du LIA et les techniques employées pour les différentes tâches des évaluations en 2002 ; tandis que l'annexe A montre les performances du LIA en VAL pour les évaluations NIST de 2000 à 2002.

5.3 Les traitements effectués avant la segmentation

Deux points sont abordés dans ce paragraphe (figure 5.2) :

- la paramétrisation
- et la segmentation initiale de l'enregistrement.

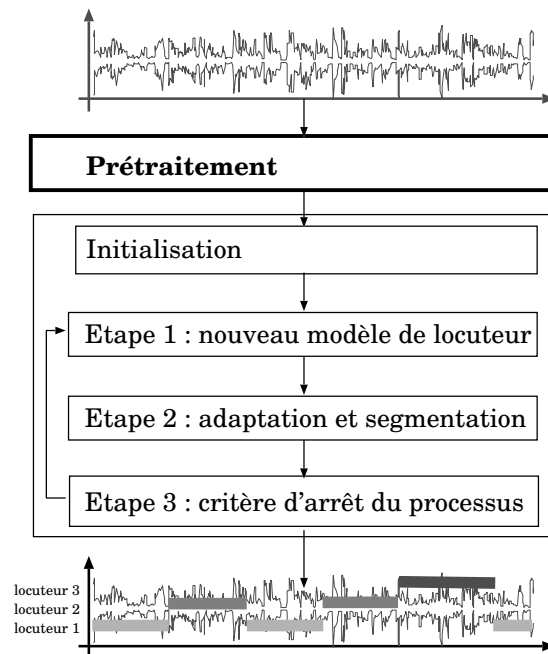


FIG. 5.2 – *Evaluation : prétraitement.*

¹Les objectifs du consortium ELISA sont donnés dans l'annexe B, ainsi qu'un descriptif de la plate-forme.

5.3.1 Paramétrisation

La paramétrisation utilisée dans le système de segmentation est issue des études faites en vérification du locuteur. Elle repose sur des coefficients *LFCC*, calculés toutes les 10ms sur une fenêtre de 20ms.

La paramétrisation employée par le LIA pour la majorité des tâches de vérification contient 16 LFCC, complétés par les dérivées premières (Delta). Les coefficients sont normalisés par le retrait de la moyenne cepstrale (CMS) et les trames de faible énergie sont supprimées.

Dans le cadre de ce travail, contrairement à ce que nous venons de voir en VAL, aucune compensation de canal (CMS) n'est appliquée sur les vecteurs acoustiques. En effet, les expériences préliminaires ont montré que la CMS dégradait les performances, les différences entre les canaux d'enregistrement aidant à la segmentation. De même, suite aux mêmes expériences, la suppression de trames de faible énergie n'a pas été appliquée.

Expériences

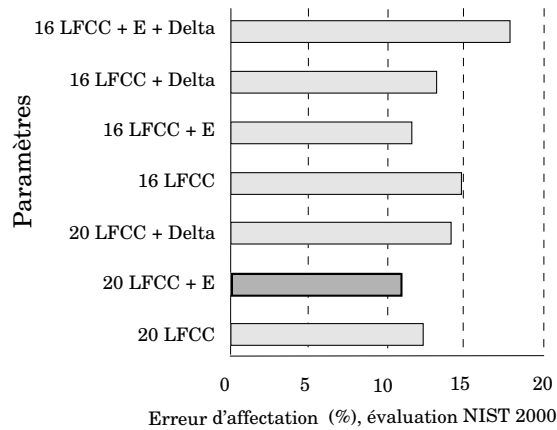


FIG. 5.3 – Résultats de segmentation en locuteurs pour différentes paramétrisations : les résultats sont obtenus à partir du corpus de développement, le taux d'erreur d'affectation est calculé à partir de la méthode NIST 2000/2001, les résultats sont donnés pour 16 ou 20 LFCC avec ou sans l'énergie (E) et les dérivées premières (Delta).

L'histogramme 5.3 montre les taux d'erreur d'affectations obtenus pour différentes paramétrisations à 16 et 20 coefficients. Pour la tâche de segmentation en locuteurs, le taux d'erreur le plus faible (10,79%) a été obtenu avec 20 LFCC et l'énergie du signal (E). L'ajout des dérivées premières (delta) améliore les résultats par rapport aux coefficients LFCC seuls, mais les delta sont moins performants que l'énergie.

Commentaires

Ces résultats expérimentaux sont en accord avec les résultats obtenus par le MIT et OGI [Reynolds 2000, Adami 2002]. D. Reynolds utilise 24 MFCC : il indique que les dérivées dégradent les performances. A. Adami utilise 24 *Line Spectral Pairs* (LSP [Itakura 1975]) : il indique qu'il observe un gain relatif de 20% par rapport aux MFCC.

5.3.2 Segmentation initiale

Dans la méthode classique, la détection des ruptures construit des segments de grande taille qui sont ensuite découpés en sous-segments de taille fixe :

- Douglas Reynolds dans [Reynolds 2000] applique cette stratégie avec la détection des silences ;
- Mouhamadou Seck dans [Seck 2001] l'utilise avec la détection des changements acoustiques ;
- Jean-Luc Gauvain dans [Gauvain 2001] l'applique avec la détection de ruptures par identification des zones.

Nous proposons de découper directement le signal en segments de petite taille sans chercher à détecter les ruptures. Bien entendu, cette méthode est applicable uniquement sur les signaux contenant essentiellement de la parole. Le temps de calcul d'une segmentation *a priori* est négligeable, les marqueurs de ruptures sont posés à des intervalles réguliers. Par contre, les segments doivent être de taille plus petite pour minimiser le nombre de segments contenant plus d'un locuteur. Dans ce travail, le signal est découpé en segments élémentaires, appelés blocs, d'une durée de 0,3 seconde.

Cette pré-segmentation du signal réduit les temps de calcul du décodage² et elle a une influence sur les transitions (*c.f.* § 5.4.2) et sur les probabilités d'émission (*c.f.* § 5.5.2).

5.4 Ajout de locuteur et initialisation d'un locuteur

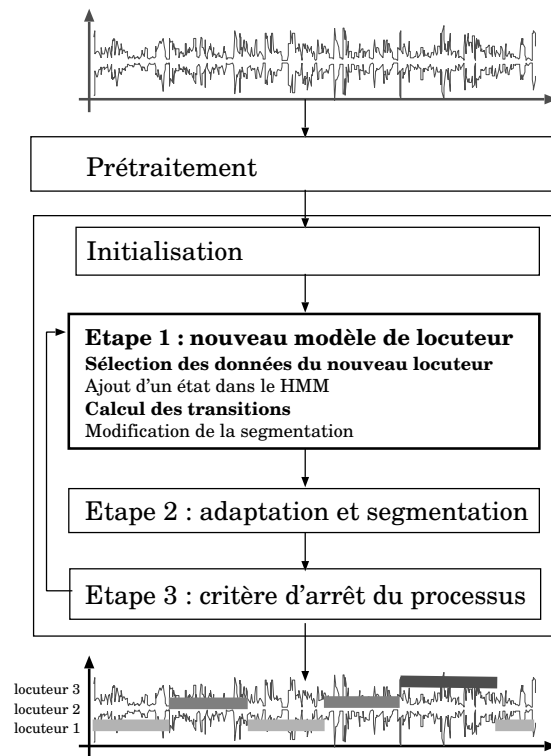


FIG. 5.4 – Evaluation, ajout de locuteur et initialisation d'un locuteur : étape 1.

La figure 5.4 rappelle que l'ajout d'un nouveau locuteur se situe dans l'étape 1. La sélection

²Qui sont déjà faibles en comparaison des décodages mis en oeuvre dans les systèmes de reconnaissance de la parole.

Pourcentage de silence (en moyenne)	10%
Pourcentage de parole (en moyenne)	90%
Pourcentage du locuteur majoritaire (en moyenne, parmi les trames de parole)	96%
Pourcentage des autres locuteurs (en moyenne, parmi les trames de parole)	4%

TAB. 5.2 – Statistique sur la séquence permettant d'initialiser un nouveau locuteur indiquant le pourcentage de silence et parole dans la séquence et le pourcentage du locuteur majoritaire et des autres locuteurs (non majoritaires) parmi les trames de parole.

de données pour l'ajout d'un nouveau locuteur et le calcul des transitions du HMM sont détaillés dans les paragraphes suivants.

5.4.1 Sélection de données

Lors de l'ajout d'un locuteur dans le modèle de conversation, un ensemble de blocs consécutifs est sélectionné parmi les segments libellés \mathcal{X}_1 ⁽³⁾. L'ensemble de blocs utilisé pour initialiser le nouveau locuteur doit être suffisamment grand pour que le modèle soit robuste, mais il ne doit contenir qu'un seul locuteur.

A l'itération i , la séquence d'observations pour initialiser le locuteur \mathcal{X}_i est sélectionnée telle que :

$$\begin{cases} r = \text{ArgMax}_{j \in L} \left(\prod_{k=j}^{j+t} b_1^i(o_k) \right) \\ L = \{j | s_j^{i-1} = s_{j+1}^{i-1} = \dots = s_{j+t}^{i-1} = \mathcal{X}_1\} \end{cases} \quad (5.1)$$

ou $r \in L$ est le rang de la première observation de la séquence $(o_r, o_{r+1}, \dots, o_{r+t})$, qui maximise le produit des probabilités d'émission $\{b_1^i(o_r), b_1^i(o_{r+1}), \dots, b_1^i(o_{r+t})\}$, calculé à partir du modèle de locuteur X_1^i .

Le critère du maximum de vraisemblance garantit de sélectionner les données les plus proches du modèle \mathcal{X}_1 . Dans les expériences, une longueur de 3 secondes a été utilisée.

Commentaires

Nous avons remarqué que les données d'initialisation d'un locuteur devaient contenir un seul locuteur. Dans la méthode présentée, cette condition n'est pas garantie. Mais l'analyse du contenu des séquences d'initialisation montre que 96% des données de parole appartiennent à un seul locuteur (*c.f.* tableau 5.2). En moyenne, 10% de la séquence correspond à du silence.

Autres méthodes

D'autres méthodes ont été testées sans apporter un gain de performance :

- L'ensemble de blocs est sélectionné tel que la vraisemblance soit minimum au lieu de choisir le bloc de vraisemblance maximum. Le nouveau locuteur est initialisé sur les données les plus éloignées du modèle \mathcal{X}_1 .
- Une initialisation affectant un bloc sur deux de \mathcal{X}_1 au nouveau locuteur a été testée. Cette stratégie fonctionne bien pour une segmentation en deux locuteurs. Cependant, elle est difficile à mettre en oeuvre pour plus de deux locuteurs.

³Le modèle X_1 contient l'ensemble des locuteurs non détectés.

5.4.2 Transition entre les états

Après la sélection des données du nouveau locuteur, le HMM est modifié. Un nouvel état représentant le locuteur est ajouté, les probabilités de transitions sont ajustées pour prendre en considération le nouvel état.

Durée des émissions

La durée des émissions suit une loi exponentielle décroissante qui est contrôlée uniquement par la valeur des probabilités de transitions. L'équation 5.2 donne la probabilité p pour que le processus soit resté pendant t observations dans l'état i avant de transiter vers un autre état. La probabilité de rester dans l'état i est donnée par $a_{i,i}$ et la probabilité de changer d'état est de $1 - a_{i,i}$ (figure 5.5).

$$p(t) = a_{i,i}^t \times (1 - a_{i,i}) \quad (5.2)$$

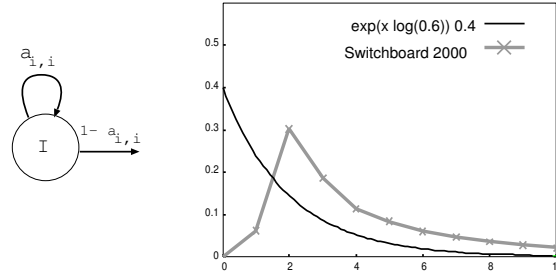


FIG. 5.5 – Durée d'émission dans un HMM : donnée pour la loi exponentielle avec une probabilité $a_{i,i} = 0,6$ et donnée pour les valeurs observées sur le corpus NIST 2000/2001 de segmentation en deux locuteurs (SwitchBoard 2000).

Cette loi ne correspond pas aux observations de la durée des segments faites sur les cinq corpus de parole spontanée et préparée (*c.f.* exemple dans la figure 5.5).

Calcul des transitions

Sans connaissance *a priori* sur la conversation, toutes les transitions entre les locuteurs sont supposées possibles; tous les états du HMM sont donc interconnectés. Aucune information sur les locuteurs n'étant disponible, aucun locuteur n'est supposé prépondérant sur les autres; les locuteurs ont des probabilités de transition identiques.

Les probabilités de transition sont établies en fonction d'un ensemble de règles en accord avec les hypothèses formulées. Elles vérifient trois conditions :

$$\begin{cases} \forall i, a_{i,i} = \gamma \\ \forall (i,j), i \neq j, a_{i,j} = \frac{1-\gamma}{N-1} \\ 0 < \gamma < 1 \end{cases} \quad (5.3)$$

Le poids γ permet de fixer la probabilité de bouclage sur un même état. Les probabilités de transition vers les autres états sont équiprobables et déduites de la probabilité de bouclage. Des exemples de probabilités de transition sont proposés dans les tableaux 5.3 et 5.4 respectivement pour un HMM à 2 et 3 états avec $\gamma = 0,6$.

	locuteur \mathcal{X}_1	locuteur \mathcal{X}_2
locuteur \mathcal{X}_1	0,6	0,4
locuteur \mathcal{X}_2	0,4	0,6

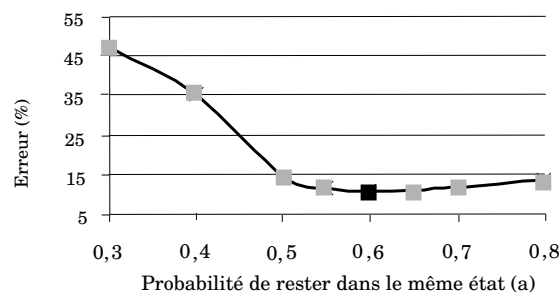
TAB. 5.3 – Probabilités de transition pour un modèle à deux états.

	locuteur \mathcal{X}_1	locuteur \mathcal{X}_2	locuteur \mathcal{X}_3
locuteur \mathcal{X}_1	0,6	0,2	0,2
locuteur \mathcal{X}_2	0,2	0,6	0,2
locuteur \mathcal{X}_3	0,2	0,2	0,6

TAB. 5.4 – Probabilités de transition pour un modèle à trois états.

Résultats expérimentaux

La courbe 5.6 montre le taux d'erreur d'affectation pour différentes valeurs de γ . Le taux d'erreur le plus faible (10,79%) est obtenu pour une valeur de γ égale à 0,6.

FIG. 5.6 – Taux d'erreur d'affectation observé pour différentes valeurs du paramètre γ régulant les probabilités de transition.

Commentaires

L'utilisation d'un HMM pour modéliser la conversation entre les locuteurs apporte une réduction du taux d'erreur de 3,45 en valeur absolue (24% de moins en valeur relative) pour $\gamma = 0,6$ par rapport à un ensemble de probabilités équiprobables ($\gamma = 0,5$). Ce gain est observé bien que la loi implicite sur la durée des émissions ne corresponde pas à la loi mesurée sur les corpus.

Aucun modèle de durée explicite n'a été introduit. De même, aucun état supplémentaire, comme les modèles gauche/droite utilisés en reconnaissance de la parole, n'a été intégré pour ajuster le modèle de durée. Des expériences sur d'autres valeurs de taille des blocs n'ont pas apporté de gain de performance.

5.5 Modèle de conversation

Les spécificités des modèles de locuteurs et des probabilité d'émission sont discutés dans cette section, ainsi que le critère d'arrêt de la phase itérative d'apprentissage/décodage du HMM (figure 5.7).

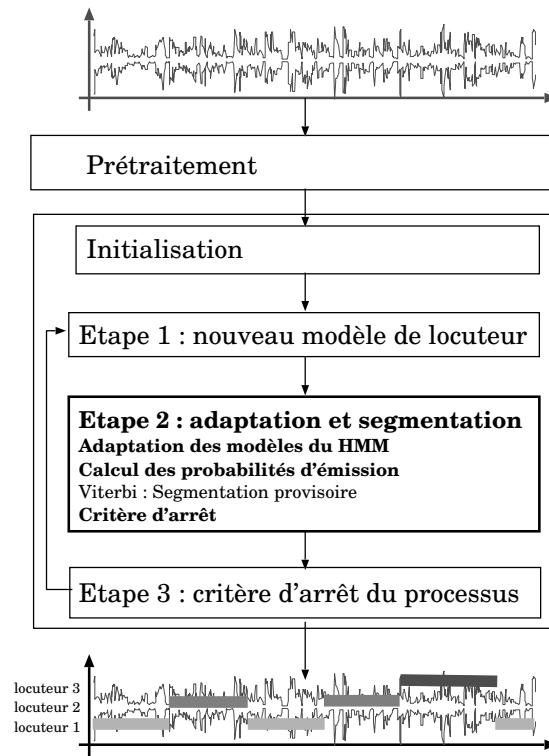


FIG. 5.7 – *Evaluation, adaptation et segmentation : étape 2, adaptation des modèles de locuteurs, probabilités d'émission, critère d'arrêt de la phase itérative d'apprentissage/décodage du modèle de conversation.*

5.5.1 Modèle de locuteur

Les modèles de locuteurs, leur apprentissage et le calcul des probabilités d'émission, reposent sur les méthodes employées par le LIA en reconnaissance du locuteur. Nous supposons que les méthodes et les paramètres employés en reconnaissance du locuteur sont adaptés à la tâche de segmentation.

Un locuteur est modélisé par un GMM à 128 composantes à matrices de covariance diagonales [Reynolds 1995]. Le nombre de composantes n'a pas fait l'objet d'une étude spécifique. Les coûts de calcul des modèles de locuteurs sont acceptables avec des modèles à 128 composantes et les résultats obtenus par le LIA lors des évaluations NIST 2001 pour la tâche de vérification du locuteur valident ce choix.

Apprentissage par adaptation MAP

Les modèles GMM sont adaptés à partir d'un modèle du monde, appris sur un corpus séparé, en utilisant l'algorithme EM-ML (optimisant le critère du maximum de vraisemblance, [Dempster 1977]). Le corpus d'apprentissage du modèle du monde contient 472 enregistrements prononcés par 100 locuteurs (hommes et femmes). Les données sont issues de la campagne d'évaluations en reconnaissance du locuteur NIST 99.

Le modèle de locuteur X_i est adapté dans un premier temps sur une séquence d'observations d'une durée fixe de 3 secondes sélectionnée lors de l'étape 1 (*c.f.* § 5.4.1). Puis après chaque application de l'algorithme de Viterbi, le modèle X_i est adapté à partir des segments étiquetés par X_i (*c.f.* § 4.3.3). La procédure d'adaptation est fondée sur la méthode du *maximum a posteriori*

(MAP, [Gauvain 1994]).

Dans le cadre de ce système, seules les moyennes du modèle sont adaptées ; le modèle utilise le vecteur de poids et les matrices de covariance du modèle du monde. Pour chaque gaussienne $g_k()$ de dimension d , la moyenne $\mu_{k,j}$ du modèle X_i à n composantes est une combinaison linéaire de la moyenne estimée $\hat{\mu}_{k,j}$ et de la moyenne correspondante $\mu_{k,j}^W$ du modèle du monde W [Meignier 2001].

$$\begin{cases} \mu_{k,j} = \alpha \mu_{k,j}^W + (1 - \alpha) \hat{\mu}_{k,j} \\ \text{avec } \alpha \in]0, 1] \\ j \in \{1, \dots, d\} \text{ et } k \in \{1, \dots, n\} \end{cases} \quad (5.4)$$

Expériences

Les résultats des expériences sur le paramètre α de l'adaptation MAP montrent que la valeur optimale se situe aux environs de 0,6. La figure 5.8 reporte les taux d'erreur obtenus avec un α variant de 0,1 à 0,9 avec un pas de 0,1. La valeur du poids α est comparable à la valeur optimale

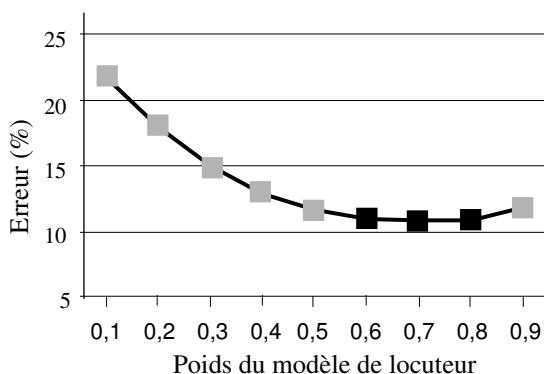


FIG. 5.8 – Différentes valeurs du paramètre alpha pour l'adaptation MAP : le taux d'erreur est obtenu sur le corpus de développement (méthode NIST 2000).

observée en reconnaissance du locuteur lors des évaluations NIST.

5.5.2 Probabilités d'émission

Le calcul des probabilités d'émission s'appuie sur les méthodes employées en RAL. La mesure entre un vecteur d'observation et un modèle repose sur la vraisemblance (*c.f.* § 3.2.2). Les vraisemblances du modèle sont communément normalisées par la vraisemblance d'un modèle du monde [Liu 1996, Furui 1997, Gravier 1998].

Les probabilités d'émission correspondent à des rapports de vraisemblances entre les modèles X_i et W calculés, pour des blocs de longueur fixe de 0,3 seconde.

Commentaires

Le décodage de Viterbi sélectionne l'état le plus vraisemblable toutes les 0,3 secondes. Généralement, les décodeurs pour la reconnaissance de la parole utilisent une référence temporelle en centiseconde de la durée de la trame au lieu des 0,3 secondes proposées. Travailler en blocs permet d'accélérer le décodage, mais surtout les moyennes de rapports de vraisemblances utilisées sont plus robustes aux variations du signal.

D'autres solutions sont envisageables pour atténuer les variations des probabilités d'émission. Par exemple, un lissage par une fenêtre glissante est proposée dans [Reynolds 2000, Adami 2002] lors de la phase de resegmentation.

5.5.3 Critère d'arrêt de la phase d'apprentissage/décodage

Comme vu au paragraphe 4.3.3, les modèles de locuteurs sont appris durant une phase itérative en deux étapes comprenant l'adaptation des modèles suivie d'un décodage par l'algorithme de Viterbi. Cette phase s'arrête lors de la validation de la valeur prise par un critère d'arrêt, calculé à partir de la dernière segmentation et de la précédente. Deux formulations du critère d'arrêt sont présentées.

Note : La notation place en exposant le numéro de l'itération du processus. L'étape 3 étant également une phase itérative, le numéro d'une itération de la phase 3 sera aussi mis en exposant : $S^{i,j}$ est une segmentation en i locuteurs dont les modèles ont été adaptés j fois.

Formulation à partir des vraisemblances

La première formulation repose sur le critère du maximum de vraisemblance. Les probabilités des segmentations pour deux itérations consécutives sont comparées :

$$P(O, S^{i,j} | A^i, B^{i,j}) \geq P(O, S^{i,j-1} | A^i, B^{i,j-1}) \quad (5.5)$$

L'apprentissage des modèles et Viterbi optimisent le critère du maximum de vraisemblance. Quand aucun gain en vraisemblance n'est observé, l'étape est interrompue.

Formulation à partir des segmentations

Une autre formulation, équivalente au critère de l'équation 5.5, est définie par : le processus d'adaptation/décodage s'arrête lorsque deux segmentations consécutives sont identiques. Autrement dit, un gain en vraisemblance au niveau des modèles est significatif s'il entraîne une modification dans la segmentation : ce gain a permis d'affecter au moins une observation à un autre locuteur.

Commentaires

Le critère d'arrêt sur l'égalité entre deux segmentations consécutives a l'inconvénient de nécessiter un calcul supplémentaire, alors que dans la première formulation la probabilité de la séquence d'états est calculée lors du décodage de Viterbi.

5.6 Détection du nombre de locuteurs

La détection du nombre de locuteurs est réalisée lors de l'évaluation du critère d'arrêt du processus de segmentation (figure 5.9). Le critère d'arrêt permet de décider quelle est la segmentation la plus probable entre deux segmentations successives (à $i - 1$ et i locuteurs).

5.6.1 Cas de la segmentation en deux locuteurs

L'enjeu du critère d'arrêt est de décider si la segmentation en un locuteur est plus probable que la segmentation en deux locuteurs.

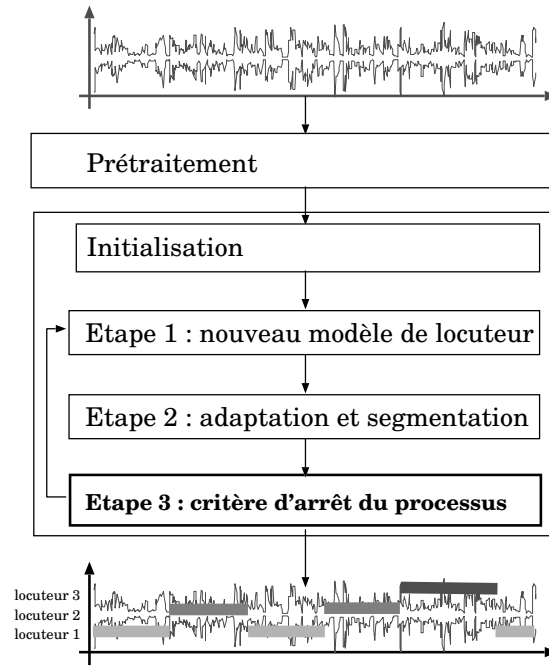


FIG. 5.9 – Evaluation de la détection du nombre de locuteurs : étape 3, critère d'arrêt.

La règle de décision, déduite du paragraphe 5.5.3, consiste à comparer la probabilité de la segmentation en deux locuteur à la probabilité de la segmentation en un locuteur :

$$P(O, S^2 | A^2, B^2) \geq P(O, S^1 | A^1, B^1) \quad (5.6)$$

Or, les HMM à un état et à deux états ne sont pas directement comparables. Les probabilités de transition sont trop différentes :

- pour un HMM à un état (un locuteur) : $a_{i,i} = 1$,
- pour un HMM à deux états (deux locuteurs) : $a_{i,i} = 0,6$ (c.f. § 5.4.2).

Les expériences montrent que la probabilité de la segmentation à un locuteur $P(O, S^1 | A^1, B^1)$ est toujours supérieure à la probabilité de la segmentation à deux locuteurs $P(O, S^2 | A^2, B^2)$. Le gain sur les probabilités d'émission du HMM à deux états ne compense pas les pertes sur les transitions (c.f. § 4.6.3).

La solution proposée est de calculer de nouveau la probabilité de S^1 à partir des transitions A^2 sans modifier la segmentation S^1 et en conservant les probabilités B^1 . La règle de décision devient :

$$P(O, S^2 | A^2, B^2) > P^*(O, S^1 | A^2, B^1) \quad (5.7)$$

5.6.2 Cas de la segmentation en n locuteurs

A la fin de l'itération i ($i > 2$), le processus évalue si la segmentation S^{i-1} en $i-1$ locuteurs est plus probable que la segmentation S^i en i locuteurs. La règle de décision pour la segmentation en deux locuteurs reste valide. Dans le paragraphe 4.6.3, il a été noté que $e_i > e_{i-1}$: l'ajout d'une classe entraîne un gain en vraisemblance pour les probabilités d'émission. Les probabilités de transition A^{i-1} sont remplacées par l'ensemble des probabilités A^i dans la séquence d'états $P(O, S^{i-1} | A^{i-1}, B^{i-1})$.

Les expériences montrent que la segmentation à i locuteurs est souvent plus probable que la segmentation à $i - 1$ locuteurs (quel que soit le nombre réel de locuteurs à détecter). Par contre, les expériences montrent aussi que, généralement, le nombre de segments $n_{k,i}^i$ dans la segmentation S^i appartenant au locuteur \mathcal{X}_i est proche de 1 pour un locuteur ajouté à tort. Une heuristique est ajoutée à la règle de décision, elle devient :

$$\begin{cases} P(O, S^i | A^i, B^i) > P^*(O, S^{i-1} | A^i, B^{i-1}) \\ \text{et} \\ n_{k,i}^i > 1 \end{cases} \quad (5.8)$$

En résumé, la règle de décision utilise les probabilités de la séquence d'états $P(O, S^i | A^i, B^i)$ et la probabilité d'états réestimée $P(O, S^{i-1} | A^i, B^{i-1})$; une condition supplémentaire est ajoutée : le nombre de segments du nouveau locuteur doit être supérieur à 1, *i.e.* le locuteur ajouté en dernier doit attirer au minimum un segment de plus que ses données d'apprentissage.

