INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : « Signal Image Parole Télécoms »

préparée au laboratoire CLIPS – IMAG (Communication Langagière et Interaction Personne Système)
dans le cadre de **l'Ecole Doctorale** « ***Electronique, Electrotechnique, Automatique, Télécommunications, Signal*** »

présentée et soutenue publiquement

par

Daniel Moraru

le 20 Décembre 2004

SEGMENTATION EN LOCUTEURS DE DOCUMENTS AUDIOS ET AUDIOVISUELS : APPLICATION A LA RECHERCHE D'INFORMATION MULTIMEDIA

Directeur de thèse : Eric Castelli
Codirecteur de thèse : Laurent Besacier

JURY

Mme. Catherine Berrut Professeur, HDR, UJF, Grenoble	, Président
M. Frédéric Bimbot Chargé de recherche, HDR, IRISA, Rennes	, Rapporteur
Mme. Jenny Benois-Pineau Professeur, HDR, Univ. Bordeaux I, Bordeaux	, Rapporteur
M. Jean-François Bonastre Maître de conférences, HDR, UAPV, Avignon	, Examinateur
M. Corneliu Burileanu Professeur, HDR, Univ. "Politehnica", Bucarest	, Examinateur
M. Eric Castelli Maître de conférences , HDR, INPG, Grenoble	, Directeur de thèse
M. Laurent Besacier Maître de conférences, UJF, Grenoble	, Codirecteur de thèse

… à mes parents

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury pour avoir accepté de participer à la soutenance de cette thèse en commençant par Madame Catherine Berrut pour avoir accepté de présider le jury.

Je remercie vivement Madame Jenny Benois-Pineau et Monsieur Frédéric Bimbot, pour avoir consacré du temps à la lecture de ce document en tant que rapporteurs de ce travail.

Je remercie également mes deux directeurs de thèse Monsieur Eric Castelli et Monsieur Laurent Besacier pour avoir accepté d'encadrer cette thèse. Je remercie très chaleureusement Monsieur Laurent Besacier pour tous ses conseils et critiques sur le plan scientifique mais également pour la relation humaine qu'il a su développer.

Un grand remerciement à Monsieur Corneliu Burileanu, examinateur de cette thèse, pour avoir guidé mes premiers pas dans le domaine du traitement de la parole et pour m'avoir proposé de venir travailler au laboratoire CLIPS à Grenoble.

Un grand remerciement également à Monsieur Jean-François Bonastre, examinateur de cette thèse, pour l'intérêt porté à mes travaux de recherche et pour m'avoir accueilli au sein du consortium ELISA. Je remercie à cette occasion les différentes personnes du Laboratoire d'Informatique d'Avignon : Sylvain, Corinne, Teva mais aussi tous les autres membres du consortium ELISA qui ont eu un rôle déterminant dans l'avancement de mon travail.

J'adresse mes remerciements à Monsieur Jean Caelen directeur du laboratoire CLIPS et à Monsieur Jean-François Serignat responsable de l'équipe GEOD pour m'avoir accueilli dans le laboratoire. Je remercie à cette occasion Audrey, Anne, Anne-Claire, Richard, Yannick, Dominique les tous les membres de l'équipe et également les autres membres du laboratoire en particulier à Monsieur Georges Quénot pour le travail réalisé ensemble.

Je remercie toute ma famille pour son soutien tout le long de cette thèse notamment dans les moments les plus difficiles.

Pour finir je tiens à remercier spécialement tous mes amis : Cristina, Monika, Catalin, Stefan, Cora, Iulian, Ioana, Radu, Flo, Nico, Daniela et Betty et plus particulièrement Dan et Vali, les premiers roumains que j'ai rencontré à Grenoble pour les bons moments passés tout le long de ces années.

Résumé

Cette thèse se situe à la frontière des domaines de la recherche d'information multimédia et du traitement automatique de la parole. Dans ce dernier domaine, une nouvelle tâche est apparue ces dernières années : la transcription enrichie d'un document audio. Une *meta-donnée* importante pour la transcription enrichie est l'information *locuteur* qui précise pour un document donné *"Qui parle et quand?"*. La segmentation d'un document en locuteurs est l'objet principal de cette thèse.

Au delà de la réalisation d'un système de segmentation en locuteurs basé sur la modélisation statistique du locuteur, notre intérêt s'est porté sur l'intégration d'informations *a priori* dans un système de segmentation et aussi sur son application à la recherche d'information multimédia. Dans tous les cas, nos travaux de recherche ont été validés dans un cadre expérimental rigoureux autour de campagnes d'évaluation internationales sur des données de différents types : enregistrements téléphoniques, journaux télévisés ou réunions.

Nos expérimentations concernant l'apport de différentes informations *a priori*, ont montré une réduction importante d'erreur de segmentation dans le cas de l'utilisation des certaines informations (une annotation incomplète disponible, une pré-segmentation acoustique obtenue automatiquement).

Concernant l'utilisation de l'information *locuteur* pour l'indexation d'une grande collection de documents audio-visuels, les résultats expérimentaux montrent son apport dans des tâches de recherche et, inversement, montrent l'intérêt du canal vidéo pour la segmentation en locuteurs Notre travail se termine avec la proposition d'un système de segmentation en histoires de documents vidéos qui utilise simultanément des données audio, vidéo et texte.

Mots-clés : indexation, segmentation en locuteurs, modèles statistiques, recherche d'information multimédia, informations *a priori*, segmentation de vidéos, segmentation en histoires.

Abstract

This thesis work is at the frontier between multimedia information retrieval and automatic speech processing. During the last years, a new task emerged in speech processing: the rich transcription of an audio document. An important *meta-data* for rich transcription is the *speaker* information which tells us *"Who spoke when?"* for a given audio document. The speaker segmentation task is the main subject of this research work.

Beyond the development of a speaker segmentation system based on speaker statistical modeling, our research interest concerned the use of *a priori* information for speaker segmentation and also its application to multimedia information retrieval. Our research work was validated in a rigorous experimental frame-work during international evaluation campaigns on different data types: telephone data, broadcast news data and meeting data.

Our experiments concerning the influence of different *a priori* information have shown a significant speaker segmentation error reduction for certain information (an incomplete speaker annotation, an automatic acoustic pre-segmentation).

Concerning the use of *speaker* information for very large audio-video database indexation, the experiments have shown its importance for information retrieval tasks but also the interest of using the video channel for speaker segmentation itself. To conclude our research work, we propose a story segmentation system for video documents which simultaneously uses audio, video and text data.

Key Words: indexation, speaker segmentation, statistical modeling, multimedia information retrieval, *a priori* information, video segmentation, story segmentation.

Table des matières

Introduction 1
1. Contexte 3
2. Définition de la tâche 5
3. Le domaine de la reconnaissance du locuteur 7
4. Problématique et méthodologie 8
5. Le manuscrit 10

Premiere partie : Etat de l'art 13
1. Architecture générale d'un système de segmentation 15
2. Outils pour le traitement de la parole utilisés en segmentation 16
2.1. Paramétrisation 16
2.1.1 Les coefficients MFCC 17
2.1.2 La moyenne de coefficients cepstraux 20
2.1.3 Coefficients différentiels et énergie 20
2.2. La modélisation du locuteur 21
2.2