Résultat

Le tableau 5.5 montre les résultats obtenus sur le corpus de développement pour la segmentation en deux locuteurs. Les résultats sont données pour les segmentations sans et avec heuristique. L'heuristique permet de diviser par deux le taux d'erreur. Cependant, la segmentation en deux locuteurs obtient 3,23% de taux d'erreur de moins que la segmentation en n locuteurs utilisant l'heuristique. Cette heuristique est loin d'être parfaite, mais elle limite l'ajout inutile de locuteurs. Il est rappelé que le corpus de développement est composé majoritairement des enregistrements contenant deux locuteurs.

Lors des évaluations NIST utilisant des corpus plus complets (contenant des enregistrements avec plus de deux locuteurs), l'heuristique apporte un gain par rapport à la segmentation en deux locuteurs pour les corpus *Broadcast News* et *Meeting*. Pour le dernier corpus, *CallHome*, les résultats présentés dans le tableau 5.6 montrent une dégradation des performances entre le système à deux locuteurs et le système à n locuteurs. Le corpus *CallHome*, bien que multi-locuteurs, contient majoritairement des enregistrements "deux locuteurs" (303 tests sur les 500 disponibles).

La dégradation en performance sur les corpus multi-locuteurs s'explique par deux raisons⁴ :

- Tous les développements ont été fait à partir d'enregistrements à deux locuteurs. Ces développements ont permis d'obtenir un système de segmentation en deux locuteurs avec des performances à l'état de l'art.
- Les corpus *Broadcast News* et *Meeting* sont des corpus enregistrés en 16Khz alors que l'ensemble de développement est constitué d'enregistrements en 8Khz. Aucun corpus de développement n'a été fourni par NIST pour définir des paramètres spécifiques à ces nouveaux corpus. Les résultats présentés sont obtenus avec un système identique au système développé sur le corpus de développement, à l'exception du modèle du monde. Ce dernier a été appris sur environ 100 minutes de parole extraite du corpus *Santa Barbara* [LDC 1998].

⁴Les résultats sont repris et commentés dans la partie IV dédiée aux évaluations NIST.

Système	Taux d'erreur
Segmentation à deux locuteurs	10,79%
Segmentation en n locuteurs sans heuristique	31,50%
Segmentation en n locuteurs avec heuristique	14,02%

TAB. 5.5 – Critère d'arrêt du processus sur le corpus de développement : la méthode d'évaluation est la méthode NIST 2000, résultats en taux d'erreur. Résultats en taux d'erreur d'affectation de trames.

Système	Nb de loc.	Total	Nombre de locuteurs ou taux d'erreur								
			1	2	3	4	5	6	7	8	9
Nombre de tests		500	4	302	134	42	10	6	2	0	0
<i>CallHome</i>	2	21,19	23,06	12,55	21,67	26,55	34,75	45,86	49,27	-	-
<i>CallHome</i>	n	23,46	23,06	17,38	23,34	28,86	31,07	37,06	41,30	-	-
Nombre de tests		76	0	16	18	19	8	10	3	1	1
<i>Broadcast News</i>	2	29,34	-	6,55	23,28	37,92	45,09	44,09	32,79	41,57	29,80
<i>Broadcast News</i>	n	28,73	-	11,93	22,97	36,18	38,48	39,63	28,44	39,97	57,25
Nombre de tests		166	0	0	0	68	30	68	0	0	0
<i>Meeting</i>	2	44,36%	-	-	-	42,79	35,91	45,49	-	-	-
<i>Meeting</i>	n	40,63%	-	-	-	42,55	32,66	42,14	-	-	-

TAB. 5.6 – Critère d'arrêt du processus sur les corpus des évaluations NIST multi-locuteurs NIST : la méthode d'évaluation est la méthode NIST 2000, résultats en taux d'erreur d'affectation de trames, le nombre d'enregistrements par nombre de locuteurs est précisé. Nb. de loc. : indique le nombre de locuteurs fixé a priori à 2 ou à n .

Chapitre 6

Conclusion

Cette partie a été dédiée à la segmentation en locuteurs. Nous avons présenté l’approche classiquement mise en oeuvre pour cette tâche. Les outils fondamentaux utilisés pour la paramétrisation du signal et la modélisation des locuteurs ont été décrits. La description de l’approche classique a permis de mettre en évidence ces limites. En particulier, cette méthode ne remet pas en cause les informations obtenues en cours de segmentation.

Nous avons présenté une nouvelle approche au problème, utilisant une modélisation de la conversation entre les locuteurs par un HMM évolutif. Cette approche permet d’intégrer dans un même processus la détection de ruptures et la classification et ainsi de pallier le problème de la remise en cause des connaissances disponibles (les modèles de locuteurs comme les segmentations provisoires). De plus, le modèle conserve lors des différentes phases une logique du maximum de vraisemblance.

Les expériences réalisées ont validé l’approche choisie. Notamment, les transitions entre les états (les locuteurs) ont apporté un gain significatif de performance par rapport à un modèle avec des transitions équiprobables (3,45 en absolu, 24% en relatif). L’adaptation des modèles de locuteurs par l’algorithme MAP a montré son efficacité, malgré la faible quantité de données d’initialisation.

Par contre, la détection du nombre de locuteurs reste problématique, rejoignant en cela les méthodes classiques. Lors des évaluations NIST 2002, notre système a obtenu les meilleurs résultats en segmentation de locuteurs sur les conversations téléphoniques cellulaires. Par contre, les performances de la segmentation en n locuteur restent en deçà de nos espoirs, cependant le système présenté est le seul système des évaluations NIST 2002 qui détecte automatiquement le nombre de locuteurs.

Troisième partie

Appariement en locuteurs d'une collection

L'appariement automatique en locuteurs d'une collection de documents sonores est le processus aboutissant à la création d'un index identifiant les locuteurs de la collection et leurs interventions respectives. Dans cette partie, la seconde phase du processus d'indexation en locuteurs est abordée. La méthode proposée repose sur une classification hiérarchique ascendante. Deux nouvelles mesures de dissimilarité sont définies, ainsi qu'une nouvelle méthode de sélection de la partition finale. Enfin, la méthode est évaluée sur un corpus issu des évaluations NIST.

Chapitre 7

Positionnement du problème

L'appariement automatique en locuteurs d'une collection de documents sonores est le processus aboutissant à la création d'une partition en locuteurs¹ des interventions de la collection. Ce processus se décompose en trois tâches (figure 7.1). La première tâche consiste à segmenter chaque document indépendamment les uns des autres. Cette phase a été étudiée dans la partie II.

La seconde tâche consiste à identifier les locuteurs apparaissant dans plusieurs documents en utilisant les informations contenues dans les segmentations produites par la première tâche. Cette tâche revient à construire un "index d'index" (*i.e.* les index construits indépendamment pour chaque segmentation sont indexés en fonction des locuteurs de la collection). La clé de l'index est un identifiant de locuteur global à la collection. Cet identifiant référence les documents dans lesquels ce locuteur parle.

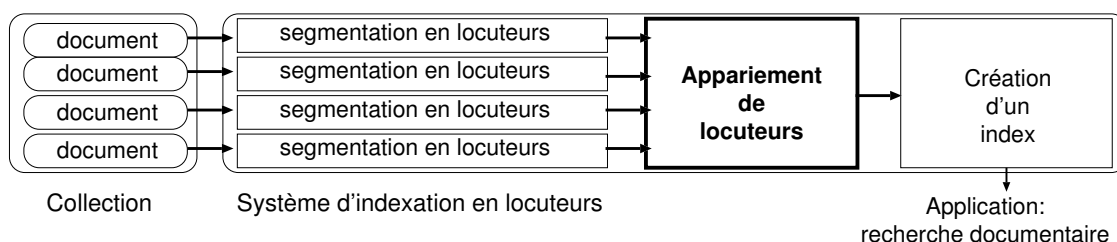


FIG. 7.1 – *Diagramme d'un système de segmentation en locuteurs.*

7.1 Segmentation et appariement en locuteurs

L'appariement en locuteurs d'une collection de documents sonores est un problème de classification proche de la tâche proposée dans [Solomonoff 1998, Reynolds 1998], où les documents sont groupés par locuteur. Dans l'appariement d'une collection, les interventions des locuteurs intervenant dans plusieurs documents sont regroupées par locuteur (figure 7.2). Il existe très peu de travaux sur le sujet ; [McLaughlin 1999] est le premier travail, à notre connaissance, sur des enregistrements segmentés.

¹Chaque classe de la partition correspondant à un locuteur.

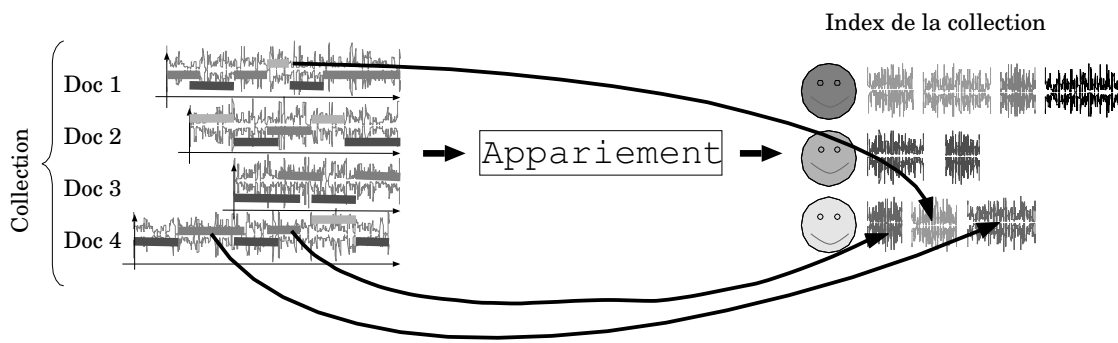


FIG. 7.2 – Exemple d'appariement de locuteurs : les interventions, extraites des documents, sont regroupées par locuteur.

7.2 Spécificités de l'appariement en locuteurs

L'appariement en locuteurs, bien que proche de la segmentation en locuteurs, montre un ensemble de caractéristiques spécifiques :

- Les hypothèses pour l'appariement en locuteurs sont identiques aux conditions de la segmentation en locuteurs d'un document sonore. Le nombre de locuteurs présents dans la collection est inconnu, mais il est supposé élevé (quelques centaines).
- Les locuteurs intervenant dans le même document ne peuvent pas être regroupés, cela remettrait en cause la segmentation du document.
- L'appariement en locuteurs opère sur une collection de documents dans lesquels un locuteur apparaît plusieurs fois. Les documents peuvent être enregistrés dans des conditions différentes. Cette variabilité (du canal de transmission pour un même locuteur) est une difficulté supplémentaire par rapport à la tâche de segmentation : lors de la segmentation en locuteurs, les systèmes tirent partie de la différence de canal ; pour l'appariement, les différences de canal risquent de dégrader les performances. Les conditions de tests sont plus proches des conditions rencontrées en identification du locuteur (IAL) en milieu ouvert que des conditions présentes lors de la segmentation d'un document.
- A la différence des tâches de VAL et IAL, la durée des interventions des locuteurs est variable : de quelques secondes à quelques minutes. Or les expériences en VAL montrent que la quantité de données influence grandement les performances, surtout pour l'apprentissage des modèles.
- Un modèle de l'intervention d'un locuteur est disponible pour chaque document. *Pour ne pas confondre le modèle de locuteur appris sur l'intervention d'un locuteur (issu d'un document) et le modèle de locuteur appris sur les interventions de la collection, nous nommerons le premier modèle "modèle d'intervention" et le second "modèle de locuteur"*. Par contre, le processus d'appariement ne peut pas recalculer les modèles de locuteurs à chaque étape, ou à chaque ajout de nouveaux documents à la collection. Le coût de calcul semble *a priori* trop important.

7.3 De l'appariement à l'application

L'appariement en locuteurs génère une partition des interventions en locuteurs. La dernière étape du processus adapte le résultat de l'appariement (la partition) à une tâche visée. Un index est construit à partir de la partition. La clé de l'index met en relation les libellés des locuteurs et

l'ensemble des interventions du locuteur. Comme dans le cadre de la segmentation, les libellés des locuteurs ne correspondent pas aux noms des locuteurs (aucune information n'est disponible pour identifier les locuteurs).

Dans le cadre des systèmes de recherche documentaire, deux moyens d'accès sont envisageables :

- Soit le système de recherche utilise des exemples d'enregistrements de la voix des locuteurs comme clé de recherche. Le système propose alors tous les documents ou toutes les parties de documents les plus similaires à l'exemple sonore du locuteur recherché.
- Soit le système de recherche utilise l'identité des locuteurs. Le mot-clé recherché est alors le nom du locuteur. Ce type de recherche implique de disposer de données externes pour attribuer un nom à chaque locuteur de la collection.

Cette dernière partie n'est pas traitée ici, elle est évoquée pour que le lecteur puisse situer la tâche dans un contexte applicatif.

7.4 Cadre de l'étude : limites

Dans ce chapitre, nous nous intéressons à l'appariement en locuteurs d'une collection de documents sonores. Nous voulons nous affranchir des difficultés relevant des erreurs commises lors de la tâche de segmentation, pour évaluer la solution d'appariement proposée. **En conséquence, les index de référence fournis par NIST avec le corpus sont utilisés comme segmentation initiale des documents en interventions de locuteur.**

Chapitre 8

Etat de l'art

Ce chapitre présente la méthode état de l'art de classification hiérarchique ascendante. Cette méthode est adaptée pour traiter le problème de l'appariement en locuteurs. Elle a été brièvement décrite dans la partie II, la description est complétée dans ce chapitre. En particulier trois éléments importants sont définis : les mesures de dissimilarité, la matrice des dissimilarités entre les classes et la méthode de sélection de la partition finale.

8.1 Principe général de la classification hiérarchique ascendante

Le principe de la classification utilisée en segmentation, décrite en particulier dans [Delacourt 1999, Moraru 2002, Reynolds 2000], est bien adapté à l'appariement d'une collection. La classification hiérarchique ascendante est la principale méthode proposée dans la littérature.

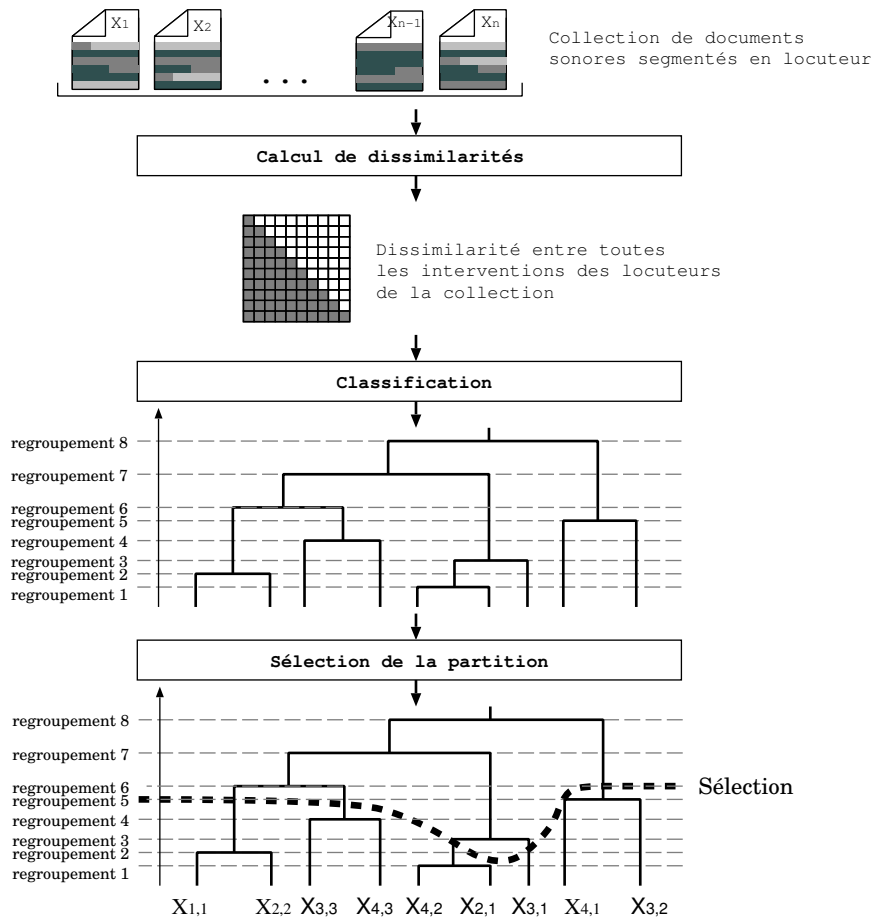


FIG. 8.1 – Appariement en locuteurs, méthode classique : classification hiérarchique.

La classification est composée de trois étapes (figure 8.1) :

1. Une matrice de dissimilarité est construite entre toutes les interventions.
2. Un processus itératif groupe les interventions : le processus est initialisé en construisant une classe par intervention $x_{k,i}$ du document X_k extraite de la collection. A chaque itération, les deux classes les plus proches sont fusionnées en une nouvelle classe. Le processus s'arrête quand une seule classe contenant toutes les interventions est construite. Les regroupements successifs forment un arbre, appelé dendrogramme.
3. A partir du dendrogramme la partition finale est sélectionnée.

Une classification hiérarchique ascendante est définie par :

- Une mesure entre les classes pour sélectionner les classes à fusionner et construire la matrice de dissimilarité.

- Une règle d'agglomération des classes. Après la fusion de deux classes, il est nécessaire de réévaluer les mesures entre cette nouvelle classe et les autres.
- Une méthode d'élagage du dendrogramme pour sélectionner la partition finale contenant toutes les interventions de la collection.

Avant d'aborder les trois points énoncés précédemment et définissant une classification hiérarchique ascendante, des éléments de combinatoire peuvent aider le lecteur à situer le problème.

8.2 Aspect combinatoire

A première vue, le problème de la classification peut être abordé en cherchant la "meilleure" partition parmi l'ensemble des partitions possibles. Or le nombre de partitions d'un ensemble à n éléments est le nombre NP_n [Saporta 1990], qui peut être obtenu par la relation de récurrence suivante :

$$\begin{aligned} NP_n &= \sum_{k=1}^n NP_{n,k} \\ \text{pour} \\ NP_{n,k} &= NP_{n-1,k-1} + kNP_{n-1,k} \\ \text{avec} \\ NP_{n,1} &= NP_{n,n} = 1 \end{aligned} \tag{8.1}$$

Par exemple pour $n = 50$, $NP_{50} = 1,87.10^{47}$ solutions ! Bien évidemment, ce résultat montre que seules des solutions approchées permettent de choisir une partition parmi l'ensemble des possibilités.

8.3 Mesure de dissimilarité

Une mesure de dissimilarité exprime la non-proximité entre deux classes contenant des interventions (*c.f.* partie II, [Saporta 1990]).

Pour les interventions $x_{k,i}$ et $x_{l,j}$ issues des documents X_k et X_l , les mesures les plus courantes sont les suivantes :

- Le rapport de vraisemblances croisé (*Cross Likelihood Ratio*, CLR, [Reynolds 1998]).

$$d_{clr}(x_{k,i}, x_{l,j}) = \frac{1}{lr(x_{k,i}|X_{l,j})} \times \frac{1}{lr(x_{l,j}|X_{k,i})} \tag{8.2}$$

- l'entropie croisée (*cross entropy*, CE, [Solomonoff 1998, Siegler 1997]) :

$$d_{ce}(x_{k,i}, x_{l,j}) = \frac{l(x_{k,i}|X_{k,i})}{l(x_{k,i}|X_{l,j})} \times \frac{l(x_{l,j}|X_{l,j})}{l(x_{l,j}|X_{k,i})} \tag{8.3}$$

$X_{k,i}$ et $X_{l,j}$ sont les modèles d'intervention appris respectivement à partir des données $x_{k,i}$ et $x_{l,j}$.

Pour les I interventions contenues dans la collection, une matrice de dissimilarité de dimension $I \times I$ est produite à partir de l'une de ces mesures. Cette matrice est composée initialement des mesures entre toutes les paires d'interventions $(x_{k,i}, x_{l,j})$ de la collection. Les dissimilarités étant symétriques, la matrice est exprimée sous la forme d'une matrice carrée triangulaire supérieure (sans la diagonale).

8.4 Estimation des dissimilarités

A l'initialisation de l'algorithme de classification, chaque intervention est placée dans une classe. Soit I le nombre d'interventions dans la collection C , l'ensemble des classes constituant la partition initiale P_I est défini par :

$$\begin{aligned}
 P_I &= \{c_y\}_I \\
 \text{pour} \\
 c_y &= \{x_{k,i}\} \\
 \text{avec} \\
 y &\in \{1, \dots, I\} \\
 k &\in \{1, \dots, K\} \\
 i &\in \{1, \dots, \text{card}(X_k)\}
 \end{aligned} \tag{8.4}$$

A chaque itération, les deux classes les plus proches c_u et c_v sont fusionnées en une classe $c_{uv} = c_u \cup c_v$. A la première fusion, une nouvelle partition P_{I-1} est construite :

$$P_{I-1} = P_I - \{c_u\} - \{c_v\} + \{c_{uv}\} \tag{8.5}$$

Itérativement, l'ensemble des partitions est obtenu de proche en proche : $\{P_I, P_{I-1}, \dots, P_i, \dots, P_2, P_1\}$. A la dernière fusion, la partition P_1 contient une seule classe contenant toutes les interventions de la collection.

Les modèles d'intervention restent coûteux à estimer, la méthode ne calcule pas de nouveaux modèles et de nouvelles dissimilarités au cours de la classification. La dissimilarité d'une nouvelle classe est estimée par rapport aux dissimilarités initiales. Pour les classes c_u et c_v fusionnées en la classe c_{uv} , la dissimilarité entre la nouvelle classe c_{uv} et une autre classe $c_w \in P_i - \{c_{uv}\}$ est calculée à partir d'une des deux règles suivantes et illustrées dans la figure 8.2 :

- *single linkage* : $d(c_{uv}, c_w)$ est la plus petite dissimilarité entre les interventions de la classe c_{uv} et des interventions d'une autre classe c_w .

$$d(c_{uv}, c_w) = \min\{d(c_u, c_w), d(c_v, c_w)\} \tag{8.6}$$

- *complete linkage* : $d(c_{uv}, c_w)$ est la plus grande dissimilarité entre les interventions de la classe c_{uv} et des interventions d'une autre classe c_w .

$$d(c_{uv}, c_w) = \max\{d(c_u, c_w), d(c_v, c_w)\} \tag{8.7}$$

8.5 Sélection de la partition finale

Le dendrogramme résultant de la fusion des différentes interventions définit une suite de partitions possibles : $\{P_I, P_{I-1}, \dots, P_2, P_1\}$. La partition finale est choisie parmi ces partitions. Plusieurs techniques existent dans la littérature [Solomonoff 1998, Everitt 1993], elles consistent à couper le dendrogramme à une hauteur donnée ou à sélectionner un ensemble de classes à différentes hauteurs.

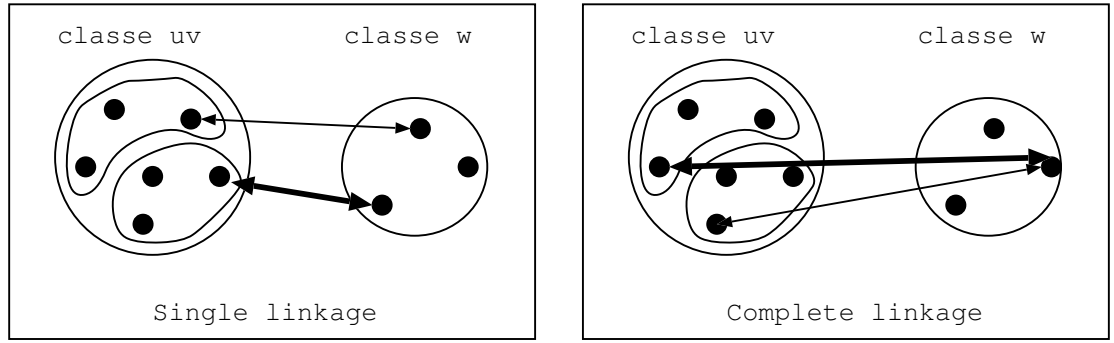


FIG. 8.2 – Estimation de la dissimilarité $d(c_{uv}, c_w)$ pour les méthodes *single linkage* et *complete linkage* de la nouvelle classe uv et une autre classe w .

8.5.1 Sélection à une hauteur donnée

Dans la méthode de sélection utilisant une hauteur de coupe horizontale du dendrogramme, la mesure I_{BBN}^1 exprimée avec la pureté des classes est une des méthodes utilisées comme critère de sélection :

$$I_{BBN} = \sum_i \sum_j n_i p_i - Q N_c \quad (8.8)$$

avec :

- n_i le nombre d'interventions dans la classe i ,
- n_{ij} le nombre d'interventions du locuteur j dans la classe i ,
- p_i est la pureté de la classe i exprimée par $p_i = \frac{n_{ii}^2}{n_i^2}$,
- N_c le nombre de classes dans la partition,
- Q un paramètre à fixer pour pondérer le nombre de classes.

La pureté mesure le nombre d'interventions provenant du même locuteur dans une classe. Une pureté égale à 1 indique que toutes les interventions sont prononcées par le même locuteur. Bien évidemment, la pureté ne peut pas être calculée directement, les libellés des classes n'étant pas disponibles. A. Solomonoff propose une estimation de cette pureté. Une pureté \hat{p}_k est estimée pour chacune des interventions de la classe c_i . Puis la pureté \hat{p}_i de la classe est définie comme la moyenne de la pureté des interventions \hat{p}_k .

$$\hat{p}_i = \frac{1}{n_i} \sum_{k \in c_i} \hat{p}_k \quad (8.9)$$

L'estimation de la pureté des interventions \hat{p}_k est calculée en trois étapes :

1. Trier toutes les interventions par ordre croissant de distance ($d(k, \cdot)$) par rapport à l'intervention k .
2. Prendre les n_i plus proches interventions et compter parmi ces interventions celles appartenant à la classe c_i . Soit $n_{c|k}$ ce nombre.
3. Définir la pureté \hat{p}_k de l'intervention k comme la fraction des premiers n_i plus proches voisins qui sont dans la classe c_i :

$$\hat{p}_k = \frac{n_{c|k}}{n_i} \quad (8.10)$$

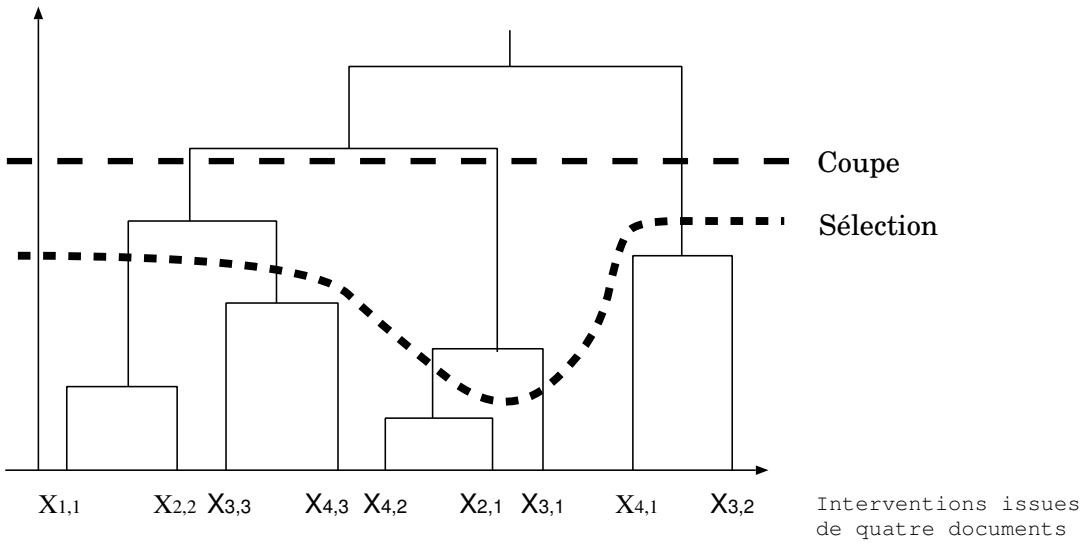
¹BBN est le nom de la société ayant proposé cette mesure.

La justification théorique ainsi que des mesures expérimentales sont données dans [Solomonoff 1998].

8.5.2 Elagage du dendrogramme

Une autre méthode de sélection de la partition finale consiste à choisir les nœuds définissant des classes dans le dendrogramme à des hauteurs différentes (figure 8.3). Une méthode itérative est proposée par [Solomonoff 1998]. Dans la suite de ce document, nous nommerons cette méthode *Des*. La méthode consiste à calculer pour chaque nœud le score s_i :

$$s_i = \hat{p}_i - \frac{Q}{n_i} \text{ avec } \hat{p}_i \text{ la pureté estimée} \quad (8.11)$$



Coupe = $\{\{X_{1,1} ; X_{2,2} ; X_{3,3} ; X_{4,3}\}, \{X_{4,2} ; X_{2,1} ; X_{3,1}\}, \{X_{4,1} ; X_{3,2}\}\}$
Sélection = $\{\{X_{1,1} ; X_{2,2}\}, \{X_{3,3} ; X_{4,3}\}, \{X_{4,2} ; X_{2,1}\}, \{X_{3,1}\}, \{X_{4,1} ; X_{3,2}\}\}$

FIG. 8.3 – Appariement en locuteurs : Elagage et coupe du dendrogramme. La méthode de sélection génère une partition en cinq classes, alors que la coupe engendre une partition en trois classes.

Puis, le dendrogramme est parcouru depuis la racine jusqu'aux feuilles. Le nœud avec le meilleur score s_i est sélectionné, ce nœud et ses fils sont supprimés de l'arbre. L'algorithme est appliqué itérativement tant qu'il reste des nœuds dans le dendrogramme.

Cette méthode pose un problème : les premiers scores s_i de l'arbre donnent obligatoirement les meilleurs scores. La méthode d'estimation de la pureté est biaisée. Une classe contenant toutes les interventions obtient une pureté estimée de 1 ! Pour résoudre cette difficulté, les auteurs coupent le dendrogramme à une hauteur donnée pour obtenir une première partition. La coupe au préalable de l'arbre impose de fixer un seuil dépendant du type de données à traiter. A partir de cette première partition, les classes sont sélectionnées en utilisant la méthode d'élagage *Des* dans les sous-arbres.

Chapitre 9

Méthode proposée

La méthode proposée dans ce chapitre reprend le cadre général de la classification hiérarchique ascendante. Nos contributions portent sur deux nouvelles mesures de dissimilarité exploitant les informations contenues dans les segmentations des documents et sur une méthode de sélection de la partition finale.

9.1 Principe général

La méthode proposée repose sur la méthode de classification hiérarchique ascendante et en particulier sur les travaux de [Solomonoff 1998]. Deux points ont retenu notre attention :

- La segmentation constitue une source d’information complémentaire par rapport aux travaux de A. Solomonoff qui travaille sur des documents mono-locuteur.
- L’élagage *Des* nécessite de couper l’arbre à une hauteur fixée au préalable.

• Information de la segmentation

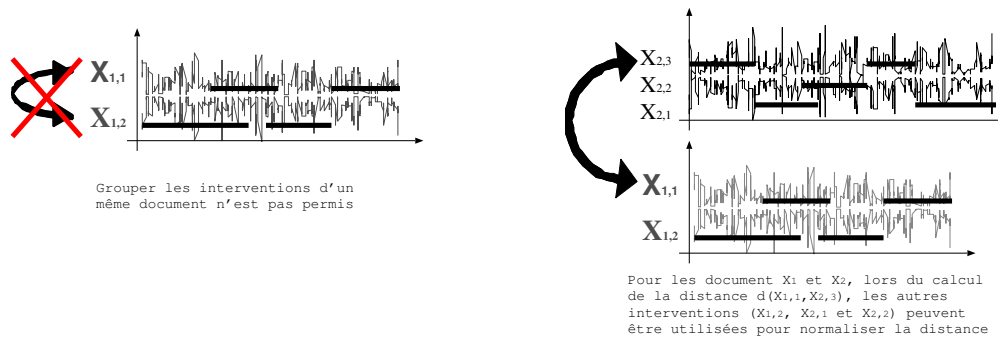


FIG. 9.1 – Appariement en locuteurs : utilisation des interventions.

Dans le cadre de l’appariement en locuteurs, la segmentation des documents, supposée sans erreur, apporte des informations supplémentaires (figure 9.1) :

- La segmentation permet de limiter le nombre de regroupements possibles. Les interventions d’un même document ne sont pas regroupées, elles n’appartiennent pas au même locuteur. Par conséquent, la classification ne génère plus un arbre complet, mais un ensemble d’arbres.
- Lors du calcul des dissimilarités, les interventions des autres locuteurs sont une source d’information. Les autres interventions d’un document peuvent être utilisées pour normaliser la dissimilarité entre deux interventions.

• Stratégie d’élagage

Une nouvelle stratégie d’élagage est proposée :

- Les scores attribués à chaque nœud reposent sur la pureté estimée, proposée par A. Solomonoff.
- Au lieu de parcourir l’arbre de la racine vers les feuilles, l’arbre est parcouru des feuilles vers la racine. Ainsi, cette méthode évite de couper l’arbre à une hauteur donnée avant d’élaguer le dendrogramme.

Pour résumer, les changements apportés à la méthode initiale, pour s’adapter à l’appariement en locuteurs d’une collection de documents multi-locuteurs, portent sur :

- La mesure de dissimilarité, deux nouvelles approches sont proposées au paragraphe 9.2.
- La matrice de dissimilarité décrite au paragraphe 9.3
- La méthode d’élagage définie au paragraphe 9.4

9.2 Mesures de dissimilarité proposées

Dans le but d'exploiter les informations apportées par la segmentation, deux mesures utilisant explicitement toutes les interventions des documents X_k et X_l sont proposées. Comme souligné dans l'état de l'art, une mesure de dissimilarité $d(x_{k,i}, x_{l,j})$ évalue la non-proximité entre deux interventions $x_{k,i}$ et $x_{l,j}$ extraites des documents X_k et X_l .

Les segmentations S_k et S_l associées aux documents X_k et X_l contiennent, en plus des interventions $x_{k,i}$ et $x_{l,j}$, les interventions correspondant aux autres locuteurs, noté $\overline{x_{k,i}}$ et $\overline{x_{l,j}}$.

Si les interventions $x_{k,i}$ et $x_{l,j}$ sont produites par le même locuteur, alors le locuteur i ne produit pas $\overline{x_{k,i}}$ et le locuteur j ne produit pas $\overline{x_{l,j}}$. Il est supposé ici que $\overline{x_{k,i}}$ et $\overline{x_{l,j}}$ sont porteuses d'informations utiles pour renforcer le facteur discriminant des dissimilarités.

• Normalisation par les données

La première dissimilarité proposée utilise les **données** ($\overline{x_{k,i}}$ et $\overline{x_{l,j}}$) des interventions des documents X_k et X_l :

$$d_x(x_{k,i}, x_{l,j}) = \frac{f(\overline{x_{l,j}}|X_{k,i}) + f(\overline{x_{k,i}}|X_{l,j})}{f(x_{l,j}|X_{k,i})f(x_{k,i}|X_{l,j})} \quad (9.1)$$

$f()$ représente soit une vraisemblance ($l()$) soit un rapport de vraisemblances ($lr()$).

• Normalisation par les modèles

La seconde dissimilarité utilise les **modèles** des interventions ($\overline{X_{k,i}}$ et $\overline{X_{l,j}}$) issues de $\overline{x_{k,i}}$ et $\overline{x_{l,j}}$:

$$d_X(x_{k,i}, x_{l,j}) = \frac{f(x_{l,j}|\overline{X_{k,i}}) + f(x_{k,i}|\overline{X_{l,j}})}{f(x_{l,j}|X_{k,i})f(x_{k,i}|X_{l,j})} \quad (9.2)$$

De même, $f()$ représente une vraisemblance ($l()$) ou un rapport de vraisemblances ($lr()$).

• Commentaires

- Lorsque $f()$ correspond à $lr()$, les deux mesures proposées reposent sur le rapport de vraisemblances croisé normalisé par les rapports de vraisemblances des modèles $\overline{X_{k,i}}$ et $\overline{X_{l,j}}$ ou par les données des interventions $\overline{x_{k,i}}$ et $\overline{x_{l,j}}$.
- $\overline{X_{k,i}}$ et $\overline{x_{k,i}}$ sont des ensembles. Ni le modèle $\overline{X_{k,i}}$, ni la dissimilarité calculée sur $\overline{x_{k,i}}$ ne sont disponibles. $l(\overline{x_{k,i}}|Y)$ et $l(y|\overline{X_{k,i}})$ sont approchées par :

$$l(\overline{x_{k,i}}|Y) = \max_{z \in \overline{x_{k,i}}} l(z|Y) \quad (9.3)$$

$$l(y|\overline{X_{k,i}}) = \max_{Z \in \overline{X_{k,i}}} l(y|Z) \quad (9.4)$$

où y (respectivement z) représente une intervention et Y (respectivement Z) le modèle associé à cette intervention.

9.3 Matrice de dissimilarité

Comme vu précédemment, la classification ne remet pas en cause le résultat de la segmentation. Les interventions d'un même document ne sont pas regroupées. La fusion doit respecter cette contrainte, aussi les dissimilarités entre les interventions d'un même document sont fixées à $+\infty$.

Après la fusion de deux classes, la matrice de dissimilarité est réévaluée suivant une des règles définies au paragraphe 8.4. La règle interdisant la fusion des interventions d'un même document

en une classe est également appliquée lors de la phase d'estimation (figure 9.2) : les dissimilarités calculées sur des interventions d'un même document sont fixées à $+\infty$. La figure 9.2 illustre les contraintes placées sur la matrice et montre la construction de la nouvelle matrice estimée.

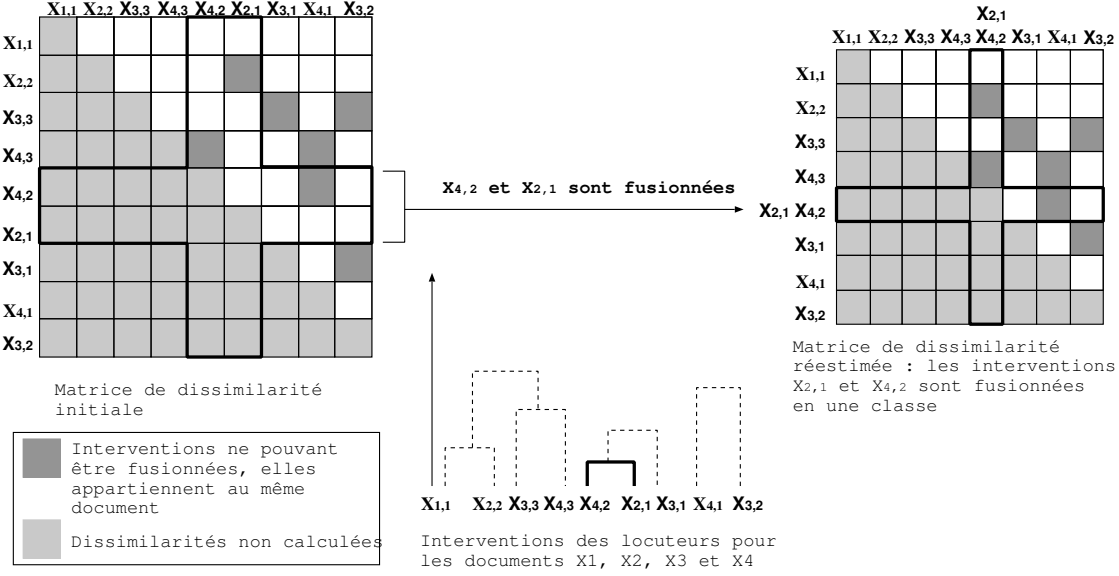


FIG. 9.2 – Appariement en locuteurs : réestimation de la matrice de dissimilarité après une fusion.

9.4 Elagage du dendrogramme

Pour résoudre le problème posé par la méthode *Des* présentée au paragraphe 8.5.2, nous proposons une nouvelle méthode, appelée *Asc*.

Méthode *Asc*

Comme dans la méthode *Des*, les scores s_i sont calculés pour chaque nœud (c.f. équation 8.11). Le dendrogramme est parcouru des feuilles jusqu'à la racine, suivant l'ordre d'agrégation des classes obtenu lors de la classification hiérarchique. Quand le score s_i n'augmente plus entre le nœud i et ses fils, les classes définies par les nœuds fils sont ajoutées à la partition. Le nœud i et ses ancêtres sont supprimés du dendrogramme. L'algorithme s'arrête quand il n'y a plus de feuille.

Commentaire

La méthode proposée est très proche de la méthode *Des* de A. Solomonoff, mais *Asc* supprime la coupe arbitraire de l'arbre à une hauteur donnée avant de procéder à l'élagage et par là même va nécessiter de fixer le seuil nécessaire à la coupe.

Les scores attribués aux nœuds sont identiques, seul le parcours du dendrogramme diffère. *Asc* élague l'arbre à des hauteurs plus basses que les hauteurs de la méthode *Des*. La méthode *Asc* génère une partition finale composée de plus de classes que la méthode *Des*.

Chapitre 10

Evaluation de la méthode

L'appariement en locuteurs est évalué dans ce chapitre à partir de données issues du corpus NIST 2000 de la tâche 2-speaker. La méthode proposée est comparée à une méthode classique. Deux évaluations sont proposées : une première expérience mesure le pouvoir discriminant des dissimilarités, puis l'appariement est évalué.

10.1 Corpus d'évaluation

L'approche proposée a été expérimentée sur des conversations téléphoniques à deux locuteurs utilisées durant la campagne d'évaluations NIST 2000/2001 pour la tâche *2-speaker* (c.f. § 12.2.2). Les segmentations de référence sont disponibles pour chacun des enregistrements. Le corpus de test est composé de 408 enregistrements extraits du corpus *Switchboard II phase II*. Le nombre de locuteurs est de 319 (132 hommes et 187 femmes). Chaque locuteur apparaît dans 1 à 4 tests (tableau 10.1) et intervient en moyenne 31 secondes ($Min \approx 14$ s., $Max \approx 53$ s.). La durée totale des tests est proche de 422 minutes.

Locuteur apparaissant dans...	1 test	2 tests	3 tests	4 tests
Nombre de locuteurs	72	90	64	93

TAB. 10.1 – Nombre de locuteurs apparaissant dans 1 à 4 tests.

10.2 Paramétrisation et modèles

La paramétrisation acoustique (16 coefficients cepstraux complétés par les dérivées premières) est calculée par le module SPRO développé par le consortium ELISA [ELISA 2000, Magrin-Chagnolleau 2001]. Les modèles et les vraisemblances sont calculés par le système de reconnaissance automatique du locuteur AMIRAL développé au LIA [Bonastre 2000, Fredouille 2000a, Fredouille 2000b]. Les interventions des locuteurs sont modélisées par un modèle de mixtures de gaussiennes (GMM) à 128 composantes (à matrices de covariances diagonales [Reynolds 1995]). Ces modèles sont adaptés depuis un modèle du monde par la méthode du *maximum a posteriori* (MAP). Seules les moyennes des modèles sont adaptées. Tous ces choix ont été guidés par les tests réalisés sur le système AMIRAL lors de la campagne d'évaluation NIST 2001 pour la tâche de vérification du locuteur.

10.3 Expériences préliminaires

La première évaluation proposée mesure la précision des dissimilarités. Des tests de vérification du locuteur, proches des conditions de la tâche de VAL de NIST (*1-Speaker*), sont réalisés entre les différentes interventions. Les dissimilarités utilisant le rapport de vraisemblances croisé d_{clr} , normalisées par les données d_x et normalisées par les modèles d_X (avec $f() = l()$ ou $f() = lr()$) sont calculées sur les interventions provenant uniquement de différents documents (pour chaque $d(u, v) \neq +\infty$ dans la matrice de dissimilarité).

Les résultats des 332112 tests de VAL sont reportés dans les courbes DET de la figure 10.1 [Martin 1997]. Une courbe DET présente les erreurs de type II (faux rejet, *Miss Probability* dans la figure) sur l'axe des ordonnées et les erreurs de type I (fausse acceptation, *False alarm* dans la figure) sur l'axe des abscisses.

Plusieurs commentaires découlent de cette expérience :

- Comme pour les résultats classiquement observés en VAL, la normalisation des scores par le modèle du monde réduit les taux d'erreurs : d_X et d_x utilisant le rapport de vraisemblances ($lr()$) obtiennent de meilleurs résultats que la vraisemblance seule ($l()$).

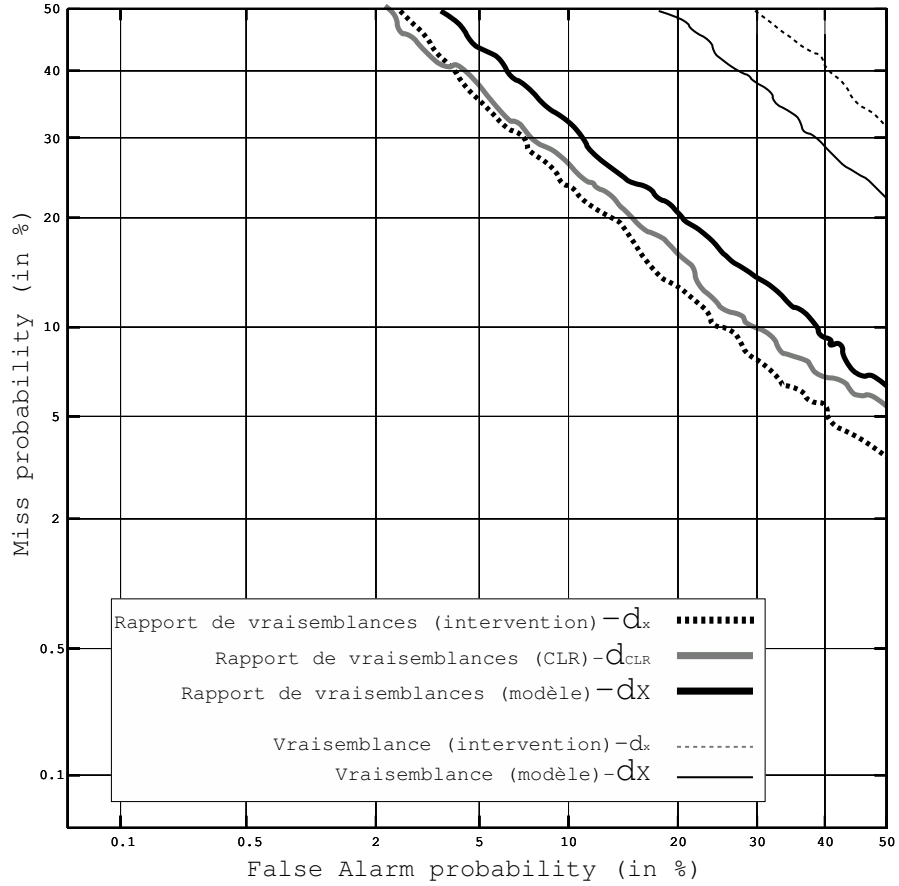


FIG. 10.1 – Courbes DET pour l'expérience en vérification du locuteur utilisant les mesures de similarité. Faux rejet = Miss Probability, fausse acceptation = False alarm. Résultats des 332112 tests pour cinq mesures différentes, seules les interventions provenant de différents documents sont utilisées.

- Les mesures étudiées obtiennent des taux d'erreurs supérieurs aux résultats obtenus lors des évaluations NIST. Le taux d'égale erreur¹ (*Equal Error Rate*, EER) est proche de 10% pour les évaluations NIST 2001 en reconnaissance du locuteur pour la tâche de VAL. Cette différence est en particulier due à la durée d'apprentissage des modèles qui est en moyenne de 31 secondes, alors que 2 minutes de parole sont disponibles dans les évaluations NIST. De plus, lors des évaluations NIST, une normalisation (ou une combinaison de normalisations) des scores de type *H-norm*, *T-norm*, *WORD+MAP* ou *D-norm* est appliquée après le calcul des vraisemblances ou des rapports de vraisemblances. Cette étape de normalisation amène une amélioration significative des performances [Gravier 2000, Auckenthaler 2000, Fredouille 1999, Ben 2002].
- La mesure d_x (avec $f() = lr()$) obtient un meilleur résultat que la mesure d_X (avec $f() = lr()$). Utiliser les données des autres locuteurs présents dans les documents augmente donc les performances. Finalement, d_x (avec $f() = lr()$) surpasse d_{clr} à l'EER. En appariement de locuteurs, l'objectif est généralement de minimiser les erreurs de classification, *i.e.* le taux de fausses acceptations (False alarm). Pour un taux inférieur à 5%, les courbes DET de d_x et d_{clr} sont trop proches pour déterminer quelle est la mesure la plus discriminante.

¹Quand le taux des faux rejets est égal aux taux de fausses acceptations.

10.4 Expériences d'appariement

L'objectif des expériences est d'évaluer les différentes mesures et les deux méthodes d'élagage. L'enjeu est bien entendu de vérifier si la classification hiérarchique permet de générer des partitions de "bonne qualité" (avec un nombre de classes proche du nombre réel tout en maximisant la pureté des classes) pour l'appariement en locuteurs. Trois points sont évalués :

- la méthode de réévaluation des dissimilarités,
- les dissimilarités,
- la méthode d'élagage.

10.4.1 Evaluation de la partition

L'évaluation de l'appariement en locuteurs est réalisée à partir de la partition obtenue après l'élagage du dendrogramme. Le calcul d'erreurs proposé dans [Chen 1998] est utilisé pour évaluer la partition. Deux types d'erreurs caractérisent la qualité d'une partition. Ces deux erreurs sont calculées à partir des valeurs suivantes :

1. Soit N_c le nombre de classes dans la partition P .
2. Soit n_i le nombre d'interventions dans la classe i .
3. Soit IN_i le nombre d'interventions du locuteur principal² de la classe i .
4. Soit OUT_i le nombre d'interventions en dehors de la classe i du locuteur principal de la classe i .

Les erreurs de type I et II s'expriment par :

$$e_I = \frac{1}{N_c} \sum_{i \in P} \frac{n_i - IN_i}{n_i}; \quad (10.1)$$

$$e_{II} = \frac{1}{N_c} \sum_{i \in P} \frac{OUT_i}{IN_i + OUT_i} \quad (10.2)$$

Enfin, pour obtenir une mesure unique, les erreurs de type I et II sont sommées :

$$e = e_I + e_{II} \quad (10.3)$$

10.4.2 Résultats

Evaluation de la méthode de réestimation des dissimilarités

Le tableau 10.2 présente les résultats des deux méthodes de réévaluation des dissimilarités. La méthode *complete linkage* engendre un nombre de classes N_c plus proche du nombre de classes réelles (319) pour un score e de même grandeur.

Evaluation des dissimilarités

Le tableau 10.3 présente les résultats des différentes mesures de dissimilarité. La mesure du rapport de vraisemblances croisé d_{clr} est la dissimilarité qui produit le meilleur score e et qui propose le nombre de classes le plus proche du nombre réel. Le nombre de classes ($N_c = 596$) est important comparé au nombre réel (319) mais le taux d'erreurs de type I, e_I , est faible ($\sim 7\%$).

²L'identité du locuteur qui minimise e_I .

Q	Méthode de réestimation							
	Complete Linkage				Single Linkage			
	Asc		Des		Asc		Des	
	e en %	N_c	e en %	N_c	e en %	N_c	e en %	N_c
0,0	63,7	596	68,6	519	64,3	643	65,1	636
0,5	67,5	537	72,8	449	70,0	598	70,6	588
1,0	70,0	505	83,5	350	77,0	544	84,5	484

TAB. 10.2 – Méthode de réestimation des dissimilarités : comparaison entre la méthode "Complete linkage" et la méthode "Single linkage", pour la mesure d_{clr} avec les valeurs du paramètre de pondération $Q \in \{0; 0,5; 1\}$; le score donné en % correspond à $e = e_I + e_{II}$. Les résultats sont donnés pour les deux méthodes d'élagage du dendrogramme "Asc" et "Des".

En choisissant une valeur du paramètre Q plus grande, moins de classes sont produites (tableau 10.2). Bien que d_x donne de meilleurs résultats que la mesure d_{clr} pour l'expérience préliminaire à l'EER, la dissimilarité d_x est moins performante que la mesure d_{clr} lors de l'appariement.

dissimilarité	Élagage Asc				Élagage Des			
	e_I	e_{II}	e	N_c	e_I	e_{II}	e	N_c
$d_X l()$	1.4	66.5	67.9	790	27.7	63.0	90.7	540
$d_x l()$	0.6	67.2	67.8	808	28.1	63.5	91.6	546
$d_X lr()$	6.6	63.8	70.4	708	21.4	60.3	81.7	563
$d_x lr()$	5.5	59.7	65.2	646	14.6	54.6	69.2	529
d_{clr}	7.4	56.3	63.7	596	15.1	53.5	68.6	519

TAB. 10.3 – Comparaison des différences mesures de dissimilarité : Les résultats sont uniquement donnés pour la méthode de réestimation des dissimilarités "Complete linkage" et suivant une valeur du paramètre de pondération $Q = 0$. Les résultats des deux méthodes d'élagage "Asc" et "Des" sont présentés. Les erreurs et le score (e , e_I , e_{II}) sont en %. $l()$ = vraisemblance, $lr()$ = rapport de vraisemblances.

Evaluation de la méthode d'élagage

La figure 10.2 montre les courbes du score e en fonction du nombre de classes pour les dissimilarités d_{clr} et d_x suivant les méthodes d'élagage *Asc* et *Des*. Les carrés et les triangles représentent les valeurs observées pour le paramètre Q variant de $[-0,3; 1,5]$. Les courbes de la méthode de sélection de partition *Asc* (les triangles) montrent que cette méthode génère plus de classes que la méthode *Des* (les carrés). Pour un nombre de classes proche, l'erreur est similaire avec l'élagage *Des* pour les deux dissimilarités d_{clr} et d_x .

Par contre, d_{clr} génère moins d'erreurs que d_x avec la méthode *Asc*, néanmoins le nombre de classes reste très important (> 500). Pour un taux d'erreurs $e < 70$, sans tenir compte du nombre de classes, la méthode *Asc* est plus performante que la méthode *Des* pour une valeur de $Q \leq 1$.

10.4.3 Commentaire

- L'expérience préliminaire a montré que les dissimilarités les plus discriminantes sont d_x et d_{clr} . Ce résultat est confirmé dans l'expérience d'appariement. Par contre, l'utilisation des données des interventions des autres locuteurs n'apporte pas un gain significatif par rapport à l'utilisation de la dissimilarité d_{clr} .

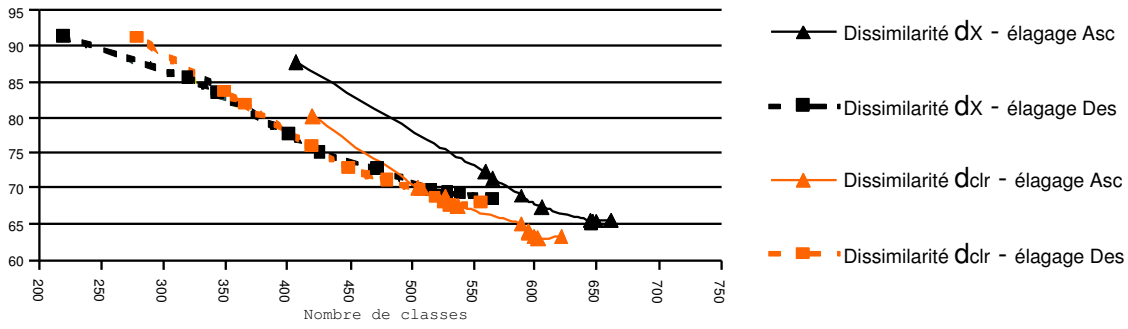


FIG. 10.2 – Courbes du score e en fonction du nombre de classes, pour les dissimilarités d_{clr} et d_x suivant les méthodes d'élagage "Asc" et "Des". Les résultats sont donnés pour les valeurs du paramètre de pondération $Q \in [-0, 3; 1, 5]$.

- La méthode d'élagage proposée, *Asc*, génère des partitions avec un nombre de classes plus important que la méthode *Des* ; mais elle ne nécessite pas de couper l'arbre à une hauteur fixée *a priori* avant l'élagage. Pour obtenir des classes très pures, la méthode *Asc* est préférable à la méthode *Des*.
- Les résultats obtenus sont satisfaisants par rapport à la difficulté de la tâche. Mais un score de 85%, cumulant les deux types d'erreurs, reste très élevé pour une partition dont le nombre de classes est proche du nombre de locuteurs réel.
- L'évaluation proposée mesure des taux d'erreurs en termes de classification des interventions. La durée des interventions n'est pas prise en compte dans le score. Une erreur sur une intervention de 30 secondes compte autant qu'une erreur sur une intervention de quelques secondes.

Chapitre 11

Conclusions et perspectives

Dans cette partie, l'intérêt de l'utilisation de la classification hiérarchique pour l'appariement en locuteurs a été évalué. Une mesure de dissimilarité usuelle en classification hiérarchique a été étudiée. Deux nouvelles mesures de dissimilarité ont été proposées dans le but de prendre en compte toute l'information disponible dans un document sonore préalablement segmenté en locuteurs. Un nouvel algorithme de sélection de la partition a été proposé. Il ne nécessite pas de couper au préalable le dendrogramme à une hauteur fixée *a priori*. Les performances obtenues sur une base de données composée de conversations téléphoniques (*Switchboard*) sont satisfaisantes au vu de la difficulté de la tâche.

Les résultats obtenus avec les dissimilarités normalisées par les informations contenues dans les segmentations sont décevants. Le potentiel de ces dissimilarités a cependant été montré lors des expériences préliminaires dans lesquelles d_x obtient le meilleur taux d'égale erreur (EER).

Il est également nécessaire d'associer un système réel de segmentation en locuteurs avec un système d'appariement en locuteurs afin d'évaluer l'influence des erreurs de segmentation sur les performances globales d'un tel système. Nous pensons que la qualité de l'appariement ne devrait pas trop être pénalisée par l'utilisation d'une segmentation réelle pour des corpus similaires à ceux utilisés dans nos expériences (conversations téléphoniques). Sur ce type de corpus, le taux d'erreurs de segmentation est faible ($\sim 15\%$ sur des enregistrements cellulaires, en incluant les erreurs provenant des locuteurs s'exprimant en simultané, *c.f.* partie IV). Cependant, les résultats de segmentation obtenus sur les réunions et les journaux télévisés montrent un taux d'erreurs proche de $\sim 30\%$, les pertes de performance pour l'appariement en locuteurs risquent d'être beaucoup plus importants dans ce cas.

La génération d'un index reposant sur les locuteurs est la dernière étape d'un système d'indexation en locuteurs. Cependant il reste encore plusieurs problèmes à résoudre pour mettre en place cette dernière phase, comme la gestion d'un grand volume de données ou le format de représentation des index afin de permettre une recherche rapide et un accès efficace aux données indexées.

Quatrième partie

Campagnes d'évaluations NIST en reconnaissance du locuteur

La méthode de segmentation en locuteurs, présentée dans la partie II, a été validée sur les tâches de segmentation des évaluations NIST en 2001/2002 en reconnaissance du locuteur. Une introduction aux évaluations NIST définit les objectifs des évaluations et les différentes tâches. Un bref descriptif des systèmes proposés par les différents participants à la tâche de segmentation en locuteurs est commenté. Les méthodes NIST d'évaluation ("scoring") des segmentations sont discutées. Enfin, les résultats des tâches de segmentation en deux locuteurs et en n locuteurs sont donnés et discutés.

Chapitre 12

Evaluations NIST

12.1 Objectifs des évaluations NIST

Depuis 1996, les évaluations annuelles en reconnaissance du locuteur sont organisées par l'institut américain NIST. Les campagnes sont ouvertes à tous les laboratoires travaillant dans ce domaine et permettent d'évaluer des conditions d'exploitation jugées problématiques. Les motivations technologiques des campagnes sont les suivantes :

- explorer de nouveaux concepts pour la RAL ;
- développer de nouvelles méthodes sur la base de ces concepts ;
- mesurer les performances de ces nouvelles méthodes.

Ces campagnes s'intéressent aussi à des problématiques spécifiques comme par exemple les durées d'enregistrements, les variations entre les sessions d'enregistrements, les variations des lignes et des combinés téléphoniques.

Les campagnes d'évaluations se déroulent en trois phases :

- Dans la première phase, NIST spécifie les tâches et les conditions d'évaluations de la campagne et définit les corpus de développement utilisables pour régler les systèmes. Cette phase a une durée variable de zéro à six mois suivant les années, la disponibilité des données, etc (pour les nouvelles tâches, il est possible qu'aucune donnée de développement ne soit fournie).
- Dans une seconde phase, les laboratoires participants reçoivent les données à évaluer. A une date fixée par NIST, les participants rendent leurs résultats. Généralement l'évaluation se déroule sur une période de 4 à 5 semaines.
- Enfin, l'ensemble des participants, des organisateurs et des sponsors se retrouvent lors d'un colloque dans lequel les résultats et les systèmes employés sont présentés et discutés.

12.2 Tâches proposées

L'historique des évaluations NIST en reconnaissance du locuteur est disponible dans [NIST 2002d]. Une vue générale des performances et des évaluations de 1996 à 2001 est donnée dans [Przybocki 1998, Martin 2000, Martin 2001]. Quatre grandes tâches ont été proposées. Ces quatre tâches nommées, *1-speaker*, *2-speaker*, *speaker tracking* et *segmentation* sont présentées dans les paragraphes ci-dessous.

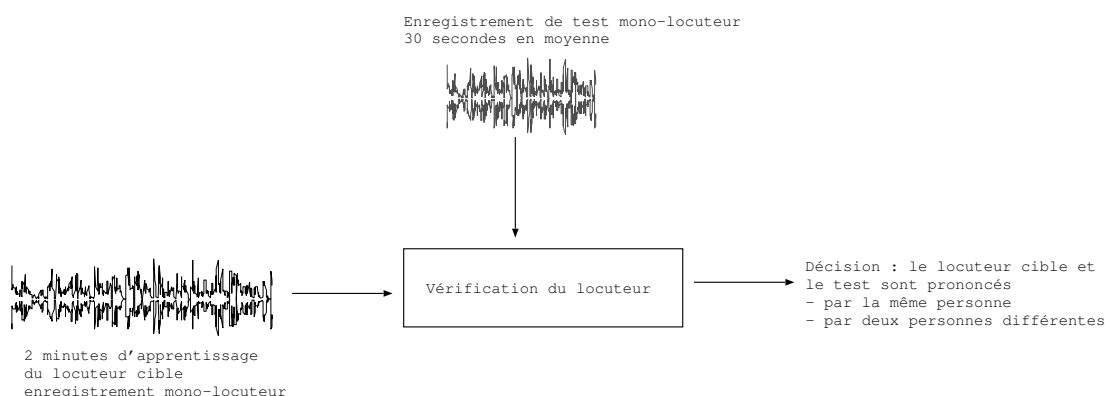


FIG. 12.1 – Tâche NIST 1-speaker : tâche de vérification du locuteur. Déterminer si un locuteur cible parle dans un enregistrement mono-locuteur de test.

12.2.1 Vérification du locuteur : 1-speaker

La tâche de vérification automatique du locuteur, appelée *1-speaker*, est la tâche de référence des évaluations NIST. Cette tâche est proposée depuis la première évaluation en 1996. La tâche de vérification du locuteur (figure 12.1) consiste à déterminer si un locuteur cible parle dans un enregistrement mono-locuteur donné (appelé test). Le système dispose d'un échantillon de la voix du locuteur cible (appelé données d'apprentissage). La tâche consiste à décider si le locuteur cible correspond au locuteur du test. Le système accepte ou rejette l'identité du locuteur cible.

Cette tâche est indirectement liée à l'indexation en locuteurs. Les méthodes mises en place dans les systèmes de vérification du locuteur (apprentissage, paramétrisation, normalisation...) sont utilisées dans la tâche de segmentation et d'appariement.

12.2.2 Vérification de locuteur : 2-speaker

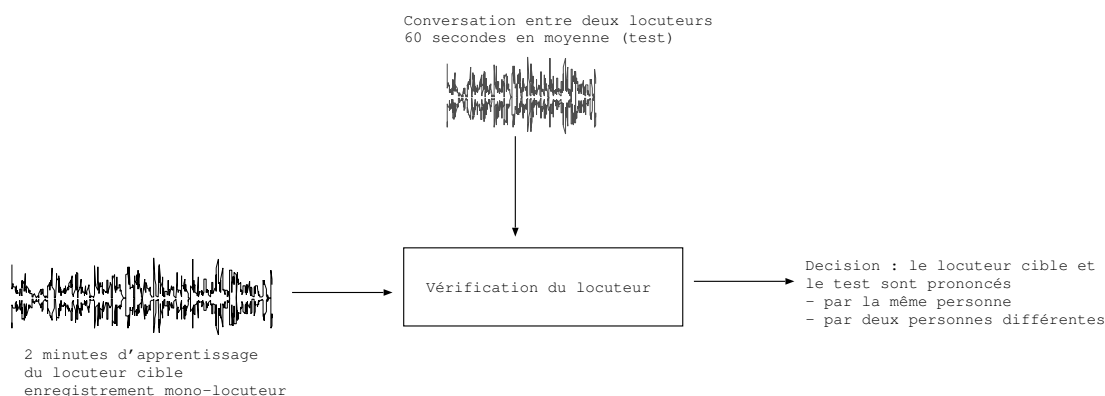


FIG. 12.2 – Tâche NIST 2-speaker en 2000 et 2001.

Cette tâche est proposée à NIST depuis 1999. La tâche consiste à déterminer si un locuteur cible parle dans une **conversation** téléphonique (appelé test). Cette tâche correspond à une tâche de vérification du locuteur (*c.f.* § 12.2.1) à l'exception des enregistrements de test qui font intervenir

des conversations téléphoniques entre deux locuteurs (*c.f.* figure 12.2). L'objectif est identique à la tâche *1-speaker*, le système accepte ou rejette l'identité du locuteur cible (NIST ne demande pas de préciser les segments où parle le locuteur cible).

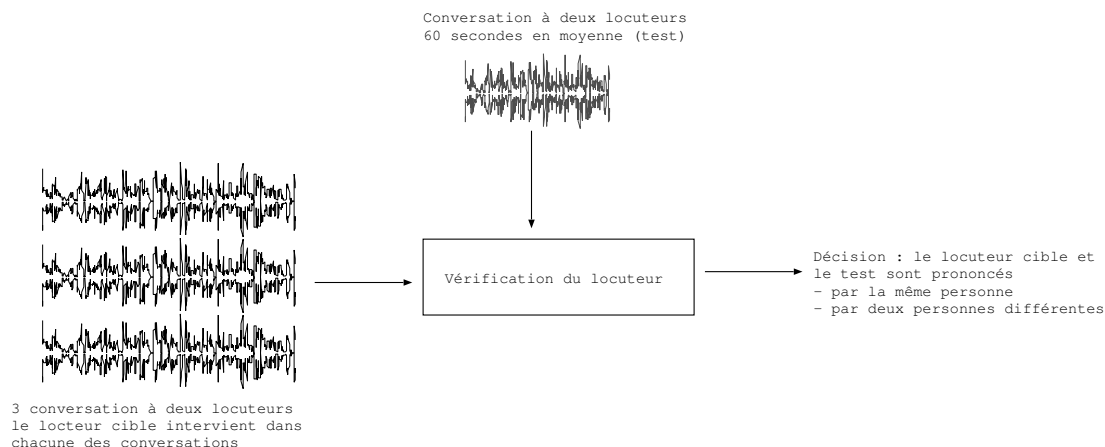


FIG. 12.3 – Tâche NIST *2-speaker* en 2002.

Une extension de la tâche *2-speaker* est proposée en 2002 (figure 12.3). Les données correspondant au locuteur cible ne sont plus constituées d'enregistrements mono-locuteur, mais de trois conversations à deux locuteurs. Le locuteur cible intervient dans les trois conversations, il parle avec trois autres locuteurs différents. Les tests sont toujours des conversations téléphoniques.

12.2.3 Suivi de locuteur : *speaker tracking*

Cette tâche a été proposée à NIST de 1999 à 2001.

La tâche consiste à déterminer si un locuteur cible parle dans une conversation téléphonique, comme pour la tâche *2-speaker* de 2000 et 2001 tout en précisant les segments où le locuteur cible intervient.

12.2.4 Segmentation en locuteurs : *2-segmentation* et *n-segmentation*

La segmentation en locuteurs correspond à la tâche définie au chapitre 3.1. Nous rappelons qu'aucune information sur les locuteurs n'est disponible : ni des échantillons de voix des locuteurs, ni le nombre de locuteurs. Toutefois, dans le cadre des évaluations le nombre de locuteurs est compris entre 1 et 10 pour la tâche *n-segmentation* et entre 1 et 2 pour la tâche *2-segmentation*. Les segmentations "hypothèses", *i.e.* les résultats rendus par les participants, sont évaluées par rapport aux segmentations de référence déterminées par NIST (*c.f.* § 13.1). Ces deux tâches ont été proposées pour la première fois en 2000 et reconduites jusqu'à présent (2002).

12.3 Systèmes proposés aux évaluations NIST

Les systèmes de segmentation du MIT, OGI, CLIPS, ELISA et du LIA sont présentés dans le tableau 12.1. Le système MIT est un système état de l'art qui est utilisé depuis les évaluations 2000 [Reynolds 2000]. OGI a un système dédié à la segmentation en deux locuteurs, avec une méthode originale de détection des locuteurs [Adami 2002]. Le système CLIPS est un système

utilisant le système AMIRAL d'ELISA pour le calcul des modèles de locuteurs et des vraisemblances [Moraru 2002]. Le système ELISA, présenté en 2002, est un système de fusion des segmentations LIA et CLIPS. Enfin, les systèmes LIA sont décrits dans la partie II et dans l'annexe C.

Modules	Option	CLIPS	MIT	OGI
Paramétrisation		16MFCC+E	23MFCC	24 LSP
Détection de ruptures		BIC	silence	GLR
Classification	taille des modèles modèle du monde estimation dissimilarité critère d'arrêt	hiérarchique GMM32 appris sur le test MAP, moyenne GLR fixé à 2 classes	hiérarchique GMM64 appris sur le test MAP, poids GLR fixé à 2 classes ou BIC	hiérarchique GMM16 appris sur le test MAP, poids GLR fixé à 2 classes
Resegmentation		oui	oui	oui
Resegmentation	modèle	GMM32	GMM <i>1-speaker</i>	GMM32
Resegmentation	modèle du monde	appris sur le test	<i>1-speaker</i>	appris sur le test
Modules	Option	LIA 2000	LIA 2001	LIA 2002
Paramétrisation		16LFCC+Delta	16LFCC+Delta	20LFCC+E
Détection de ruptures		BIC	<i>a priori</i> (0,3s)	<i>a priori</i> (0,3s)
Classification	taille des modèles modèle du monde estimation dissimilarité critère d'arrêt	VAL + seuil GMM128 données externes MAP lr() - -	HMM GMM128 données externes MAP, moyenne lr() $P(Q)$ + heuristique ou fixé à 2 classes	HMM GMM128 données externes MAP, moyenne lr() $P(Q)$ + heuristique ou fixé à 2 classes
Resegmentation		non	non	non
Resegmentation	modèle	-	-	-
Resegmentation	modèle du monde	-	-	-

TAB. 12.1 – Résumé des différentes méthodes proposées à NIST de 2000 à 2002.

Une dernière comparaison sur les temps d'exécution obtenus lors de l'évaluation NIST 2002 est reportée au tableau 12.2. Le système proposé en 2002 par le LIA a un temps de calcul en dessous du temps réel aussi bien avec un nombre de classes fixé à 2 (cas de la segmentation en deux locuteurs) qu'avec une détection en automatique du nombre de locuteurs (cas de la segmentation en n locuteurs). La stratégie descendante de ce système réduit le nombre d'opérations à effectuer par rapport à la stratégie ascendante.

Système	Nombre de locuteurs	× temps réel	Machine	temps d'exécution
CLIPS	fixé à 2	2,52	Pentium III 750MHz	2220m
CLIPS		1,83 (estimé)	<i>Pentium III 1GHz</i>	
LIA	fixé à 2	0,29	Pentium III 1GHz	249m
LIA	déTECTÉ en automatique	0,63	Pentium III 1GHz	555m

TAB. 12.2 – Comparaison des temps d'exécution des systèmes de segmentation LIA et des méthodes CLIPS : les temps de calcul sont donnés en fonction de la tâche de segmentation en locuteur lors de l'évaluation NIST 2002. La deuxième valeur du CLIPS correspond à la valeur estimée du temps d'exécution par rapport à une machine identique à celle du LIA (un facteur de $\frac{3}{4}$ a été appliqué). Les systèmes CLIPS et LIA reposent sur la même bibliothèque pour le calcul des modèles de locuteurs et des vraisemblances (AMIRAL-ELISA) et sur les mêmes outils de paramétrisation (SPRO-ELISA).

Chapitre 13

Evaluation du résultat de la segmentation

Ce chapitre présente les méthodes d'évaluation d'une segmentation. Les méthodes proposées reposent pour la plupart sur les travaux de NIST. L'évaluation d'une segmentation "hypothèse" suppose qu'une segmentation de référence est disponible. L'évaluation de la segmentation doit répondre aux besoins de l'application visée. Les critères de qualité d'une segmentation sont à spécifier. A partir de ces critères, une métrique est définie pour mesurer les différences entre l'hypothèse et la référence.

13.1 Création des segmentations de référence

Les segmentations de référence sont construites par une méthode automatique ou manuellement suivant les types d'enregistrement disponibles. Lors de la création des références, il est difficile de déterminer avec précision les frontières entre les segments. Par exemple, les respirations des locuteurs et les débuts de segments commençant par une plosive sont subjectifs.

NIST détermine les segments de référence à partir du résultat d'un détecteur d'énergie appliqué à chaque canal d'enregistrement ¹. Le détecteur d'énergie donne alors la segmentation individuelle de chaque locuteur. Si les enregistrements de part et d'autre de la conversation ne sont pas disponibles, les enregistrements sont segmentés manuellement (NIST ne précise pas la méthode).

Les 250 ms à la fin de chaque segment sont ignorés pour pallier les erreurs de positionnement des frontières entre les segments. Les segments d'un même locuteur séparés par un silence d'au plus 0,05 seconde sont fusionnés.

13.2 Critères de qualité d'une segmentation

La méthode d'évaluation doit répondre aux besoins de la tâche visée. Nous considérons qu'une segmentation de qualité doit remplir les critères suivant :

1. Tous les locuteurs sont présents, le nombre de locuteurs dans la segmentation "hypothèse" est identique au nombre de locuteurs de la segmentation de référence.
2. Les changements entre les locuteurs sont correctement détectés (les ruptures).
3. Les segments sont affectés au "bon" locuteur.
4. Seules les zones de parole sont segmentées (les zones de silence n'apparaissent pas dans la segmentation).
5. Les zones contenant plusieurs locuteurs sont correctement détectées et affectées aux locuteurs correspondant.

13.3 Evaluation de la détection du nombre de locuteurs

L'évaluation du nombre de locuteurs permet de mesurer les différences en nombre de locuteurs entre la segmentation "hypothèse" et de référence. Soit $L_{X_k}^{hyp}$ le nombre de locuteurs dans la segmentation "hypothèse" et $L_{X_k}^{ref}$ le nombre de locuteurs dans la segmentation de référence. $|L_{X_k}^{hyp} - L_{X_k}^{ref}|$ indique la différence en valeur absolue entre les deux nombres de locuteurs. Pour le corpus C , la moyenne des différences est calculée par :

$$E_{loc} = \sum_{k=1}^K |L_{X_k}^{hyp} - L_{X_k}^{ref}| \quad (13.1)$$

La valeur absolue est nécessaire, sinon le nombre de locuteurs détectés à tort est compensé par le nombre de locuteurs non détectés. Cette moyenne de valeurs absolues est difficile à interpréter, elle n'indique pas si le système de segmentation génère trop ou pas assez de locuteurs.

Une autre approche, résolvant le problème, est de s'intéresser à la répartition des locuteurs détectés plutôt qu'au nombre réel de locuteurs. Un tableau est construit avec en ligne les nombres de locuteurs de référence et en colonne les valeurs $L_{X_k}^{hyp} - L_{X_k}^{ref}$. Pour chaque couple $(L_{X_k}^{hyp}, L_{X_k}^{ref})$

¹Les canaux sont ensuite ajoutés pour obtenir le fichier de test.

$L_{X_k}^{hyp} - L_{X_k}^{ref}$	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
$L_{X_k}^{ref} = 1$	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 2$	0	0	0	0	0	0	0	127	124	41	8	2	0	0	0	0	0
$L_{X_k}^{ref} = 3$	0	0	0	0	0	0	27	57	36	12	1	1	0	0	0	0	0
$L_{X_k}^{ref} = 4$	0	0	0	0	0	0	27	7	5	3	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 5$	0	0	0	0	1	5	2	2	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 6$	0	0	0	0	3	1	1	1	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 7$	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 8$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 9$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 10$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TAB. 13.1 – Exemple de la répartition du nombre de locuteurs détectés sur le corpus CallHome. En ligne : le nombre de locuteurs dans la référence. En colonne : la différence entre le nombre de locuteurs dans l'hypothèse et le nombre de locuteurs dans la référence.

du document X_k , la case correspondante dans le tableau est incrémentée de 1. Le tableau 13.3 propose un exemple de répartition du nombre de locuteurs. La valeur la plus élevée sur une ligne montre que le système de segmentation ne trouve dans la plupart des cas que deux locuteurs.

13.4 Evaluation des ruptures

Les ruptures dans la segmentation correspondent à la frontière entre deux segments. Perrine Delacourt propose d'évaluer les erreurs de positionnement des ruptures par la méthode décrite dans [Delacourt 2000a]. Deux types d'erreurs de positionnement des ruptures peuvent être commis :

- Une frontière est présente dans la référence mais elle est absente de l'hypothèse. L'erreur est une détection manquée (*MD*, *miss detection*), le système n'a pas trouvé la rupture.
- Une frontière est absente de la référence mais elle est présente dans l'hypothèse. L'erreur est une fausse alarme (*FA*, *false alarm*), le système a détecté une rupture alors qu'elle n'existe pas.

Ces deux erreurs sont usuelles dans le domaine de la théorie de la détection, le lecteur est invité à consulter [Bonastre 2000], par exemple, pour plus d'information. *MD* et *FA* sont calculées suivant les équations données ci-dessous :

$$MD = \frac{n_{MD}}{r_{ref}} \quad (13.2)$$

$$FA = \frac{n_{FA}}{r_{ref} + n_{FA}} \quad (13.3)$$

Où n_{MD} est le nombre de détections manquées, n_{FA} est le nombre de fausses alarmes et r_{ref} est le nombre de ruptures dans la segmentation de référence.

Le positionnement des frontières est subjectif ou sujet à erreur. Généralement, une tolérance d'erreurs de positionnement de l'ordre de quelques dizaines de millisecondes est acceptée lors de l'évaluation.

Une autre méthode d'évaluation des ruptures est proposée par T. Kemp dans [Kemp 2000]. Cette mesure reprend les méthodes d'évaluation utilisées dans le domaine de la recherche documentaire : la précision et le rappel.

13.5 Evaluation des segments

L'évaluation des segments mesure la qualité des segments en termes de détection et d'affectation au bon locuteur. Cette évaluation, plus générale, intègre indirectement l'évaluation du nombre de locuteurs et des ruptures : les erreurs sur le nombre de locuteurs ou sur les ruptures génèrent des erreurs d'affectation.

13.5.1 NIST 2000 et 2001

NIST en 2000 a proposé une méthode d'évaluation des segmentations pour leurs campagnes d'évaluations [NIST 2001, NIST 2000]. Cet outil est public.

Evaluation d'une segmentation

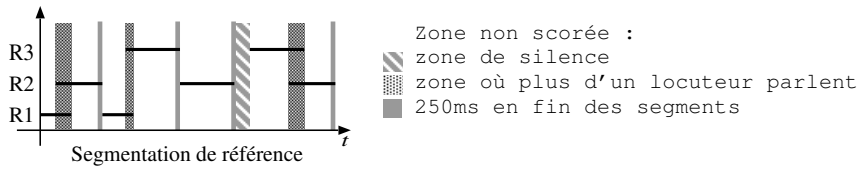


FIG. 13.1 – NIST 2000/2001, erreurs d'affectation : zones rejetées du calcul d'erreur.

La segmentation S_k est évaluée en comparant les segments "hypothèse" avec les segments de référence. Les zones à évaluer sont déterminées à partir de la segmentation de référence (*c.f.* figure 13.1). Seules les zones de signal contenant de la parole et où un seul locuteur parle sont utilisées dans le calcul.

Deux types d'erreurs sont possibles sur les segments : le faux rejet et la fausse acceptation des segments. Ces deux erreurs, qui sont communément utilisées en RAL, ne sont pas indépendantes dans le cas de la segmentation en locuteurs. Le faux rejet d'un segment provoque une fausse acceptation pour un autre locuteur.

Les interventions de chaque locuteur "hypothèse" sont comparées avec celles des locuteurs de référence pour calculer une erreur d'affectation. L'erreur d'affectation $E1_{S_k}$ est calculée en minimisant l'erreur d'affectation des locuteurs "hypothèse" aux locuteurs de référence (*c.f.* figure 13.2). $E1_{S_k}$ est calculée en recherchant les paires de locuteurs (un locuteur "hypothèse" et un locuteur de référence) parmi l'ensemble des paires de locuteurs possibles pour que l'erreur d'affectation soit minimale.

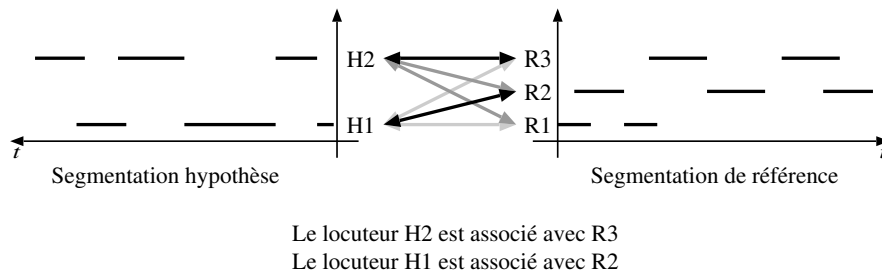


FIG. 13.2 – NIST 2000/2001, erreur d'affectation : construction des paires de locuteurs.

L'erreur d'affectation mesure une erreur de classification en durée sur les segments. Un locuteur parlant peu génère peu d'erreurs. En effet, dans le cas d'un enregistrement contenant un locuteur dominant (en durée) la conversation est plus facile à segmenter. Ce biais est quantifié en calculant l'erreur d'affectation à partir d'une segmentation triviale appelée *Baseline*, contenant un seul segment couvrant la totalité du fichier (*i.e.* un seul locuteur). La valeur obtenue reflète la difficulté du test.

Les changements de locuteurs sont indirectement mesurés par cette méthode d'évaluation. Un locuteur ajouté à tort (ou non détecté) entraîne une erreur de classification. Il en est de même pour les ruptures entre deux segments appartenant à des locuteurs différents. Un décalage de la frontière engendre une erreur d'affectation proportionnelle à la durée du décalage. Par contre, une frontière placée entre deux segments appartenant au même locuteur n'entraîne pas d'erreur.

Cependant, ni les zones de silence, ni les zones contenant plusieurs locuteurs ne sont prises en compte dans le calcul de l'erreur d'affectation. Cette méthode d'évaluation ne respecte pas l'ensemble des critères de qualité d'une segmentation défini au paragraphe 13.2.

Evaluation d'un corpus segmenté

L'erreur d'affectation d'une collection de documents segmentés correspond à la moyenne des erreurs d'affectation :

$$E1 = \sum_{S_k} \frac{E1_{S_k}}{d_k} \quad (13.4)$$

Avec $E1_{S_k}$ l'erreur d'affectation de la segmentation S_k et d_k la durée des segments évalués dans S_k . Cette moyenne prend en compte la durée de chaque test. $E1$ est une valeur comprise entre 0 et 1.

13.5.2 NIST 2002

L'évaluation des segmentations proposée lors de NIST 2002 [NIST 2002b] comble une partie des manques de la méthode d'évaluation NIST 2001. Le score intègre les erreurs sur les zones de silence et les erreurs sur les zones contenant plusieurs locuteurs (*c.f.* figure 13.3). Seule la suppression des 0,25 s en fin de segment est conservée ainsi que la fusion des segments d'un même locuteur séparés de moins de 0,05 s.

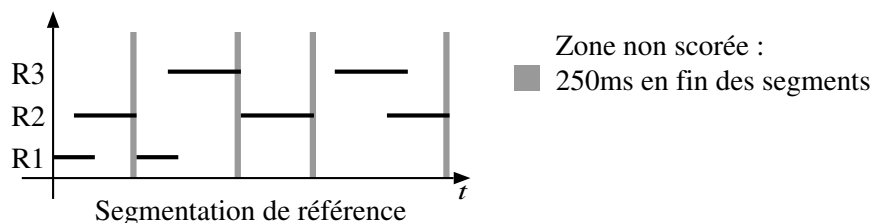


FIG. 13.3 – NIST 2002, erreurs d'affectation : zones rejetées du calcul d'erreur.

Le score final est composé de la somme pondérée des différentes erreurs suivantes :

- Les erreurs Parole/Silence
- $E2_{FR}$: l'erreur de non-détection d'un segment de parole. Cette erreur peut être vue comme les faux rejets des segments de parole.

- $E2_{FA}$: l'erreur de non-détection d'un segment de silence. Cette erreur correspond à une fausse alarme sur un segment de parole.
- Les erreurs sur les segments de parole
 - $E2_{FA\ spk}$: taux d'erreurs généré par les locuteurs non détectés.
 - $E2_{FR\ spk}$: taux d'erreurs généré par des locuteurs inexistant dans la segmentation de référence.
 - $E2_{spk}$: taux d'erreurs d'assignation entre deux locuteurs détectés.

Le score global est obtenu par la somme pondérée des cinq types d'erreurs. Pour l'évaluation 2002, les coûts des différentes erreurs sont fixés à 1 (c.f. équation 13.5).

$$E2_{S_k} = (c_{FR} \cdot E2_{FR} + c_{FA} \cdot E2_{FA}) + (c_{FA\ spk} \cdot E2_{FA\ spk} + c_{FR\ spk} \cdot E2_{FR:spk} + c_{spk} \cdot E2_{spk})$$

avec

$$c_{FR} = c_{FA} = c_{FA\ spk} = c_{FR\ spk} = c_{spk} = 1 \quad (13.5)$$

Le score $E2$ d'un ensemble de segmentations est obtenu par la moyenne des scores pondérée par la durée des scores de chaque fichier.

Pour intégrer la difficulté du corpus, NIST propose de normaliser $E2_{S_k}$ par le taux d'erreurs obtenu par la segmentation *Baseline*² :

$$S = \frac{E2_{S_k}}{E2_{Baseline}} \quad (13.6)$$

Cette nouvelle méthode d'évaluation est plus précise que la méthode proposée en 2000 et 2001. Tout en conservant un score unique, de nouvelles informations sont prises en compte. Les erreurs de segmentation Parole/Silence ($E2_{FA}$, $E2_{FR}$) sont maintenant évaluées. Cependant, NIST fournit une segmentation *Parole* et *Silence*, qui peut être utilisée dans le système de segmentation. L'erreur d'affectation $E1$ est répartie en trois types d'erreurs ($E2_{FA\ spk}$, $E2_{FR\ spk}$, $E2_{spk}$) permettant ainsi d'évaluer en termes de taux d'erreurs le nombre de locuteurs détectés. Si le nombre de locuteurs a été correctement détecté alors les taux d'erreurs $E2_{FA\ spk}$ et $E2_{FR\ spk}$ sont nuls. La figure 13.4 illustre les liens entre les deux scores d'évaluation $E1$ et $E2$.

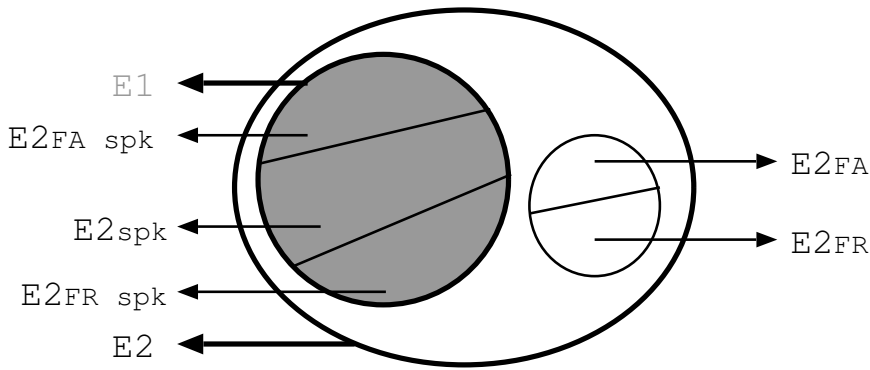


FIG. 13.4 – Erreur NIST 2002 v.s. erreur 2000 & 2001 : $E1$: erreur obtenue par la méthode NIST 2000/2001. $E2$: erreur obtenue par la méthode NIST 2002 $E2$ se décompose en erreurs sur parole/silence ($E2_{FA}$, $E2_{FR}$ et en erreurs sur les locuteurs ($E2_{FA\ spk}$, $E2_{FR:spk}$, $E2_{spk}$).

²Une segmentation Baseline est une segmentation contenant tous les segments de parole affectés à un seul locuteur.

Chapitre 14

Segmentation en locuteurs : *n-segmentation*

Les résultats aux évaluations NIST de 2000 à 2002 obtenus par notre système sur la tâche de segmentation en n locuteurs sont donnés. Dans un premier temps, le corpus utilisé au cours de ces trois évaluations est décrit.

14.1 Corpus d'évaluation

14.1.1 NIST 2000 et 2001 : *CallHome*

Le corpus utilisé en 2000 et réutilisé en 2001 est composé de 500 enregistrements extraits de la base *CallHome* [LDC 1998]. Chaque test, d'une durée inférieure à 10 minutes, contient une conversation spontanée en milieu téléphonique filaire (*landline*). De 2 à 7 locuteurs interviennent dans les conversations (*c.f.* tableau 14.1) et la base contient six langues (*c.f.* tableau 14.2).

Nombre de locuteurs	Nombre d'enregistrements
2	303
3	136
4	43
5	10
6	6
7	2

TAB. 14.1 – *CallHome* : Répartition des tests en fonction du nombre de locuteurs intervenant.

Langue	Nombre d'enregistrements
arabe	95
anglais	56
allemand	67
japonais	68
mandarin	118
espagnol	96

TAB. 14.2 – *CallHome* : Répartition des tests en fonction de la langue des intervenants.

14.1.2 NIST 2002 : *Meeting* et *Broadcast News*

La tâche de *n-segmentation* est évaluée sur deux nouveaux corpus : 166 réunions (*Meeting*, [NIST 2002a]) enregistrées par NIST et 76 enregistrements d'émissions de télévisions extraits du corpus *Hub 4 Broadcast News* [LDC 1998] :

- Le corpus *Meeting* contient des conversations spontanées multi-locuteurs enregistrées à partir de micro-casques ou d'un microphone de table. Le corpus *Meeting* est composé d'enregistrements en bande large de bonne qualité.
- Le corpus *Broadcast News* est composé d'enregistrements en bande large (studio) et en bande étroite (téléphone). Le présentateur du journal est enregistré en studio alors que les reporters extérieurs peuvent être enregistrés à partir de téléphones. De plus, les conversations ne sont pas forcément spontanées, les journalistes lisant les nouvelles qu'ils présentent. Les enregistrements *Broadcast News* ne contiennent que de la parole, les génériques d'introduction ont été supprimés des enregistrements.

Pour ces deux corpus, NIST fournit aussi une segmentation Parole/Silence obtenue à partir d'un détecteur de silence.

Le nombre de tests en fonction du nombre de locuteurs est donné dans le tableau 14.1.2. Nous remarquons que dans le corpus *Meeting* la répartition du nombre de locuteurs n'est pas uniforme : les tests contiennent uniquement 4 ou 6 locuteurs.

Nombre de locuteurs	Nombre de tests	
	<i>Meeting</i>	<i>Broadcast News</i>
2	0	16
3	0	18
4	68	19
5	0	8
6	98	10
7	0	3
8	0	1
9	0	1

TAB. 14.3 – *Meeting et Broadcast News : Répartition des tests en fonction du nombre de locuteurs.*

14.2 Comparaison des corpus

Trois différences majeures entre les corpus sont à noter, elles portent :

- sur la qualité des enregistrements,
- sur la quantité de tests par nombre de locuteurs,
- et sur la nature des conversations.

Les données d'évaluation 2000 et 2002 se distinguent par la qualité d'enregistrement. Le corpus de 2000 est composé de conversations téléphoniques alors que le corpus de 2002 est enregistré en qualité studio (excepté pour certains reportages présents dans les enregistrements *Broadcast News*). Pour la première année, NIST introduit en 2002 des données enregistrées en bande large. Aucun corpus de développement n'a été fourni avec les données *Meeting* et *Broadcast News*. Les systèmes de segmentation ont été développés à partir des corpus en parole téléphonique.

Le corpus 2002 atténue le biais de la répartition du nombre de locuteurs par test *CallHome*. *CallHome* contient majoritairement des tests à deux locuteurs (66% des tests) alors que le corpus 2002 contient 4,7 locuteurs en moyenne par test (contre 2,5 pour *CallHome*). Toutefois, le corpus 2002 est majoritairement constitué de tests à 4 et 6 locuteurs. Dans les corpus 2000 et 2002, la répartition du nombre de tests en fonction du nombre de locuteurs est loin d'être uniforme.

Les conversations des corpus *CallHome* et *Meeting* sont spontanées alors que le corpus *Broadcast News* contient des conversations préparées. Les différences entre ces corpus sont montrées dans l'histogramme de durées des segments (*c.f.* figure 14.1). La durée des segments en moyenne est de 2,1 secondes et 1,1 secondes respectivement pour *CallHome* et *Meeting*. Pour le corpus *Broadcast News*, la durée moyenne des segments atteint 12 secondes.

14.3 Résultats

14.3.1 NIST 2000 et 2001

Les résultats donnés dans le tableau 14.4 sont obtenus avec la méthode d'évaluation NIST 2000 (*c.f.* § 13.5.1). Un gain de performance absolu de 12% (33% en relatif) est obtenu entre le système

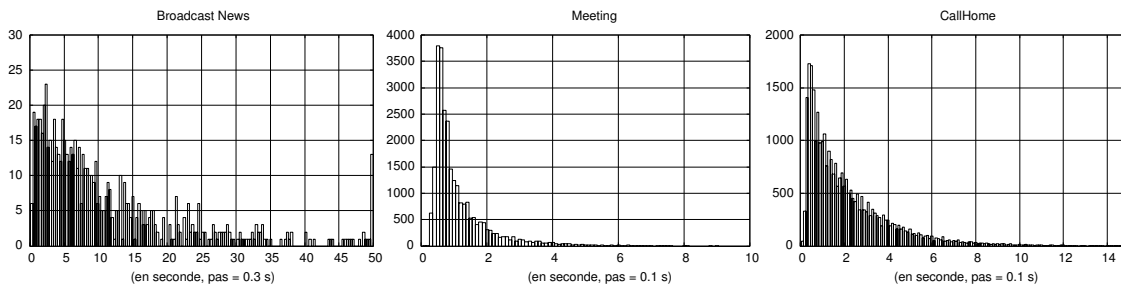


FIG. 14.1 – Histogrammes de durées des segments pour les corpus CallHome, Broadcast News et Meeting.

Système	CallHome	Nombre de locuteurs					
		2	3	4	5	6	7
Baseline	0,38	0,26	0,39	0,49	0,50	0,56	0,61
LIA 2000	0,36	0,26	0,37	0,46	0,47	0,53	0,59
LIA 2001	0,24	0,24	0,24	0,22	0,29	0,38	0,34
LIA 2002 (en n loc.)	0,23	0,17	0,23	0,28	0,31	0,37	0,41
LIA 2002 (en 2 loc.)	0,21	0,12	0,21	0,26	0,34	0,45	0,49
Meilleur système	0,20	0,13	0,21	0,22	0,32	0,38	0,40

Système	CallHome	Arabe	Anglais	Allemand	Japonais	Mandarin	Espagnol
Baseline	0,38	0,40	0,24	0,30	0,33	0,42	0,42
LIA 2000	0,36	0,37	0,25	0,29	0,33	0,40	0,41
LIA 2001	0,24	0,22	0,26	0,24	0,26	0,24	0,25
LIA 2002 (en n loc.)	0,23	0,25	0,14	0,18	0,26	0,23	0,25
LIA 2002 (en 2 loc.)	0,21	0,22	0,10	0,16	0,22	0,20	0,24
Meilleur système	0,20	0,20	0,14	0,14	0,22	0,23	0,22

TAB. 14.4 – Résultat NIST *n*-segmentation 2000 et 2001 sur la base CallHome. Les résultats représentent une erreur de classification *E1* en retirant 0,25 seconde à la fin de chaque segment. Les segments séparés de 0,5 seconde et appartenant à même locuteur sont fusionnés. Les résultats sont donnés en fonction du nombre de locuteurs dans le test et en fonction de la langue du test pour le système LIA de 2000 à 2002. Le résultat "Baseline" correspond au résultat d'une segmentation triviale contenant un seul segment par enregistrement. Le résultat du meilleur système est aussi donné.

2000 et le système 2001. Le système 2000 a des performances très proches du système Baseline¹. En effet, seul un gain de 4% en absolu est observé entre le système Baseline et le système de 2000, ce faible écart étant essentiellement dû au fait que le système 2000 a été développé pendant la période d'évaluation et qu'aucun test préliminaire n'a pu être réalisé.

Le système 2001 est adapté pour la segmentation de fichiers multi-locuteurs pour un nombre de locuteurs inférieur ou égal à quatre². Le score du système est stable bien que le modèle du monde soit uniquement appris sur des données de langue anglaise issues de NIST 99 (*Switchboard II phase II*). Sur l'ensemble du corpus, le système obtient un score de 24%, suivant la langue ce taux variant de $\pm 2\%$ en absolu.

En complément des taux d'erreurs par langue et par nombre de locuteurs, la figure 14.2 montre l'histogramme des erreurs par test.

La détection du nombre de locuteurs est représentée sous forme de tableau (*c.f.* § 13.3). Les

¹Pour chaque test, la segmentation est composée d'un segment couvrant la totalité du fichier. Le système *Baseline* mesure la difficulté du corpus.

²Au-dessus de quatre, le nombre de tests n'est pas assez important pour être significatif.

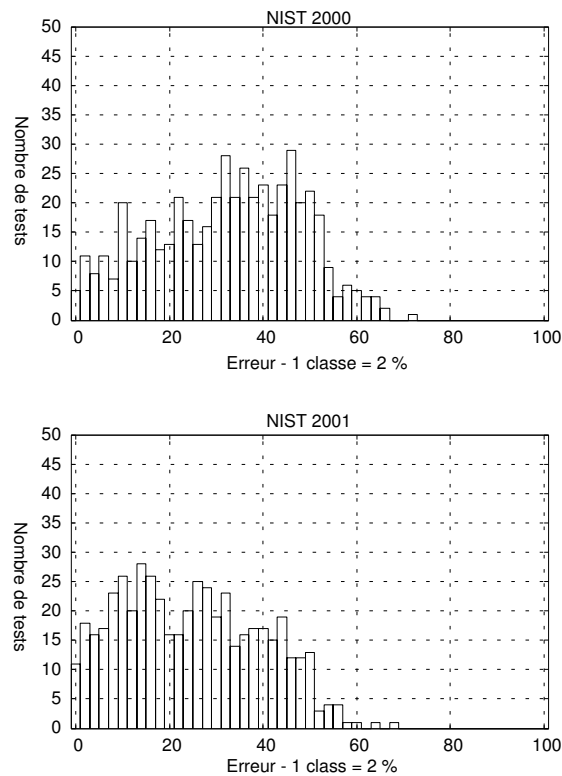


FIG. 14.2 – Histogramme des taux d’erreurs d’affectation ($E1$) par tests pour le corpus *CallHome* lors des évaluations NIST *n-segmentation* en 2000 et 2001.

tableaux 14.5 et 14.6 représentent respectivement l’histogramme des erreurs par test pour les campagnes 2000 et 2001.

14.3.2 NIST 2002

La méthode de calcul des erreurs a changé en 2002. Cette méthode est décrite au paragraphe 13.5.2.

Les résultats obtenus sur les corpus *Broadcast News* et *Meeting* sont donnés dans le tableau 14.7. Sur les deux corpus, un grand nombre de locuteurs ne sont pas détectés ; le taux de locuteurs non détectés est supérieur à 15%. Pour le corpus *Meeting*, l’erreur la plus importante provient des erreurs d’affectation des segments aux locuteurs détectés : le taux d’erreurs d’assignation entre deux locuteurs détectés est de 31.8%. En revanche pour le corpus *Broadcast News* ce taux d’erreurs (10,7%) est inférieur au taux d’erreurs généré par les locuteurs non détectés sur la tâche *2-segmentation* (15.6%). Bien que les segments de silence signalés par NIST aient été retirés en fin de segmentation, les taux d’erreurs de non-détection d’un segment de parole et de non-détection d’un segment de silence ne sont pas nuls ! NIST n’a pas su expliquer le problème lors du colloque 2002. Ces deux erreurs cumulées ($E2_{FA} + E2_{FR}$) représentent environ 5% et 8% de taux d’erreurs pour le corpus *Meeting* et *Broadcast News*.

Le tableau 14.8 montre les résultats des deux précédents corpus évalués avec la méthode NIST 2000. La méthode d’évaluation NIST 2000 produit un score plus faible que la méthode NIST 2002. Cette différence est de l’ordre de 12% pour le corpus *Meeting* et de 10% pour le corpus *Broadcast News*. Elle provient des erreurs parole/silence et des zones où des locuteurs parlent en simultané.

$L_{X_k}^{hyp} - L_{X_k}^{ref}$	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
$L_{X_k}^{ref} = 1$	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 2$	0	0	0	0	0	0	0	208	57	20	11	3	2	0	1	0	0
$L_{X_k}^{ref} = 3$	0	0	0	0	0	0	53	33	20	10	8	3	2	3	1	1	0
$L_{X_k}^{ref} = 4$	0	0	0	0	0	8	14	7	4	4	2	2	1	0	0	0	0
$L_{X_k}^{ref} = 5$	0	0	0	0	5	3	0	2	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 6$	0	0	0	1	1	1	2	0	0	1	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 7$	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 8$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 9$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 10$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TAB. 14.5 – Répartition des locuteurs détectés sur le corpus CallHome NIST pour la tâche n-segmentation en 2000, $L_{X_k}^{hyp} - L_{X_k}^{ref}$ indique la différence entre le nombre de locuteurs détectés dans la segmentation "hypothèse" et le nombre de locuteurs réel de la référence. $L_{X_k}^{ref}$ indique le nombre de locuteurs réel.

$L_{X_k}^{hyp} - L_{X_k}^{ref}$	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
$L_{X_k}^{ref} = 1$	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 2$	0	0	0	0	0	0	0	127	124	41	8	2	0	0	0	0	0
$L_{X_k}^{ref} = 3$	0	0	0	0	0	0	27	57	36	12	1	1	0	0	0	0	0
$L_{X_k}^{ref} = 4$	0	0	0	0	0	0	27	7	5	3	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 5$	0	0	0	0	1	5	2	2	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 6$	0	0	0	0	3	1	1	1	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 7$	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 8$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 9$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$L_{X_k}^{ref} = 10$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TAB. 14.6 – Répartition des locuteurs détectés sur le corpus CallHome NIST pour la tâche n-segmentation en 2001. $L_{X_k}^{hyp} - L_{X_k}^{ref}$ indique la différence entre le nombre de locuteurs de la référence et le nombre de locuteurs détectés dans la segmentation "hypothèse". $L_{X_k}^{ref}$ indique le nombre de locuteurs dans la référence.

Types de taux d'erreurs	<i>Meeting</i>	<i>Broadcast News</i>
$E2_{FR}$: non-détection d'un segment de parole	4.8	6.5
$E2_{FA}$: non-détection d'un segment de silence	0,0	1.9
$E2_{FA\ spk}$: généré par les locuteurs non détectés	15.3	15.2
$E2_{FR\ spk}$: généré par des locuteurs inexistantes dans la segmentation de référence	1.0	4.2
$E2_{spk}$: assignation entre deux locuteurs détectés	31.8	10,7
E_2 : taux d'erreurs total	52.69	38.6

TAB. 14.7 – Taux d'erreurs NIST 2002 sur les corpus *Meeting* et *Broadcast News* : tâche *n-segmentation* NIST 2002, nombre de locuteurs maximum : 10.

Système	Total	Nombre de locuteurs ou taux d'erreurs								
		1	2	3	4	5	6	7	8	9
Nombre de tests	76	0	16	18	19	8	10	3	1	1
Corpus <i>Broadcast News</i>	28,73	-	11,93	22,97	36,18	38,48	39,63	28,44	39,97	57,25
Nombre de tests	166	0	0	0	68	30	68	0	0	0
Corpus <i>Meeting</i>	40,63	-	-	-	42,55	32,66	42,14	-	-	-

TAB. 14.8 – Résultats pour les corpus *Meeting* et *Broadcast News* : le taux d'erreurs est obtenu par la méthode NIST 2000 ($E1$, en %). Le nombre d'enregistrements en fonction du nombre de locuteurs est précisé. Attention : le score indiqué pour le corpus *Meeting* n'est pas représentatif. En moyenne, seulement 32 secondes des enregistrements sont utilisées dans le calcul des scores sur les 2 minutes des enregistrements.

Chapitre 15

Segmentation en locuteurs : *2-segmentation*

Les résultats aux évaluations NIST de 2001 et 2002 obtenus par notre système sur la tâche de segmentation en deux locuteurs sont donnés. Dans un premier temps, le corpus utilisé au cours de ces deux évaluations est décrit.

15.1 Corpus d'évaluation

15.1.1 NIST 2000 et 2001

La population des tests est composée de fichiers extraits de la base *Switchboard II phase II*. Ce corpus est composé de 1000 conversations téléphoniques filaires d'une durée de 60 secondes :

- 271 tests contiennent deux locuteurs hommes,
- 321 tests contiennent deux locuteurs femmes
- et 408 tests sont mixtes.

La figure 15.1 montre l'histogramme des durées des interventions pour le corpus.

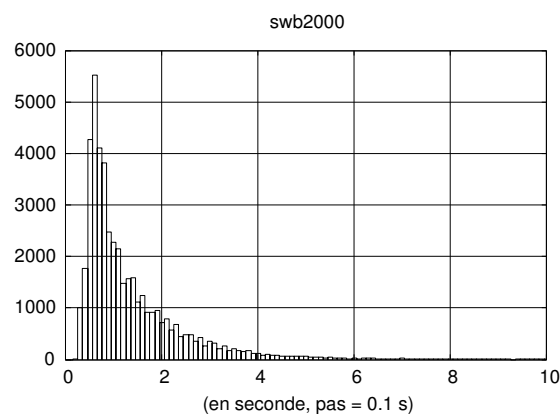


FIG. 15.1 – Histogramme des durées des interventions pour le corpus NIST 2-speaker en 2000 et 2001.

15.1.2 NIST 2002

Le corpus est composé de conversations téléphoniques cellulaires extraites de *Switchboard II phase III* [LDC 2002]. Le corpus contient 199 fichiers d'une durée moyenne de 120 secondes. La répartition des tests par genre n'a pas été fournie en 2002. La figure 15.2 illustre la durée des segments pour ce corpus.

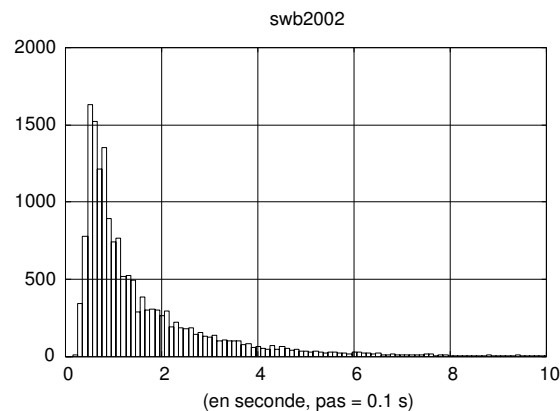


FIG. 15.2 – Histogramme des durées des interventions pour le corpus NIST 2-speaker en 2002.

Système	Corpus	2 hommes	2 femmes	homme/femme
Baseline	0,33	0,33	0,33	0,32
LIA 2000	0,31	0,32	0,33	0,29
LIA 2001	0,26	0,25	0,28	0,26
Meilleur	0,07	0,06	0,10	0,05
LIA 2002	0,10	-	-	-

TAB. 15.1 – Résultats de la tâche 2-segmentation aux évaluations NIST 2000 et 2001. Les résultats sont donnés avec la méthode d'évaluation NIST 2000 pour les systèmes du LIA de 2000 à 2002 et pour le système ayant obtenu les meilleurs résultats aux évaluations (Meilleur).

Système	Méthode d'évaluation NIST 2000	Méthode d'évaluation NIST 2002
Baseline	-	34,51
LIA	9,99	15,60
CLIPS	-	16,58
LIA+CLIPS	8,05	14,24

TAB. 15.2 – Résultats de la tâche 2-segmentation à l'évaluation NIST 2002. Les résultats sont donnés avec la méthode d'évaluation 2000 et 2002. Le système ayant obtenu les meilleurs résultats est le système LIA.

15.2 Résultats

15.2.1 NIST 2000 et 2001

Les résultats obtenus aux évaluations 2000 et 2001 sur la tâche 2-segmentation sont loin des résultats obtenus par le meilleur système (*c.f.* tableau 15.1). Le système utilisé est identique au système de la tâche *n-segmentation* hormis le nombre de locuteurs maximum fixé à deux. En collaboration avec Eurecom, le système 2000 a été développé en aveugle lors de l'évaluation, tandis que le système 2001 a été développé à partir du corpus de développement multi-locuteur *CallHome*.

15.2.2 Résultats NIST 2002

Le LIA, lors de l'évaluation 2002, a obtenu les meilleures performances pour la tâche 2-segmentation (15,60%). Ce résultat est amélioré en utilisant un système de fusion conjuguant les systèmes du CLIPS et du LIA [Moraru 2003]. Le système de fusion *LIA + CLIPS* obtient 14,24% d'erreurs. La méthode d'évaluation NIST 2002 génère environ 5% d'erreurs de plus que la méthode d'évaluation NIST 2000, la différence provenant des erreurs sur les segments où les locuteurs parlent en simultané.

15.2.3 Comparaison des corpus

Les deux corpus sont composés de conversations téléphoniques spontanées entre deux locuteurs, seul le type de canal de transmission changeant d'un corpus à l'autre. Les performances obtenues sont proches (environ 10% avec le système 2002 et la méthode d'évaluation NIST 2000). Par contre, une différence de performances importante a été constatée en VAL avec un système identique (*c.f.* annexe D). Les enregistrements cellulaires sont plus difficiles à traiter en VAL. Le système 2002, appliqué sur le corpus d'évaluation 2000, obtient des performances proches du meilleur système (LIA 2002 \sim 10%).

Chapitre 16

Conclusion

Les participations aux évaluations NIST 2001 et 2002 nous ont permis d'évaluer la méthode de segmentation en locuteurs proposée dans la partie II. Actuellement aucune tâche d'appariement en locuteurs n'est envisagée dans ces évaluations. La diversité des corpus a montré la difficulté de la tâche de segmentation, en particulier pour la segmentation des réunions (corpus *Meeting*).

Lors des évaluations NIST 2002, le système du LIA a obtenu le plus faible taux d'erreur pour la tâche de segmentation en deux locuteurs. Des problèmes restent encore non résolus ou mal solutionnés :

1. Les segments où plus d'un locuteur interviennent sont un problème qui n'a pas été abordé.
2. Aucun développement spécifique n'a été fait pour les deux corpus en bande large. Des gains de performances sont envisageables, en particulier en utilisant des modèles du monde spécifique aux corpus ainsi qu'une paramétrisation acoustique adaptée.
3. La détection du nombre de locuteurs reste l'enjeu majeur de la segmentation. Le système proposé est le seul système offrant une détection en automatique du nombre de locuteurs. OGI, le MIT et le CLIPS utilisent des systèmes qui fixent *a priori* le nombre de locuteurs à 2 ou 3. Bien que la solution proposée de détection du nombre de locuteurs ne soit pas parfaite, elle a apporté un gain de performance sur deux des trois corpus multi-locuteurs (méthode d'évaluation NIST 2001).

Cinquième partie

Conclusions et perspectives

Chapitre 17

Conclusions et perspectives

Ce travail de thèse s'inscrit dans le cadre de l'indexation en locuteurs de documents sonores. L'indexation en locuteurs consiste à déterminer automatiquement le nombre de locuteurs et à spécifier les interventions de chaque locuteur. L'indexation, abordée en vue d'une recherche d'information, est un thème dont les enjeux applicatifs sont de plus en plus importants avec l'accroissement de la quantité des documents numériques disponibles sur le Web ou sur les média numériques (télévisions et radios). L'accès efficace à l'information suppose que les collections de documents soient archivées, classées et décrites. Les travaux présentés dans cette thèse proposent des méthodes pour l'indexation des documents sonores en vue d'une recherche suivant les intervenants présents dans les documents. Deux thèmes sont abordés : la segmentation en locuteurs d'un document sonore et l'appariement en locuteurs d'une collection de documents sonores.

17.1 Conclusions

Segmentation en locuteurs

La segmentation en locuteurs d'un document sonore consiste à déterminer le nombre de locuteurs et à détecter les segments de chaque locuteur. Trois points demeurent problématiques :

- La détection du nombre de locuteurs de façon automatique butte sur le problème plus général de la détection du nombre de classes lors d'une classification en aveugle. Les méthodes usuelles n'obtiennent pas de résultats satisfaisants et ce malgré l'introduction d'heuristiques ou de facteurs correcteurs.
- L'évaluation des performances est toujours sujet à critique malgré les efforts de NIST pour définir une mesure de référence. Les différents critères à évaluer restent très difficiles à synthétiser. Les méthodes d'évaluation doivent satisfaire deux critères antinomiques :
 - La mesure doit être synthétique (un seul score) pour comparer aisément les performances des méthodes.
 - Les différents types d'erreurs doivent pouvoir être analysés finement. Par exemple, la méthode actuelle n'intègre qu'indirectement la détection du nombre de locuteurs. Lors des évaluations NIST, il est apparu préférable, pour minimiser le taux d'erreur, de fixer *a priori* (à deux) le nombre de locuteurs au lieu de chercher à détecter celui-ci. La méthode d'évaluation repose uniquement sur un taux d'erreur d'affectation mesuré en centiseconde. Par conséquent, les systèmes gagnent à détecter correctement le ou les locuteurs majoritaires quitte à ne pas trouver les locuteurs s'exprimant peu. Le problème peut être résolu en fixant des coûts plus élevés pour les locuteurs non détectés. Mais il demeure le choix des

critères de qualité d'une segmentation. Les critères fixés par NIST ne correspondent pas nécessairement aux critères pour les applications visées.

- La détection des segments dans lesquels plusieurs locuteurs s'expriment simultanément n'est pas abordée par les méthodes proposées dans la littérature, incluant la nôtre. Les méthodes supposent que les locuteurs ne parlent pas en même temps ou qu'ils ne se coupent pas la parole. Cette hypothèse semble peu réaliste comme le montre le corpus *Meeting* des évaluations NIST 2002 (les locuteurs parlent en simultané pendant 75% du temps).

La majorité des travaux consacrés à la segmentation en locuteurs repose sur un enchaînement de modules dont les deux principaux sont la détection de ruptures et la classification hiérarchique ascendante. Nous avons proposé une approche différente, s'apparentant à une méthode de classification descendante, qui permet de pallier les problèmes des approches classiques. Dans un même processus, les ruptures, les segments et les locuteurs (leurs modèles) sont remis en cause itérativement jusqu'à l'obtention de la solution optimale :

- La conversation est modélisée par un modèle de Markov caché (HMM) évolutif dans lequel les états correspondent aux locuteurs et les transitions modélisent les changements de locuteurs.
- A chaque itération du processus, un nouveau locuteur est détecté et ajouté au modèle HMM dont les paramètres sont réévalués (d'où le qualificatif d'"évolutif").

La méthode a été développée à l'aide d'un corpus issu des évaluations NIST, puis validée lors des évaluations 2001 et 2002. Les expériences réalisées ont montré l'intérêt d'un modèle de conversation reposant sur un HMM (un gain de 3,45% en absolu et de 24% en relatif est obtenu grâce aux contraintes portées par les transitions du HMM). De plus, l'approche segmente les documents en un temps de calcul réduit (en dessous du temps réel).

Lors des évaluations NIST 2001 et 2002, les résultats de l'approche proposée sont comparables aux méthodes concurrentes pour les tâches de segmentation en n locuteurs, bien que la méthode soit pénalisée par l'absence de données complémentaires pour l'apprentissage du modèle du monde, *i.e.* un modèle générique de locuteurs.

Durant la campagne d'évaluation NIST 2002, notre système a obtenu les meilleures performances pour la tâche de segmentation en deux locuteurs et ce avec un temps de calcul de 0,3 fois le temps réel¹.

Appariement en locuteurs

L'appariement automatique en locuteurs d'une collection de documents préalablement segmentés consiste à déterminer le nombre de locuteurs de la collection et à retrouver les interventions de chaque locuteur dans l'ensemble des documents. Peu de travaux traitent de ce sujet bien que cette tâche soit nécessaire dans la majorité des applications pratiques d'indexation en locuteurs.

Dans ce travail, la méthode d'appariement employée repose sur une méthode de classification hiérarchique, dans laquelle nous avons introduit les connaissances apportées par la segmentation des documents. Notre contribution porte sur deux points :

- Deux nouvelles mesures ont été proposées, elles permettent lors du calcul des mesures de dissimilarité entre les classes de tirer profit de toutes les interventions contenues dans les documents segmentés. Aucun gain de performance significatif n'a été relevé pour la tâche visée, cependant ces mesures ont montré un intérêt lors des expériences de discrimination.
- Une nouvelle méthode d'élagage pour la détermination du nombre de locuteurs a été proposée. Contrairement à l'approche classique, qui nécessite de couper l'arbre avant d'entreprendre

¹Les autres systèmes présentés sont entre 1 et 2 fois le temps réel.

l'élagage, cette méthode détecte directement le nombre de locuteurs sans engendrer de perte de performances.

17.2 Perspectives

Segmentation en locuteurs

Les recherches actuelles en segmentation de locuteurs ne proposent des solutions obtenant des performances satisfaisantes pour la détection des locuteurs que lorsque le nombre de locuteurs est un paramètre connu du système. L'approche proposée reposant sur un HMM évolutif a montré un potentiel intéressant pour explorer ce problème, notamment grâce au respect du maximum de vraisemblance tout au long du procédé. De nouvelles investigations concernant les critères d'arrêt et la sélection des données initiales des modèles doivent permettre d'améliorer ces points.

Avec la méthode proposée, le cas des locuteurs s'exprimant simultanément semble plus facile à prendre en compte qu'avec les méthodes concurrentes. Par exemple, ce phénomène peut être abordé en conservant les chemins alternatifs lors de la segmentation ; une autre voie à explorer concerne les méthodes de séparation de sources.

Enfin, l'approche proposée peut être améliorée à court terme sur deux points particuliers :

- Il serait préférable que les émissions évoluent dans un espace probabiliste comme les transitions. Une solution serait de mettre en œuvre une méthode de type WORD+MAP [Fredouille 1999, Fredouille 2000b] permettant de normaliser les scores à partir d'une distribution *a posteriori* des vraisemblances. Celle-ci pourrait être combinée avec les approches *D-Norm* [Ben 2002] qui permettent de s'affranchir du besoin de données supplémentaires pour estimer les paramètres de la normalisation.
- Les modèles de Markov fournissent un cadre théorique qui a fait ses preuves dans des domaines variés allant de la reconnaissance de la parole aux codes correcteurs. Il existe cependant des limitations dans les HMM conventionnels. En particulier, le modèle suppose que la probabilité de la durée de passage dans un état suive une loi exponentielle décroissante. Les probabilités de transition dépendent seulement des états d'origine et de destination ; chaque observation dépend uniquement de l'état qui l'a générée, *i.e.* le modèle ne tient pas compte des observations voisines. De nombreuses recherches dans le domaine de la reconnaissance de la parole ont été menées pour remédier au problème [Crystal 1982, Levinson 1986, Russel 1985]. Les solutions proposées n'ont pas significativement augmenté la précision de la reconnaissance dans les applications pratiques. Pour les appliquer à la méthode de segmentation présentée ici, elles devront faire l'objet d'une étude dans le but d'adapter au mieux la loi de durée au problème traité.

Appariement en locuteurs

Ce travail étant l'un des premiers sur le thème de l'appariement en locuteurs de documents segmentés, il reste bien entendu de nombreuses voies à explorer :

- Pour l'évaluation des performances, la méthode présentée consiste à mesurer l'erreur de classification au niveau des interventions sans tenir compte de la durée de celles-ci. Une solution permettant d'intégrer cette durée consiste à utiliser les méthodes d'évaluation issues de la tâche de segmentation, cependant la complexité algorithmique peut être un problème (l'évaluation reposant sur un algorithme *a priori* non polynomial). Bien entendu, comme pour la segmentation, les critères à évaluer restent à définir. . .

- La normalisation des mesures de dissimilarité à partir des valeurs de la matrice de dissimilarité est à explorer. La normalisation pourrait s’inspirer des méthodes comme *H-norm*, *T-norm*, *WORD+MAP* ou *D-norm* ([Gravier 2000], [Auckenthaler 2000], [Fredouille 1999], et [Ben 2002]).
- Bien qu’il s’agisse davantage d’une suite logique, il sera nécessaire d’évaluer l’appariement à partir des segmentations obtenues avec la méthode proposée.

Indexation de documents

• Dans cette thèse, nous avons étudié l’indexation uniquement pour des documents sonores et suivant une seule caractéristique (les locuteurs). En vue d’une intégration dans un système de recherche documentaire, notre système pourra être étendu :

- à des caractéristiques supplémentaires, par exemple la musique ;
- ou à de nouvelles informations sur les locuteurs (*c.f.* § 1.5).

L’introduction d’informations supplémentaires devrait renforcer la qualité de la segmentation et de l’appariement des documents. Cependant, la détection de nouvelles caractéristiques, différentes des locuteurs, complique la tâche d’indexation.

• Un autre enjeu consiste à intégrer les paroles prononcées par les locuteurs dans l’indexation en locuteurs. Jusqu’à présent, la segmentation est utilisée en amont des tâches de reconnaissance de la parole (*Broadcast news*, [NIST 2000]) pour l’adaptation des modèles de parole aux locuteurs. Cependant, les récents travaux du groupe *SuperSID* de l’école d’été de l’Université John Hopkins montre l’intérêt de l’intégration d’informations complémentaires, en particulier textuelles, pour la vérification du locuteur [Reynolds 2002]. Il est envisageable d’adapter les techniques proposées par le groupe *SuperSID* à l’indexation en locuteurs pour la segmentation de longs documents comme pour l’appariement.

• Les segmentations et les partitions produites par les deux tâches abordées dans ce travail ont été uniquement évaluées en fonction des segmentations et des partitions de référence fournies avec les corpus d’évaluation. Les méthodes d’indexation proposées devront être intégrées dans une application de recherche documentaire pour évaluer la qualité des réponses apportées aux requêtes des utilisateurs en termes de précision et de rappel. Il reste, bien entendu, à mettre en place des index et un moteur de recherche adaptés aux documents traités, ainsi qu’à définir les modalités d’interrogation suivant les locuteurs (identité des locuteurs, exemples de voix).

Si jusqu’à présent le temps de calcul de l’indexation n’était pas une contrainte majeure, l’indexation étant réalisée *a priori*. Toutefois il devient une contrainte dans un système d’interrogation, la réponse doit être rendue à l’utilisateur en un temps raisonnable, en plus d’être de qualité.

• Au vu de la nature des documents rencontrés (films et émissions télévisées) sur le Web, un système multimodal reposant sur le principe du modèle HMM et associant l’image au son représente une piste intéressante à explorer.

Annexes

Annexe A

Vérification du locuteur, NIST 2000 à 2002

Un résumé des systèmes de vérification du locuteur présentés aux évaluations NIST de 2000 à 2002 est présenté dans cette annexe. Le tableau A.1 montre les paramètres des systèmes, la figure A.1 illustre les résultats obtenus.

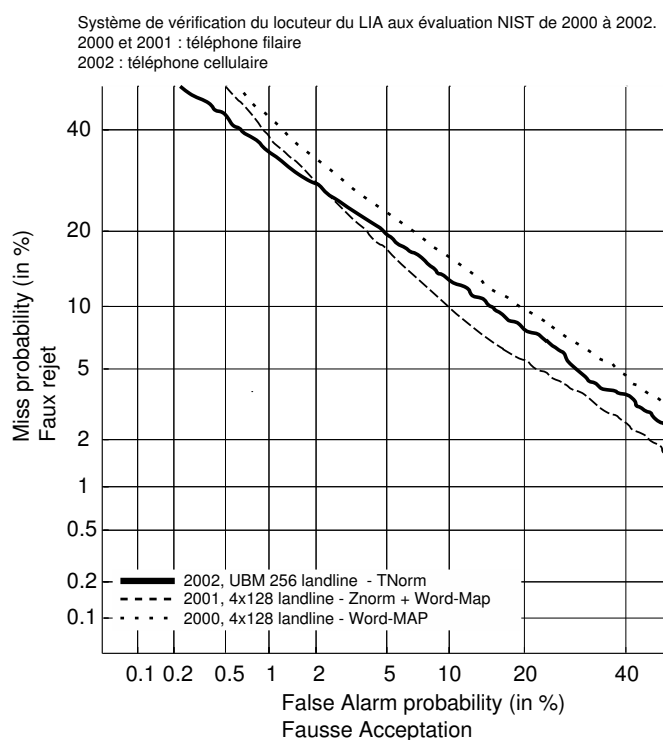


FIG. A.1 – Résultats NIST de 2000 à 2002

Module	Paramètre	2000	2001	2002
Corpus		filaire	filaire	cellulaire
Paramétrisation	Type	LFCC	LFCC	LFCC
	Nb Coef. statiques	16	16	16
	Delta	oui	oui	oui
	Limitation de bande passante (300-3400Hz)	oui	oui	oui
	CMS	oui	oui	non
	Suppression de trames	non	oui	oui
	Normalisation	non	non	oui en fin de processus
Modèle du monde	Nb de composantes	128	128	128
	Dependant du genre	oui	oui	oui
	Dependant du micro	oui	oui	non
	Nb de modèles	4 indépendant	4 indépendant	2 UBM
	Algorithme	EM-ML	EM-ML	EM-ML
	Suppression de trames	oui	oui	non
	Contrôle de la variance	non	non	oui
Modèle de locuteur	Nb de composantes	128	128	256
	Adaptation	MAP	MAP	MAP
	Durée d'apprentissage	2m	2m	2m
	Nb de sessions	1	1	1
Test	Durée	15-45s	15-45s	15-45s
Score	Type	SWGMM	lr()	lr()
	Normalisation	WMAP	ZNorm + WMAP	TNorm

TAB. A.1 – Paramètres des systèmes de vérification du locuteur aux évaluations NIST de 2000 à 2002.

Annexe B

Plate-forme de développement ELISA

B.1 Consortium ELISA

”Le LIA est membre fondateur du consortium ELISA. Ce consortium a été créé par l’ENST, le LIA et l’IRISA, auxquels se sont joints différents laboratoires, comme l’EPFL, l’IDIAP et l’ERM. ELISA a pour objectifs de faciliter les recherches coopératives en reconnaissance du locuteur. Le consortium organise des réunions régulières, maintient une plate-forme logicielle pour la reconnaissance du locuteur et aide ses membres à se présenter aux évaluations internationales, comme les campagnes NIST, en fournissant logiciels et aide technique et scientifique. La majorité des laboratoires du consortium participent conjointement aux évaluations NIST et présentent leur travaux dans des publications communes [ELISA 2000, Gravier 1999].

Enfin, en dehors des participants fondateurs, ELISA est un consortium à contours variables, différents laboratoires rejoignant celui-ci pour une action précise à durée limitée. Cette structure souple est facilitée par la dernière caractéristique du consortium : ELISA est auto-financé par ses participants.”[Bonastre 2000].

B.2 Plate-forme pour la tâche de vérification du locuteur

La plate-forme ELISA a été développée pour faciliter la mise en place des expériences de vérification du locuteur dans le cadre des évaluations NIST. Elle permet aussi bien d’assurer la phase de développement des systèmes de vérification que de faciliter la mise en place des évaluations. Cette plate-forme est composée de trois parties :

- Un PEC (Plan d’évaluation Commun) qui définit les ensembles de données de développement, les données d’évaluations étant définies par NIST. Chaque année, le consortium met en place un nouveau PEC à partir des données de l’évaluation NIST précédente.
- Un ensemble de logiciels¹ :
 - Sphere est une bibliothèque mise à disposition par NIST pour la manipulation des fichiers sons au format SPHERE.
 - SPRO pour la paramétrisation : ces logiciels sont développés par Guillaume Gravier (IRISA).

¹Les logiciels cités sont ceux inclus dans la plate-forme de 2000 à 2002

- AMIRAL pour la modélisation, le calcul des vraisemblances, la normalisation... Ces programmes sont développés par le LIA depuis 1998 (Teva Merlin, Jean-François Bonastre, Corinne Fredouille, Sylvain Meignier).
- SilRemove, un module de suppression de trames est proposé à l'origine par Ivan Magrin-Chagnolleau (DDL); le développement de ce module a depuis été poursuivi par l'équipe du LIA, intégrant différents algorithmes de suppression de trames.
- DetCurve est un module de calcul des courbes DET provenant de NIST.
- Un ensemble de scripts permettant d'interconnecter les différentes phases d'une expérience (*c.f.* figure B.1). Ces modules ont été développés par le LIA avec la contribution de l'IRISA pour la partie normalisation.

Depuis 2001, une nouvelle plate-forme est développée et maintenue par le LIA. Dans cette plate-forme, une expérience se résume à un script enchaînant les différentes tâches et à un ensemble de fichiers de configuration. La plate-forme a l'avantage de prendre en charge les expériences de leur définition jusqu'à la génération des compte-rendus d'expérience. Les interventions de l'utilisateur se limitent à la définition des paramètres de l'expérience et à l'ajout ou au remplacement d'un des modules de la chaîne opératoire.

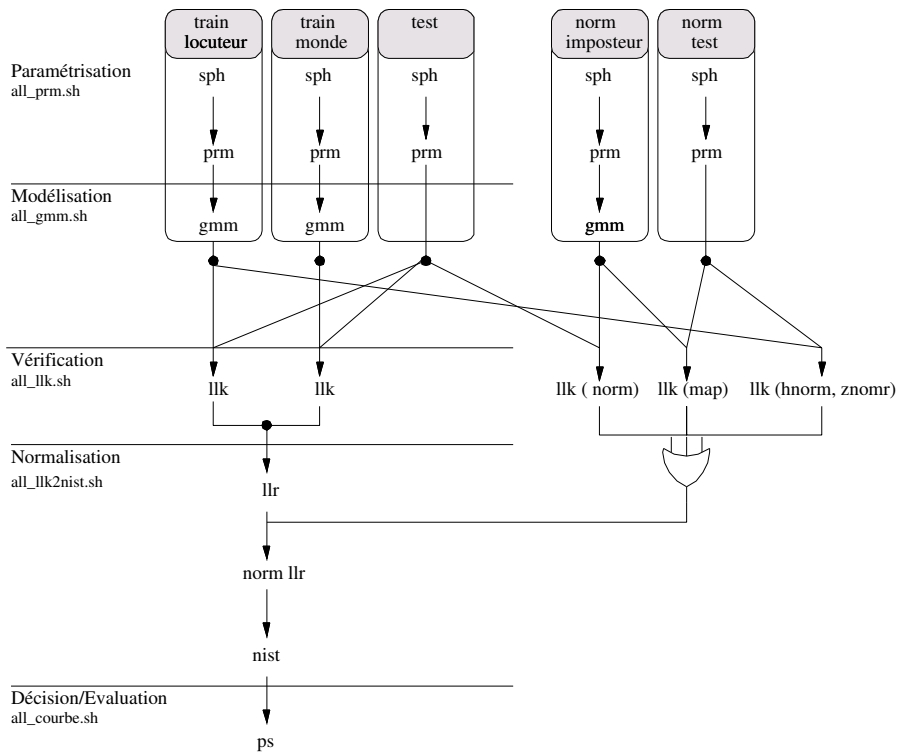


FIG. B.1 – Vue générale de la plate-forme de vérification du locuteur

B.3 Plateforme pour la tâche de segmentation en locuteurs

Sur le même principe que la plate-forme de vérification du locuteur, une plate-forme destinée aux évaluations de segmentation en locuteurs a été développée à partir de 2001. La complexité des expériences de segmentation est moins importante que pour la vérification du locuteur (*c.f.* figure

B.2). Cette plate-forme et le PEC associé pour la tâche de segmentation ont été définis par le LIA en 2001 et 2002.

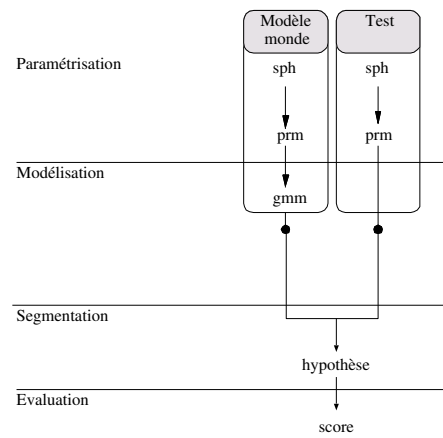


FIG. B.2 – *Vue générale de la plate-forme de segmentation en locuteurs*

Annexe C

Système LIA 2000 et 2001 de segmentation en locuteurs

C.1 Système développé par le LIA pour NIST 2001

Le système proposé en 2000 résulte de la collaboration du LIA avec l'Institut Eurecom. Ces travaux ont été réalisés par Teva Merlin (LIA) et Perrine Delacourt (Eurecom). Ce système a été développé en aveugle (aucun test préliminaire n'a été fait) lors des évaluations. Le processus de segmentation est réalisé en deux étapes :

- Une détection de rupture permet d'obtenir des segments de parole contenant un seul locuteur. La détection calcule une distance entre deux fenêtres glissantes sur le signal. Les frontières entre les segments sont déterminées par un seuil sur la suite des distances ([Delacourt 2000b]).
- Un système de détection des locuteurs repose sur les modèles GMM. Un processus itératif sélectionne le plus long segment trouvé dans la phase précédente. Un modèle de locuteur est appris sur ce segment. Enfin, tous les segments, dont le rapport de vraisemblance pour ce modèle est supérieur à un seuil, sont associés au locuteur. L'algorithme recherche un nouveau locuteur s'il reste des segments non attribués à un locuteur.

C.2 Système développé par le LIA pour NIST 2001

Le système proposé en 2001 est proche du système exposé au paragraphe 4. Le modèle de conversation est identique : la conversation est modélisée par un HMM où les états représentent les locuteurs et les transitions les changements de locuteurs. Seul quelques paramètres diffèrent :

- Les vecteurs acoustiques sont composés de 16LFCC complétés par les Delta.
- Les modèles de locuteur sont initialisés à partir d'un modèle du monde et les moyennes des modèles de locuteur sont appris par EM-ML.
- Le critère d'arrêt de la convergence des modèles utilise la probabilité du chemin de décodage.
- Le critère d'arrêt pour l'ajout des modèles utilise la probabilité du chemin de décodage, si le locuteur n'est associé qu'à un seuil segment d'une durée inférieure à 3 secondes alors le processus d'ajout de locuteur s'arrête.

Les paramètres de ce système ont été fixés sur le corpus de développement *CallHome* composé de 56 enregistrements différents des enregistrements d'évaluation.

Annexe D

Descriptif des systèmes proposés lors des évaluations NIST 2002

NIST 2002 SPEAKER RECOGNITION EVALUATION : LIA RESULTS

Sylvain Meignier, Teva Merlin, Raphaël Blouet, Jean-François Bonastre*

LIA / CERI

Université d'Avignon - Agroparc - BP 1228 - 84911 Avignon Cedex 9 - France

{sylvain.meignier, teva.merlin, raphael.blouet, jean-francois.bonastre}@lia.univ-avignon.fr

Contents

1	LIA "One speaker" system	2
1.1	Overview	2
1.1.1	Parameterization	2
1.1.2	Modeling	2
1.1.3	Scoring and normalization	2
1.2	Development results	2
1.3	System specifics	2
1.3.1	LIA_1 Primary system	2
1.3.2	LIA_2	2
1.3.3	LIA_3	2
1.3.4	LIA_4	2
1.3.5	Results	2
2	LIA Segmentation system	4
2.1	Overview	4
2.1.1	Parameterization	4
2.1.2	Background model	4
2.1.3	Speaker models	4
2.1.4	Scoring	4
2.1.5	Segmentation system	4
2.2	System specifics	6
2.2.1	Primary system LIA_1_seg	6
2.2.2	Secondary system LIA_2_seg	6
2.3	Development results	6
2.3.1	Development corpus	6
2.3.2	Baseline system	6
2.3.3	Parameterization	6
2.3.4	Transition weights in the Markov Model	6
2.3.5	Alpha parameter for MAP	7
2.4	NIST 2002 results	7
2.4.1	Switchboard results	7
2.4.2	Meeting & Broadcast results	7
3	LIA 2sp segmentation system	7
4	LIA 2sp Detection system	8
4.1	Overview	8
4.1.1	Split of the samples	8
4.1.2	Speaker matching	8
4.1.3	Scoring	8
4.2	Development results	8
4.3	System specifics	8
4.3.1	Primary system LIA_1_2sp	8
4.3.2	Secondary system LIA_2_2sp	9
4.3.3	NIST 2002 Results	9
5	References	9

* RAVOL project: financial support from Conseil général de la région Provence Alpes Côte d'Azur and DigiFrance.

1. LIA "ONE SPEAKER" SYSTEM

1.1. Overview

1.1.1. Parameterization

The signal is characterized as follows: 24 filter bank coefficients over 20ms Hamming-windowed frames are computed at a 10 ms frame rate. Bandwidth is limited to the 300-3400 Hz range. Filter bank coefficients are then converted to 16th order cepstral features using Discrete Cosine Transformation. Finally, the cepstral features are augmented by delta coefficients.

A frame removal algorithm is then applied on feature vectors to discard non significant frames. This algorithm is based on a bi-gaussian modeling of the energy distribution, learnt from non-zero energy frames using EM algorithm. The threshold used to determine which frames to discard is computed based only on the characteristics of the gaussian of high energy.

The final step consists in centering and reducing the distribution of all coefficients for the remaining frames.

1.1.2. Modeling

All models are based on GMMs, with diagonal covariance matrices.

Background model(s): Two kinds of background models are used, depending on the system:

- 128 components, gender-dependent models;
- 256 components Universal Background Models, derived from the gender-dependent models cited above.

A standard EM algorithm with ML criterion is used to learn the background models. Variance flooring is applied during the training so that variance for each gaussian is no less than 0.5 \times the global variance. The data used for background models come from the 99 NIST evaluation (landline telephone) and are composed of 171 male test segments and 268 female test segments (with a mean duration of 30 seconds). For some systems, new background models are computed using cellular data from the 2001 NIST evaluation (74 male test segments and 100 female test segments), using the above background models as initialization for the EM algorithm.

Speaker models: All speaker models are derived from the background model using a variant of MAP adaptation of the means. The relative weights of the background model and the estimation data result from a combination of the amount of data available in both cases and a priori weights (0.25 for the background model, 0.75 for the estimation data).

1.1.3. Scoring and normalization

Once log-likelihood scores are estimated at the frame level, each speech signal is split into fixed length temporal blocks (0.3s duration) within which an arithmetic mean is applied to merge log-likelihood score and to get block segmental scores. T-norm is applied on the block scores, which are then merged again (arithmetic mean) to provide decision scores. The impostor population required to compute the score distributions for T-norm is extracted from cellular data of 2001 evaluation campaign (74 male speakers and 100 female speakers). Finally, the decision scores are compared to a threshold, tuned to minimize DCF value on a development data set (2001 evaluation campaign).

1.2. Development results

1.3. System specifics

1.3.1. LIA_1 Primary system

This system relies on a 256 components Universal Background Model computed on landline data (from the NIST 1999 evaluation).

T-norm is applied on block scores as described in 1.1.3. The decision threshold is set at 1.5.

1.3.2. LIA_2

This system uses the same UBM as the primary system. No normalization is applied here.

The decision threshold is set at 0.45.

1.3.3. LIA_3

The background model is a 256 components UBM learnt on cellular data from the NIST 2001 evaluation, using the landline UBM as initialization for the EM algorithm.

T-norm is applied on block scores as described in 1.1.3. The decision threshold is set at 1.5.

1.3.4. LIA_4

This system uses two 128 components, gender-dependent background models computed on landline (NIST 99) data.

T-norm is applied on block scores as described in 1.1.3. The decision threshold is set at 1.5.

1.3.5. Results

Systems:

Figure 1 shows the results for all LIA 1-sp systems.

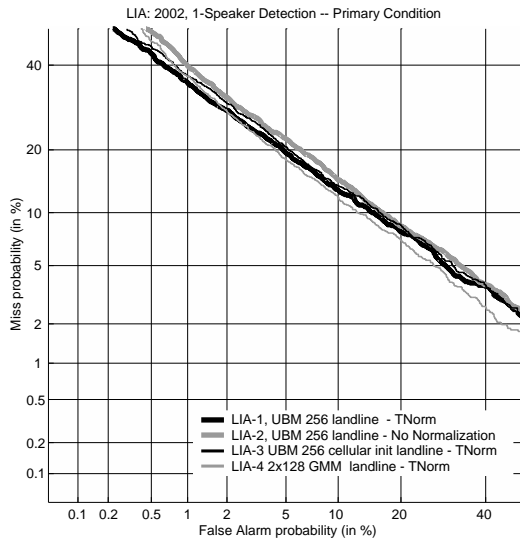


Fig. 1. 1-sp: Results by system on 2002 evaluation

It is interesting to notice the different behaviors between LIA_1 and LIA_4, which differ by the kind of background model they use (256 components UBM in one case, 2 gender-dependent, 128 components models in the other case) and the size of speaker models. The UBM-based system shows better performance around the *minimum DCF* point, while the other system appears to be more interesting at *EER*.

Also to be noticed is the comparable results obtained by LIA_1 (UBM computed on landline data) and LIA_3 (UBM computed on cellular data). The use of cellular data for the background model doesn't help here (though it is worth mentioning that the cellular UBM was computed with the landline UBM as initialization).

Gender:

Figure 2 shows results by gender for LIA_2 system (similar to the primary system but without normalization) and ELISA system (basically, LIA system for the 2001 evaluation) on both 2001, landline data, and 2002, cellular data.

For both systems, male and female results are comparable on 2001 data. On 2002 data, the results for females are slightly better than for males in the case of ELISA system, and significantly better in the case of

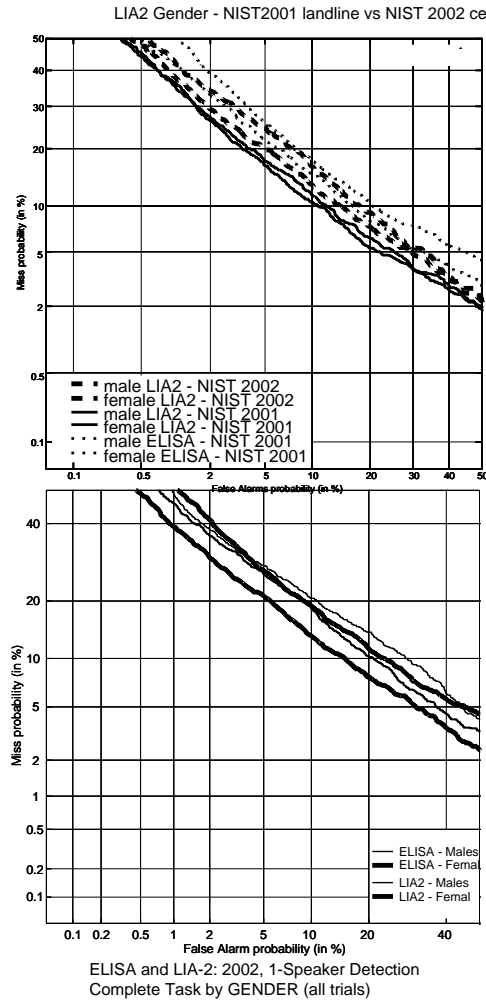


Fig. 2. 1-sp: Results by gender for current (LIA2) and last year (ELISA) systems, on cellular (2002) and landline (2001) data

LIA_2 system. It will be interesting to find out which factor leads to this behavior with the 2002 data.

2. LIA SEGMENTATION SYSTEM

2.1. Overview

2.1.1. Parameterization

The signal is characterized as follows: 24 filter bank coefficients over 20ms Hamming-windowed frames are computed at a 10 ms frame rate. Filter bank coefficients are then converted to 20th order cepstral features using Discrete Cosine Transformation. Finally, the cepstral features are augmented by the energy.

Unlike what is done for the lsp task, no frame removal or any coefficient normalization are applied here.

2.1.2. Background model

Background Modeling scheme is based on classical GMMs. Background model is characterized by 128 components, each summarized by a diagonal covariance matrix. This model is neither gender- nor handset-dependent. It is trained with a classical EM algorithm based on ML principle. At each iteration of the training algorithm, frames which are too "close" to a particular gaussian are discarded, such preventing variance from getting too low (but in a finer way than a mere variance flooring, which proved to be less efficient for this task).

Two background models are learnt from data sets defined as:

- 8kHz training data extracted from data of 1999 evaluation campaign (the same subset we used for ELISA system, smaller than the dataset used for our lsp systems).
- 16kHz training data extracted from the "Santa Barbara Corpus of Spoken American English, Part one" (LDC, 31 Jul 2000). 10 minutes are extracted from each of the 14 files. Each test is downsampled from 22kHz to 16kHz.

Cepstral Mean Subtraction is applied on training data before learning.

The 8kHz background model is used with Switchboard corpus. Whereas the 16kHz background model is used with Meeting and Broadcast News corpus.

2.1.3. Speaker models

Speaker models are derived from one of the two background models by applying MAP adaptation (means only are adapted). The adaptation technique used here is not the same one as for lsp. It relies on fixed, a priori weights (background weight = 0.4 and speaker weight = 0.6).

2.1.4. Scoring

For each speaker model, a likelihood ratio is computed for each fixed duration (0.3s) block.

2.1.5. Segmentation system

Segmentation using a Markov Model: The segmentation is modeled by a Markov Model [1]. During the segmentation, the Markov Model (MM) is generated by an iterative process, which detects and adds a new state (i.e. a new speaker) at each stage. Each MM state characterizes a speaker, and the transitions model the changes between speakers (figure 3).

Example: First speaker "L0": A first speaker "L0" is learnt on the whole test. The segmentation is modeled by a one-state MM and all the 0.3s blocks are set to speaker "L0".

Second speaker "L1":

- Selecting data for the new model "L1": A new speaker model is learnt on 3 seconds of test that maximize the sum of likelihood ratio for model "L0". A new state labeled "L1" is added to the MM. The 3 seconds of test are given to speaker "L1" in the segmentation (and removed from "L0").
- Adapting speaker models and computing the segmentation: First, the speaker models are adapted according to the segmentation data. Then, Viterbi decoding produces a new segmentation. Adaptation and decoding are performed while the segmentation differs between two successive steps.

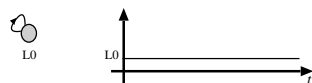
Next speakers "L2", "L3", ..., "LX": Next speakers are obtained the same way as speaker "L1". 3 seconds of test labeled "L0" are selected. A new state is added to the MM, the 3 seconds of test are labeled "LX". Model "LX" and the segmentation are computed.

When does the process stop ?

- If the Viterbi probability path decreased between two steps. The Viterbi probability path of the previous iteration is estimated again according to the new transitions probabilities. The transitions probabilities depend on the number of states in the MM (see below). The previous segmentation is kept (the last speaker is removed) and we stop.
- If more than the maximum number of speakers allowed is reached.
- If there is no more 3 seconds segment labeled "L0" left in the segmentation.

Step 1: adding speaker L0

Process initialisation



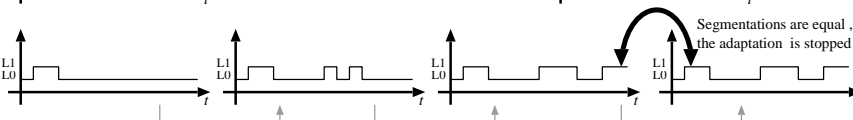
Step 2: adding speaker L1

Process : steps 1 & 2



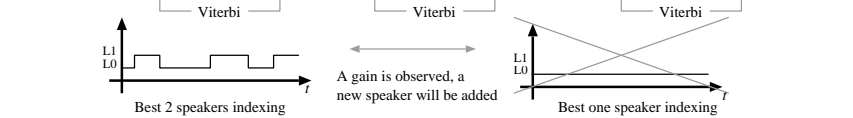
Process : step 3

Models Adaptation



Process : step 4

Stop criterion



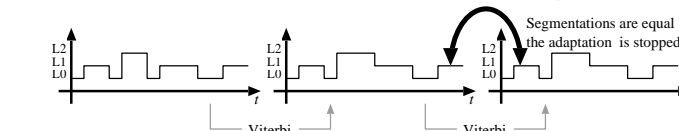
Step 3: adding speaker L2

Process : steps 1 & 2



Process : step 3

Models Adaptation



Process : step 4

Stop criterion

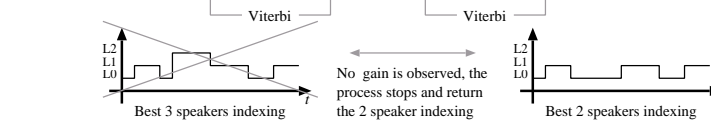


Fig. 3. Example of segmentation

- Dirty trick: If only one segment is labeled as the last speaker. The previous segmentation is kept and we stop.

At the end of the process, only the segments labeled as speech by NIST Speech Activity Detector are kept.

System Parameters:

- The transition probabilities of the MM are:
 - the probability of staying in the same state is "0.6";
 - the probability of switching to the next step is equal to "0.4/(n-1)", where n is the number of states.
- The maximum number of speakers allowed is set to 2 or 10, depending on the corpus.

2.2. System specifics

2.2.1. Primary system LIA_1_seg

Switchboard: the maximum number of speakers allowed is set to 2.

Meeting and Broadcast News: the maximum number of speakers allowed is set to 10.

2.2.2. Secondary system LIA_2_seg

Switchboard: the maximum number of speakers allowed is set to 10.

Meeting and Broadcast News: the maximum number of speakers allowed is set to 2.

2.3. Development results

2.3.1. Development corpus

The corpus used for development is composed of:

- 100 files drawn from *Switchboard* 2001 2sp evaluation;
- 62 English files extracted from *CallHome* corpus (evaluation and development).

2.3.2. Baseline system

Tables 1 and 2 show the results obtained with the system described in 2.1.5. All the development was done with the maximum number of speakers allowed set to 2. Error rates were computed with the **NIST 2001 scoring**. The baseline system obtained 10.79% of error with a majority of scores lower than 10% (fig. 4). The values given to the various parameters (MAP, parametrization, HMM) for the NIST 2002 system were selected according to the results obtained on the *Switchboard* corpus.

6326.23	Total time in a state to be segmented (in seconds)
5643.38	Time in a correctly segmented state (in seconds)
682.85	Time in an incorrectly segmented state (in seconds)
10.79	Segmentation error

Table 1. Scoring results on development corpus.

Corpus	Error	number of tests #
Switchboard	9.95	100
CallHome	11.57	62
All	10.79	162

Table 2. Error rates by corpus, on development corpus.

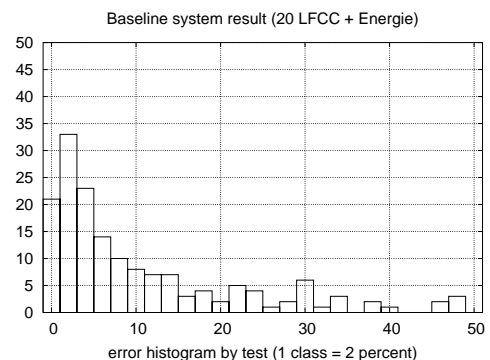


Fig. 4. Histogram of segmentation error: number of files (y axis) for each percentage of error (x axis), on development corpus

2.3.3. Parameterization

Table 3 shows the segmentation error for various kinds of parameterization. In the case of *LFCC+Energy*, the error rate decreases as the number of LFCC coefficients increases. Energy helps to segment the signal if speakers or their records have different volumes. *LFCC+Delta* coefficients obtain worse results than *LFCC+Energy* but a gain is observed between *16LFCC* and *16LFCC+Delta*.

2.3.4. Transition weights in the Markov Model

We recall that the choice of speaker is done every 0.3 second. Table 4 shows the influence of the transition weights. The best result is obtained with a weight of 0.6 for staying in the same state and 0.4 for switching to the next step. A lower weight favors small segments whereas higher weights tend to produce large segments.

Parameterization	Error
20LFCC	12.29
20LFCC+Energy	10.79
20LFCC+Delta	14.03
16LFCC	14.67
16LFCC+Energy	11.63
16LFCC+Delta	13.02
16LFCC+E+Delta	17.77

Table 3. Error rates on development corpus, for various kinds of parameterization.

b_{ii} transition weight	Switchboard	CallHome	Sum
Markov Model (0.3)	45.12	48.97	47.13
Markov Model (0.4)	31.22	38.78	35.17
Markov Model (0.5)	12.59	15.74	14.24
Markov Model (0.55)	9.98	12.84	11.47
Markov Model (0.6)	9.95	11.57	10.79
Markov Model (0.65)	10.43	11.81	11.15
Markov Model (0.7)	10.51	13.14	11.89
Markov Model (0.8)	13.78	13.05	13.40

Table 4. Error rates on development corpus, depending on the transition weights.

2.3.5. Alpha parameter for MAP

Table 5 shows the segmentation error for various values of the α parameter for MAP (which represents the weight of the background model in the adaptation process). The best MAP parameter value appears to be lower when the world data is closer to the corpus: the best parameter is 0.6 for *Switchboard* and 0.8 for *CallHome* and the world data was extracted from *Switchboard*.

2.4. NIST 2002 results

2.4.1. Switchboard results

Table 6 shows the results on Switchboard corpus according to the NIST 2002 scoring. The error score of

MAP parameter value	Switchboard	CallHome	Sum
0.1	20.10	23.34	21.79
0.2	15.49	19.86	17.77
0.3	12.18	17.21	14.81
0.4	12.40	13.14	12.79
0.5	10.33	12.66	11.54
0.6	9.95	11.57	10.79
0.7	10.96	10.48	10.71
0.8	11.29	10.06	10.65
0.9	12.58	11.04	11.77

Table 5. Error rates on development corpus, depending on the value of MAP α parameter.

Error	Switchboard
Missed speech	0.0
False alarm speech	0.0
Missed speaker	0.0
False Alarm speaker	0.0
Speaker error	15.6
<i>Total cost</i>	<i>15.6</i>

Table 6. Error rates on NIST 2002 *Switchboard* segmentation evaluation, system LIA1, max. number of speakers = 2 speakers.

Error	Meeting	Broadcast
Missed speech	4.8	6.5
False alarm speech	0.0	1.9
Missed speaker	15.3	15.2
False Alarm speaker	1.0	4.2
Speaker error	31.8	10.7
<i>Total cost</i>	<i>52.69</i>	<i>38.6</i>

Table 7. Error rates on NIST 2002 *Meeting & Broadcast News* segmentation evaluation, system LIA1, max. number of speakers = 10 speakers.

10.19% obtained on *Switchboard* corpus is close to the development results (10.79%) with the NIST 2001 scoring.

2.4.2. Meeting & Broadcast results

Table 7 shows the results on *Meeting* and *Broadcast News* corpus according to the NIST 2002 scoring.

Meeting and *Broadcast* have different types of errors:

- On both corpus, too many speakers are not detected (Missed Speaker $\sim 15\%$). The speaker detection is the main problem of blind segmentation and the proposed solution does not assess this problem very well.
- For the *Meeting* corpus, the most important error comes from the *Speaker Error* (31.8%). Whereas in the case of *Broadcast* corpus, this error (10.7%) is lower than the *Missed Speaker Error* and is even lower than the error obtained on *Switchboard* (15.6%).

3. LIA 2SP SEGMENTATION SYSTEM

The segmentation is processed using the segmentation system applied to Switchboard corpus (i.e. 8kHz background model, 20LFCC+E, the maximum number of speakers allowed is set to 2).

4. LIA 2SP DETECTION SYSTEM

The 2sp detection system combines the 1sp system and the segmentation system (fig. 5).

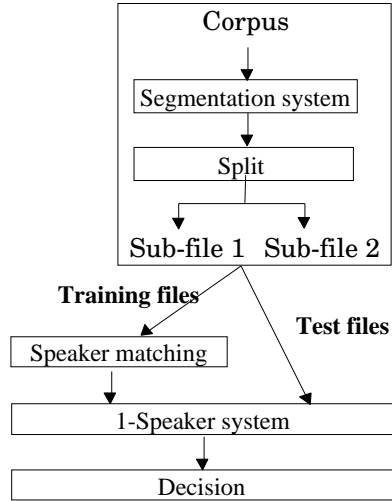


Fig. 5. 2-Speaker system overview

4.1. Overview

4.1.1. Split of the samples

The 1sp parameterization is performed on test data and train data (16LFCC). But delta coefficients, frame removal and frame normalization are not done yet. Each parameter file is split into at most two segments using the segmentation results (c.f. LIA 2sp Segmentation system). Frame removal and normalization are now applied on each sub-file.

4.1.2. Speaker matching

Speaker models are learnt on each train sub-file as a 1sp test by adapting the data to the 256 components UBM used for system LIA_1_1sp.

The speaker matching is a "clustering algorithm" based on the cross likelihood ratio (CLR). Figure 6 shows the two merging steps.

The cross likelihood ratio [2] is expressed in terms of similarity as:

$$d_{clr}(X_i, Y_j) = \frac{l(Y_j|\lambda(X_i))}{l(Y_j|\lambda(W))} \cdot \frac{l(X_i|\lambda(Y_j))}{l(X_i|\lambda(W))}$$

with:

X_i being the set of segments corresponding to speaker

i in the file X ;

$\lambda(X_i)$ being the segment-speaker model corresponding to data X_i ;

$\lambda(W)$ being the background model.

A similarity matrix between the segments is computed. Only CLR from models derived from different train files are computed (white squares in fig. 5). $d_{clr}(X_i, Y_i)$ are impossible (darkgrey squares in fig. 5). And $d_{clr}(X_i, Y_j)$ and $d_{clr}(X_j, Y_i)$ are equal then only one of them is computed (lightgrey squares in fig. 5).

The models with the best CLR score (highlighted square in fig. 5) are merged. A new speaker model is learnt from the merged speaker data.

A new similarity matrix is computed. Only two CLR are computed between the new model and the last two sub-files. The sub-file with the best score is kept.

Last, the speaker model is learnt from the 3 selected sub-files, using the same method as for 1sp systems.

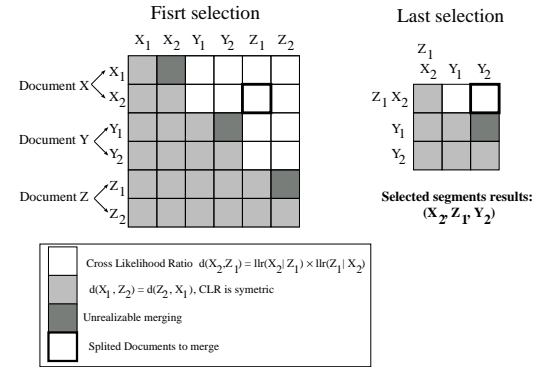


Fig. 6. Speaker matching

4.1.3. Scoring

1sp test is performed on each sub-test, the best score between the two sub-tests is kept.

4.2. Development results

No specific developments were done for this task. We only used the segmentation and 1-sp development results to select the parameters for 2-sp.

4.3. System specifics

4.3.1. Primary system LIA_1_2sp

2sp Detection system without normalization.

4.3.2. Secondary system LIA-2-2sp

Primary system followed by a T-norm (c.f. 1-sp).

4.3.3. NIST 2002 Results

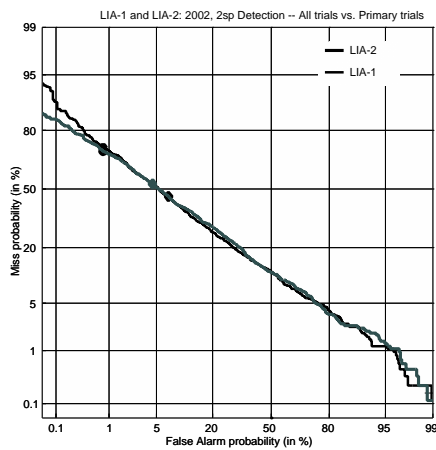


Fig. 7. 2-sp: Results for LIA1 & LIA2 systems on 2002 evaluation

5. REFERENCES

- [1] Meignier Sylvain, Jean-François Bonastre, and Stéphane Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *2001 : a Speaker Odyssey*, 2001, pp. 175–180.
- [2] D.A. Reynolds, E. Singer, B.A. Carlson, J.J. McLaughlin G.C. O'Leary, and M.A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proceedings of ICSLP 98*.

Annexe E

Bibliographie personnelle

Conférences internationales

- Moraru D., Meignier S., Besacier L. et Bonastre J., Combining experts for automatic speaker segmentation, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003)*, 2003, soumis.
- Meignier S., Bonastre J. et Magrin-Chagnolleau I., Speaker Utterances tying among speaker segmented audio documents using hierarchical classification : towards speaker indexing of audio databases, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2002)*, 2002.
- Meignier S., Bonastre J.-F. et Igounet S., E-HMM approach for learning and adapting sound models for speaker indexing, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 175–180, 2001.
- Magrin-Chagnolleau I., Gravier G. et Blouet R. and for the ELISA consortium, Overview of the ELISA consortium research activities, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 67–72, 2001.
- Meignier S., Bonastre J.-F., Fredouille C. et Merlin T., Evolutive HMM for speaker tracking system, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2000)*, 2000.

Conférences francophones

- Meignier S., Bonastre J. et Magrin-Chagnolleau I., Appariement de locuteurs entre des documents sonores préalablement segmentés en utilisant la classification hiérarchique, dans *XXIVèmes Journées d'Etudes sur la Parole (JEP)*, 2002.
- Meignier S., Bonastre J. et Igounet S., Modèle de markov évolutif pour les tâches de segmentation et d'indexation, dans *18ième Colloque GRETSI'01 sur le traitement du signal et des images*, 2001.
- Meignier S., Bonastre J., Fredouille C. et Merlin T., Modèle de markov évolutif pour les tâches de suivi de locuteurs, dans *XXIIIèmes Journées d'Etudes sur la Parole (JEP)*, 2000.
- Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T. et Wellekens C. J., Différentes stratégies pour le suivi du locuteur, dans *Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 123–129, 2000.

Glossaire

1-speaker : *1-speaker* est une tâche de vérification du locuteur proposée aux évaluations NIST de reconnaissance du locuteur.

2-segmentation : *2-segmentation* est une tâche de segmentation en deux locuteurs aux évaluations NIST de reconnaissance du locuteur.

2-speaker : *2-speaker* est une tâche de vérification du locuteur pour des conversations à deux locuteurs aux évaluations NIST de reconnaissance du locuteur.

n-segmentation : *n-segmentation* est une tâche de segmentation en plusieurs locuteurs aux évaluations NIST de reconnaissance du locuteur.

speaker tracking : *speaker tracking* est une tâche de poursuite de locuteur aux évaluations NIST de reconnaissance du locuteur.

Appariement en locuteurs : A partir d'une collection de documents, l'appariement en locuteurs détermine le nombre de locuteurs et leurs interventions dans les documents.

Caractéristique : Une caractéristique est une meta-information faisant l'objet d'une indexation.

Classification : Une classification est une partition d'un ensemble d'objets en groupes homogènes. Chaque groupe est identifié par un descripteur.

Index : Un index permet l'accès à des documents ou à des parties de documents en fonction d'un identifiant. Les caractéristiques servent d'identifiant dans l'index.

Indexation : L'indexation est un processus visant à la mise en place d'index.

Intervention : L'intervention d'un locuteur dans un document est l'ensemble des segments où il parle.

Précision : La précision en recherche documentaire est la quantité d'information pertinente trouvée, divisée par la quantité d'information trouvée.

Rappel : Le rappel en recherche documentaire est la quantité d'information pertinente trouvée, divisée par la quantité d'information pertinente à trouver.

Segment : Un segment correspond à une partie de document qui est définie par ses instants de début et de fin.

Segmentation : Une segmentation correspond à :

- la tâche consistant à trouver les segments correspondant à une caractéristique prédéfinie,
- au résultat retourné par un système de segmentation.

Segmentation de référence : Une segmentation de référence est une segmentation livrée avec un corpus ; cette segmentation permet d'évaluer une segmentation hypothèse.

Segmentation en locuteurs : La segmentation en locuteurs est une tâche dans laquelle le nombre de locuteurs est déterminé et où les segments de chaque locuteur (l'intervention d'un locuteur) sont spécifiés.

Segmentation hypothèse : Une segmentation hypothèse correspond au résultat donné par un système de segmentation en locuteurs.

Vérification du locuteur : La vérification du locuteur consiste à décider si deux échantillons de voix sont prononcés par le même locuteur.

Acronymes

EER *Equal Error Rate*, taux d'égale erreur

BBN Société Bolt, Beranek & Newman.

CMS *Cepstral Mean Subtraction*, soustraction de la moyenne cepstrale.

EM-ML *Expectation Maximization of Maximum Likelihood*, apprentissage d'un modèle par l'algorithme *Expectation Maximization* optimisant le critère du maximum de vraisemblance.

GMM *Gaussian Mixture Model*, modèle à mixture de gaussiennes.

HMM *Hidden Markov Model*, modèle de Markov caché.

LFCC *Linear Frequency Cepstral Coefficients*, coefficients cepstraux issus d'une analyse en banc de filtre à échelle linéaire.

MFCC *Mel Frequency Cepstral Coefficients*, coefficient cepstraux utilisant une échelle de Mel.

MAP *Maximum A Posteriori*, adaptation d'un modèle par la méthode du maximum *a posteriori*.

MIT *Massachusetts Institut of Technology*.

MLLR *Maximum Likelihood Linear Regression*, adaptation d'un modèle utilisant une régression linéaire.

NIST *National Institute of Standards and Technology*.

OGI *Oregon Graduate Institute*.

IAL Identification Automatique du Locuteur.

VAL Vérification Automatique du Locuteur.

RAL Reconnaissance Automatique du Locuteur.

RAP Reconnaissance Automatique de la Parole.

RT-02 *Rich Transcription 2002 evaluation*, campagne d'évaluation organisée par NIST pour la transcription et l'annotation de documents sonores.

W3C *World Wide Web Consortium*, consortium de l'Internet.

Table des figures

1.1	<i>Principe de l'indexation automatique.</i>	6
1.2	<i>Principe de l'indexation automatique en locuteurs.</i>	7
1.3	<i>Indexation en locuteurs : segmentation et appariement.</i>	8
3.1	<i>Principe général de la segmentation en locuteurs</i>	20
3.2	<i>Exemple de modèle de Markov à 3 états.</i>	25
3.3	<i>Exemple de segmentation en 5 locuteurs</i>	26
3.4	<i>Exemple de rupture.</i>	27
3.5	<i>Exemple de rupture sur les silences</i>	28
3.6	<i>Histogramme des durées suivant les types de décalages entre les segments</i>	28
3.7	<i>Exemple de détection des changements de caractéristique acoustique</i>	29
3.8	<i>Méthode de détection des changements de caractéristique acoustique</i>	29
3.9	<i>Classification hiérarchique : approche descendante v.s. approche ascendante.</i>	31
3.10	<i>Classification hiérarchique ascendante : regroupement en classes.</i>	32
3.11	<i>Classification hiérarchique : sélection de la partition finale par coupe horizontale ou par élagage.</i>	34
3.12	<i>Histogramme des durées des segments pour différents corpus.</i>	36
4.1	<i>Méthode proposée.</i>	40
4.2	<i>Comparaison entre les modules d'un système classique où l'information est propagée de module en module et la méthode proposée où l'information est remise en cause à chaque itération. En grisé : étapes optionnelles.</i>	41
4.3	<i>Exemple de segmentation en locuteurs d'un document sonore : initialisation du modèle de conversation.</i>	42
4.4	<i>Exemple de segmentation en locuteurs d'un document sonore : ajout de locuteurs.</i>	44
4.5	<i>Classification : différences entre l'approche proposée et de la méthode classique.</i>	45
4.6	<i>Exemple de segmentation en locuteurs d'un document sonore : itération 1, initialisation. Les zones grisées correspondent aux segments.</i>	46
4.7	<i>Exemple de segmentation en locuteurs d'un document sonore : initialisation du locuteur X_2. Les zones grisées correspondent aux segments.</i>	47
4.8	<i>Exemple de segmentation en locuteurs d'un document sonore : adaptation du locuteur X_2. Les zones grisées correspondent aux segments.</i>	48
4.9	<i>Exemple de segmentation en locuteurs d'un document sonore : initialisation du locuteur X_3. Les zones grisées correspondent aux segments.</i>	49
4.10	<i>Exemple de segmentation en locuteurs d'un document sonore : adaptation du locuteur X_3. Les zones grisées correspondent aux segments.</i>	50

4.11	<i>Exemple de segmentation en locuteurs d'un document sonore : la segmentation en deux locuteurs est conservée, le processus s'arrête. Les zones grisées correspondent aux segments.</i>	50
4.12	<i>HMM évolutif utilisant une stratégie ascendante ou descendante pour la sélection des locuteurs.</i>	52
5.1	<i>Méthode d'évaluation : rappel des étapes de la méthode de segmentation proposée.</i>	54
5.2	<i>Evaluation : prétraitement.</i>	56
5.3	<i>Résultats de segmentation en locuteurs pour différentes paramétrisations</i>	57
5.4	<i>Evaluation, ajout de locuteur et initialisation d'un locuteur</i>	58
5.5	<i>Durée d'émission dans un HMM</i>	60
5.6	<i>Taux d'erreur d'affectation observé pour différentes valeurs du paramètre γ régulant les probabilités de transition.</i>	61
5.7	<i>Evaluation, adaptation et segmentation : étape 2</i>	62
5.8	<i>Différentes valeurs du paramètre α pour l'adaptation MAP</i>	63
5.9	<i>Evaluation de la détection du nombre de locuteurs</i>	65
7.1	<i>Diagramme d'un système de segmentation en locuteurs.</i>	73
7.2	<i>Exemple d'appariement de locuteurs : les interventions, extraites des documents, sont regroupées par locuteur.</i>	74
8.1	<i>Appariement en locuteurs, méthode classique : classification hiérarchique.</i>	78
8.2	<i>Estimation de la dissimilarité $d(c_{uv}, c_w)$ pour les méthodes single linkage et complete linkage de la nouvelle classe uv et une autre classe w.</i>	81
8.3	<i>Appariement en locuteurs : élagage ou coupe du dendrogramme</i>	82
9.1	<i>Appariement en locuteurs : utilisation des interventions.</i>	84
9.2	<i>Appariement en locuteurs : réestimation de la matrice de dissimilarité après une fusion.</i>	86
10.1	<i>Courbes DET pour l'expérience en vérification du locuteur utilisant les mesures de similarité</i>	89
10.2	<i>Courbes du score e en fonction du nombre de classes</i>	92
12.1	<i>Tâche NIST 1-speaker : tâche de vérification du locuteur</i>	98
12.2	<i>Tâche NIST 2-speaker en 2000 et 2001.</i>	98
12.3	<i>Tâche NIST 2-speaker en 2002.</i>	99
13.1	<i>NIST 2000/2001, erreurs d'affectation : zones rejetées du calcul d'erreur.</i>	104
13.2	<i>NIST 2000/2001, erreur d'affectation : construction des paires de locuteurs.</i>	104
13.3	<i>NIST 2002, erreurs d'affectation : zones rejetées du calcul d'erreur.</i>	105
13.4	<i>Erreur NIST 2002 v.s. erreur 2000 & 2001</i>	106
14.1	<i>Histogrammes de durées des segments pour les corpus CallHome, Broadcast News et Meeting.</i>	110
14.2	<i>Histogramme des taux d'erreurs d'affectation ($E1$) par tests pour le corpus CallHome lors des évaluations NIST n-segmentation en 2000 et 2001.</i>	111
15.1	<i>Histogramme des durées des interventions pour le corpus NIST 2-speaker en 2000 et 2001.</i>	116
15.2	<i>Histogramme des durées des interventions pour le corpus NIST 2-speaker en 2002.</i>	116
A.1	<i>Résultats NIST de 2000 à 2002</i>	129
B.1	<i>Vue générale de la plate-forme de vérification du locuteur</i>	132

<i>TABLE DES FIGURES</i>	155
B.2 <i>Vue générale de la plate-forme de segmentation en locuteurs</i>	133

Liste des tableaux

3.1	<i>Comparaison du nombre de coefficients statiques pour les tâches de VAL et de segmentation</i>	22
3.2	<i>Comparaison des performances de systèmes de segmentation utilisant une classification hiérarchique ascendante</i>	35
5.1	<i>Les paramètres du système de référence.</i>	55
5.2	<i>Statistique sur la séquence permettant d'initialiser un nouveau locuteur</i>	59
5.3	<i>Probabilités de transition pour un modèle à deux états.</i>	61
5.4	<i>Probabilités de transition pour un modèle à trois états.</i>	61
5.5	<i>Critère d'arrêt du processus sur le corpus de développement</i>	67
5.6	<i>Critère d'arrêt du processus sur les corpus des évaluations NIST multi-locuteurs NIST</i>	67
10.1	<i>Nombre de locuteurs apparaissant dans 1 à 4 tests.</i>	88
10.2	<i>Méthode de réestimation des dissimilarités</i>	91
10.3	<i>Comparaison des différences mesures de dissimilarité</i>	91
12.1	<i>Résumé des différentes méthodes proposées à NIST de 2000 à 2002.</i>	100
12.2	<i>Comparaison des temps d'exécution des méthodes LIA et des méthodes CLIPS</i>	100
13.1	<i>Exemple de la répartition du nombre de locuteurs détectés sur le corpus CallHome</i>	103
14.1	<i>CallHome : Répartition des tests en fonction du nombre de locuteurs intervenant.</i>	108
14.2	<i>CallHome : Répartition des tests en fonction de la langue des intervenants.</i>	108
14.3	<i>Meeting et Broadcast News : Répartition des tests en fonction du nombre de locuteurs.</i>	109
14.4	<i>Résultat NIST "n-segmentation" 2000 et 2001 sur la base "CallHome"</i>	110
14.5	<i>Répartition des locuteurs détectés sur le corpus CallHome NIST pour la tâche n-segmentation en 2000</i>	112
14.6	<i>Répartition des locuteurs détectés sur le corpus CallHome NIST pour la tâche n-segmentation en 2001</i>	112
14.7	<i>Taux d'erreurs NIST 2002 sur les corpus Meeting et Broadcast News</i>	113
14.8	<i>Résultats pour les corpus Meeting et Broadcast News</i>	113
15.1	<i>Résultats de la tâche 2-segmentation aux évaluations NIST 2000 et 2001</i>	117
15.2	<i>Résultats de la tâche 2-segmentation à l'évaluation NIST 2002</i>	117
A.1	<i>Paramètres des systèmes de vérification du locuteur aux évaluations NIST de 2000 à 2002.</i>	130

Bibliographie

- [Adami 2002] Adami A., Kajarekar S. S. et Hermansky H., A new speaker change detection method for two-speaker segmentation, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*, tome IV, pages 3908–3911, 2002.
- [Auckenthaler 2000] Auckenthaler R., Carey M. et Lloyd-Thomas H., Score normalization for text-independent speaker verification system, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Backer 1975] Backer J., The dragon system-an overview, dans *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1975.
- [Banziger 2000] Banziger T., Klasmeier G., Johnstone T., Kamceva T. et Scherer K. R., Améliorer les systèmes de vérification automatique du locuteur en intégrant la variabilité émotionnelle : Méthodes et premières données, dans *XXIIIèmes Journées d'Etudes sur la Parole (JEP)*, pages 341–344, 2000.
- [Barras 1996] Barras C., *Reconnaissance de la parole continu : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés*, Thèse de doctorat, Université Paris VI, 1996.
- [Ben 2002] Ben M., Blouet R. et Bimbot F., A monte-carlo method for score normalization in automatic speaker verification using kullback-leiber distances, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2002)*, 2002.
- [Berrut 1997] Berrut C., *Indexation des données multimedia, utilisation dans le cadre d'un système de recherche d'informations*, Habilitation à diriger des recherches, Université Joseph Fourier, 1997.
- [Besacier 2000] Besacier L., Bonastre J.-F. et Fredouille C., Localization and selection of speaker specific information with statistical modelling, dans *Speech Communication*, tome 31, pages 89–106, 2000.
- [Bett 2000] Bett M., Gross R., Yu H., Zhu X., Yang J. et Waibel A., Multimodal meeting tracker, dans *Conference on Content-Based Multimedia Information Access, RIAO 2000*, 2000.
- [Bimbot 1995] Bimbot F., Magrin Chagnolleau I. et Mathan L., Second-order statistical measures for text-independent speaker identification, dans *Speech Communication*, tome 17(1-2), pages 177–192, 1995.
- [Bonastre 1994] Bonastre J.-F., *Stratégie analytique orientée connaissances pour la caractérisation et l'identification du locuteur*, Thèse de doctorat, Université d'Avignon, 1994.
- [Bonastre 2000] Bonastre J.-F., *Reconnaissance du locuteur et approche statistique : description, limites, et potentialités*, Habilitation à diriger des recherches, Université d'Avignon, 2000.
- [Carey 1992] Carey M. J. et Parris E. S., Speaker verification using connected words, dans *Proceedings of Institute of Acoustics*, tome 14(6), pages 95–100, 1992.

- [Charlet 1997] Charlet D., *Authentification vocale par téléphone en mode dépendant du texte*, Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications (ENST), 1997.
- [Chen 1997] Chen F., Hearst M., Kimber D., Kupiec J., Pederson J. et Wilcox L., *Managing Multimedia data : using metadata to integrate and apply digital data*, chapitre Metadata for mixed-media access, Mc Graw Hill, 1997.
- [Chen 1998] Chen S. et Gopalakrishnan P., Speaker, environment and channel change detection and clustering via the bayesian information criterion, dans *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [Crystal 1982] Crystal T. et House A., Segmental durations in connected speech signals : Preliminary results, *Journal of the Acoustical Society of America*, pages 29–45, 1982.
- [Delacourt 2000a] Delacourt P., *La segmentation et le regroupement par locuteurs pour l'indexation de document audio*, Thèse de doctorat, ENST-Eurecom, 2000.
- [Delacourt 2000b] Delacourt P. et Welkens C. J., DISTBIC : A speaker based segmentation for audio data indexing, *Speech Communication*, 32 :111–126, 2000.
- [Delacourt 2000c] Delacourt P. et Wellekens C., Regroupement par locuteurs de messages vocaux, dans *CORESA2000 : 6èmes Journées d'Études et d'Échanges COMpression et REprésentation des Signaux Audiovisuels*, Futuroscope Poitiers, France, 2000.
- [Delacourt 1999] Delacourt P. et Wellekens C. J., A first step into speaker-based indexing, dans *European Workshop on Content-Based Multimedia Indexing, CBMI 1999*, 1999.
- [Dempster 1977] Dempster A. P., Laird N. M. et Rubin D. B., Maximum-likelihood from incomplete data via the EM algorithm, dans *Journal of Acoustical Society of America (JASA)*, tome 39, pages 1–38, 1977.
- [ELISA 2000] ELISA, The ELISA systems for the NIST 99 evaluation in speaker detection and tracking, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Everitt 1993] Everitt B., *Cluster analysis*, Oxford University Press Inc., New York, third édition, 1993.
- [Fredouille 2000a] Fredouille C., *Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayésienne des Vraisemblances*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 2000.
- [Fredouille 1999] Fredouille C., Bonastre J.-F. et Merlin T., Similarity normalization method based on world model and a posteriori probability for speaker verification, dans *European Conference on Speech Communication and Technology (Eurospeech)*, tome 2, pages 983–986, 1999.
- [Fredouille 2000b] Fredouille C., Bonastre J.-F. et Merlin T., AMIRAL : a block-segmental multirecognizer architecture for automatic speaker recognition, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Furui 1981] Furui S., Cepstral analysis technique for automatic speaker verification, dans *IEEE Transactions on Acoustics, Speech, and Signal Processing*, tome 29(2), pages 254–272, 1981.
- [Furui 1997] Furui S., Recent advances in speaker recognition, dans *Audio, Video-based Biometric Person Authentication (AVBPA)*, pages 237–252, 1997.
- [Gauvain 1999] Gauvain J., de Kercadio Y., Lamel L. et Adda G., The LIMSI SDR system for TREC-8, dans *Proceeding of 8th Text Retrieval Conference TREC-8*, 1999.
- [Gauvain 2001] Gauvain J., Lamel L. et Adda G., Audio partitioning and transcription for broadcast data indexation, *Multimedia Tools and Applications*, pages 187–200, 2001.

- [Gauvain 1994] Gauvain J. L. et Lee C. H., Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, dans *IEEE Transactions on Speech and Audio Processing*, tome 2(2), pages 291–298, 1994.
- [Gish 1986] Gish H., Krasner M., Russel W. et Wolf J., Methods and experiments for text-independent speaker recognition over telephone channels, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 86)*, pages 865–868, 1986.
- [Gravier 1998] Gravier G. et Chollet G., Comparison of normalization techniques for speaker recognition, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 97–100, 1998.
- [Gravier 2000] Gravier G., Kharroubi J. et Chollet G., On the use of prior knowledge in normalization schemes for speaker verification, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Gravier 1999] Gravier G., Kharroubi J., Chollet G., Bimbot F., Blouet R., Seck M., Bonastre J.-F., Fredouille C., Merlin T., Pigeon S., Verlinde P., Cernocky J., Petrovska D., Nedic B., Magrin-Chagnolleau I. et Durou G., The ELISA'99 speaker recognition and tracking, dans *Workshop on Automatic Identification Advanced Technologies (AutoId)*, 1999.
- [Grenier 1980] Grenier Y., Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur, dans *IXèmes Journées d'Etudes sur la Parole (JEP)*, pages 163–171, 1980.
- [Hain 1998] Hain T. et Woodland P., Segmentation and classification of broadcast news audio, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.
- [Hermansky 1994] Hermansky H. et Morgan N., RASTA processing of speech, dans *IEEE Transactions on Speech and Audio Processing*, tome 2, pages 578–589, 1994.
- [Homayounpour 1995] Homayounpour M. M., *Vérification vocale d'identité : dépendante et indépendante du texte*, Thèse de doctorat, Université de Paris-Sud centre d'Orsay, 1995.
- [Homayounpour 1994] Homayounpour M. M. et Chollet G., Performance comparison of some relevant spectral representations for speaker verification, dans *Workshop on Automatic Speaker Recognition, Identification, Verification*, pages 27–30, 1994.
- [Itakura 1975] Itakura F., Line spectrum representation of linear predictive coefficients of speech signals, *Journal of the Acoustical Society of America*, 1975.
- [Jain 1999] Jain A. K., Murty M. N. et Flynn P. J., Data clustering : A review, *ACM Computing Surveys*, 31(3) :264 – 323, 1999.
- [Jelinek 1976] Jelinek F., Continuous speech recognition by statistical methods, dans *Proceedings of the IEEE*, 1976.
- [Johnson 1997] Johnson S. E., *Speaker Tracking*, Masters thesis, University of Cambridge, 1997.
- [Johnson 1999] Johnson S. E., Who spoke when ? - automatic segmentation and clustering for determining speaker turns, dans *Proceedings of EUROSPEECH 99*, 1999.
- [Johnson 1998] Johnson S. E. et Woodland P. C., Speaker clustering using direct maximisation of the mllr-adapted likelihood, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.
- [Jones 2002] Jones D. et Zissman M., Metadata ad-hoc committee for EARS (MacEARS) MacEARS terms of reference, Rapport technique Version 1.2, MIT Lincoln Laboratory, 2002.

- [Kajarekar 2002] Kajarekar S., Adami A. et Hermansky H., OGI submission - NIST 2002 one-speaker detection task, dans *NIST 2002 Speaker Recognition Evaluations*, www.asp.ogi.edu/~sachin/presentations/spver02.ppt, 2002.
- [Karlsson 1998] Karlsson I., Banziger T., Dankovicová J., Johnstone T., Lindberg J., Melin H., Nolan F. et Scherer K., Speaker verification with elicited speaking-styles in the verivox project, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 207–210, 1998.
- [Kemp 2000] Kemp T., Schmidt M., Westphal M. et Waibel A., Strategies for automatic segmentation of audio data, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2000)*, 2000.
- [Koolwaaij 2000] Koolwaaij J. et Boves L., Local normalization and delayed decision making in speaker dection and tracking, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [LDC 1998] LDC, Lingistic data consortium, <http://www ldc.upenn.edu>, 1998.
- [LDC 2002] LDC, SWB cellular project status pages, <http://www ldc.upenn.edu./Projects/SWB/cellular>, 2002.
- [Lefevre 2000] Lefevre F., *Estimation de probabilité non-paramétrique pour la reconnaissance markovienne de la parole*, Thèse de doctorat, Université Pierre et Marie Curie, 2000.
- [Leggetter 1995] Leggetter C. et P. W., Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, *Computer Speech and Language*, pages 171–185, 1995.
- [Levinson 1986] Levinson S., Continuously variable duration hidden markov models for automatic speech recognition, *Computer Speech and Language*, pages 29–45, 1986.
- [Li 2001] Li D., Sethi I., Dimitrova N. et McGee T., Classification of general audio for content-based retrival, *Pattern Recognition letters*, 2001.
- [Liu 1996] Liu C.-S., Wang H.-C. et Lee C.-H., Speaker verification using normalized log-likelihood score, dans *IEEE Transactions on Speech and Audio Processing*, tome 4(1), pages 56–60, 1996.
- [Magrin-Chagnolleau 2001] Magrin-Chagnolleau I., Gravier G., Blouet R. et for the ELISA consortium, Overview of the ELISA consortium research activities, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 67–72, 2001.
- [Martin 2000] Martin A. et Przybocki M., The NIST 1999 speaker recognition evaluation - an overview, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, 10(1-3), 2000.
- [Martin 2001] Martin A. F. et Przbocki M., The NIST speaker recognition evaluations : 1996-2001, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 36–42, Crete, Greece, 2001.
- [Martin 1997] Martin A. F. et Przybocki M. A., The DET curve in assessment of detection task performance, dans *Proceedings of EUROSPEECH 97*, pages 1895–1898, 1997.
- [McLaughlin 1999] McLaughlin J., Reynolds D., Singer E. et O’Leary G. C., Automatic speaker clustering from multi-speaker utterances, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 99)*, 1999.

- [Meignier 2001] Meignier S., Bonastre J.-F. et Igounet S., E-HMM approach for learning and adapting sound models for speaker indexing, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 175–180, 2001.
- [Moraru 2001] Moraru D., *Indexation de signaux en locuteurs*, DEA, CLIPS-IMAG, Institut National Polytechnique de Grenoble, 2001.
- [Moraru 2002] Moraru D., Segmentation en locuteurs, Rapport technique, CLIPS-IMAG, Université Joseph Fourier, 2002.
- [Moraru 2003] Moraru D., Meignier S., Besacier L. et Bonastre J., Combining experts for automatic speaker segmentation, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003)*, 2003, , soumis.
- [NIST 2000] NIST, The 2000 NIST speaker recognition evaluation plan, <http://www.nist.gov/speech/tests/spk/2000/doc/spk-2000-plan-v1.0.htm>, 2000.
- [NIST 2000] NIST, rédacteur, *Proceedings of the 2000 Speech Transcription Workshop*, University College Conference Center, University of Maryland, 2000, <http://www.nist.gov/speech/publications/tw00/index.htm>.
- [NIST 2001] NIST, The NIST 2001 speaker recognition evaluation plan, <http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrrec-evalplan-v53.%pdf>, 2001.
- [NIST 2002a] NIST, Automatic meeting transcription project, http://www.nist.gov/speech/test_beds/mr_proj/, 2002.
- [NIST 2002b] NIST, The NIST year 2002 speaker recognition evaluation plan, <http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrrec-evalplan-v60.%pdf>, 2002.
- [NIST 2002c] NIST, RT-2002 evaluation plan, http://www.nist.gov/speech/tests/rt/2002/rt02_eval_plan_v3.pdf, 2002.
- [NIST 2002d] NIST, <http://www.nist.gov/speech/tests/spk/index.html>, 2002.
- [Oglesby 1990] Oglesby J. et Mason J. S., Optimisation of neural models for speaker identification, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 90)*, pages 261–264, 1990.
- [Pelecanos 2001] Pelecanos J. et Sridharan S., Feature warping for robust speaker verification, dans *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, pages 213–218, 2001.
- [Poritz 1982] Poritz A., Linear predictive hidden markov models dans speech signal, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 82)*, 1982.
- [Przybocki 1998] Przybocki M. A. et Martin A. F., NIST speaker recognition evaluation - 97, dans *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 120–123, 1998.
- [Rabiner 1989] Rabiner L. R., A tutorial on Hidden Markov Models and selected applications in speech recognition, dans *IEEE transactions Speech Audio Processing*, tome 77(2), pages 257–285, 1989.
- [Reynolds 2002] Reynolds D., Peskin B., Navratil J., Campbell J., Andrews W., Klusacek D., ans Adami A., Jin Q., Abramson J. et Mihaescu R., *SuperSID : Exploiting High-Level Information for High-Performance Speaker Recognition*, The Center for Language and Speech Processing, The Johns Hopkins University, 2002, <http://www.clsp.jhu.edu/ws2002/groups/supersid/>.

- [Reynolds 1998] Reynolds D., Singer E., Carlson B., G.C. O'Leary J. M. et Zissman M., Blind clustering of speech utterances based on speaker and language characteristics, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 98)*, 1998.
- [Reynolds 1992] Reynolds D. A., *A Gaussian mixture modeling approach to text-independent speaker identification*, Phd thesis, Georgia Institute of Technology, 1992.
- [Reynolds 1994] Reynolds D. A., Experimental evaluation of features for robust speaker identification, dans *IEEE transactions Speech Audio Processing*, tome 2, pages 639–643, 1994.
- [Reynolds 1995] Reynolds D. A., Speaker identification and verification using gaussian mixture speaker models, dans *Speech Communication*, tome 17(1-2), pages 91–108, 1995.
- [Reynolds 1997] Reynolds D. A., Comparison of background normalization methods for text-independent speaker verification, dans *Proceedings of EUROSPEECH 97*, 1997.
- [Reynolds 2000] Reynolds D. A., Dunm R. B. et Laughlin J. J., The lincoln speaker recognition system : NIST EVAL2000, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2000)*, 2000.
- [van Rijsbergen 1979] van Rijsbergen C., *Information Retrieval*, Butterworths, New York, 1979.
- [Rosenberg 1990] Rosenberg A. E., Lee C. et Soong F., Sub-word talker verification using hidden markov modeling, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 90)*, 1990.
- [Russel 1985] Russel M. J. et Moore R. K., Explicit modeling of state occupancy in hidden markov models for automatic speech recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 85)*, pages 5–8, 1985.
- [Salton 1983] Salton G. et McGill M., *Introduction to Modern Information Retrieval.*, McGraw Hill Book Company, New York, 1983.
- [Saporta 1990] Saporta G., *Probabilité Analyse des données et statistique*, Tecnip, 1990.
- [Scherer 1998] Scherer K. R., Johnstone T. et Sangsue J., L'état émotionnel du locuteur : facteur négligé mais non négligeable pour la technologie de la parole, dans *XXIIèmes Journées d'Etudes sur la Parole (JEP)*, pages 249–257, 1998.
- [Schroeder 2000] Schroeder J. et Campbell J., rédacteurs, *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Academic Press, 2000.
- [Seck 2001] Seck M., *Détection de rupture et suivi de classe de sons pour l'indexation sonore*, Thèse de doctorat, Université de Rennes 1, IRISA-sigma2, 2001.
- [Siegler 1997] Siegler M., U. Jain U., Raj B. et Stern R., Automatic segmentation and clustering of broadcast news audio, dans *the DARPA Speech Recognition Workshop*, Westfields, Chantilly, Virginia, 1997.
- [Siu 1992] Siu M.-H., Rohlicek R. et Gish H., An unsupervised, sequential learning algorithm for segmentation of speech waveforms with multi speakers, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 92)*, 1992.
- [Siu 1991] Siu M.-H., Yu G. et Gish H., Segregation of speakers for speech recognition and speaker identification, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 91)*, 1991.
- [Smeaton 2001] Smeaton A. F., Indexing, browsing and searching of digital video and digital audio information, *Lecture Notes on Information Retrieval*, M. Agosti et al. Eds, Springer-Verlag, 2001, page 0029, 2001.

- [Solomonoff 1998] Solomonoff A., Mielke A., Schmidt M. et Gish H., Clustering speakers by their voices, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 98)*, 1998.
- [Soong 1988] Soong F. K. et Rosenberg A. E., On the use of instantaneous and transitional spectral information in speaker recognition, dans *IEEE Acoustics Transactions, Speech, and Signal Processing (ASSP)*, tome 36(6), pages 871–879, 1988.
- [Soong 1992] Soong F. K., Rosenberg A. E., Rabiner L. R. et Juang B. H., A vector quantization approach to speaker recognition, dans *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 92)*, pages 387–390, 1992.
- [Struyf 1997] Struyf A., Hubert M. et Rousseeuv P. J., Integrating robust clustering techniques in S-Plus, *Computational Statistic and data analysis*, pages 17–37, 1997.
- [Thong 2000] Thong J.-M. V., Goddeau D., Litvinova A., Logan B., Moreno P. et Swain M., Speechbot : a speech recognition based audio indexing system for the web, dans *Conference on Content-Based Multimedia Information Access, RIAO 2000*, 2000.
- [Van Vuuren 1996] Van Vuuren S., Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 96)*, pages 1788–1791, 1996.
- [Viswanathan 2000] Viswanathan M., Maali F., Beigi H. et Tritschler A., Multimedia information access using multiple speaker classifiers, dans *Conference on Content-Based Multimedia Information Access, RIAO 2000*, 2000.
- [Viterbi 1967] Viterbi A., Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, dans *IEEE Transactions on Information Theory*, 1967.
- [W3C 1998] W3C, Synchronized multimedia integration language (SMIL) 1.0 specification, <http://www.w3.org/TR/REC-smil/>, 1998.
- [Weber 2000] Weber M. et Kemp T., Evaluating different information retrieval algorithms on real-world data, dans *Proceedings of International Conference on Spoken Language Processing (ICSLP 2000)*, 2000.
- [Wilcox 1994] Wilcox L., Kimber D. et Chen F., Audio indexing using speaker identification, *SPIE*, pages 149–157, 1994.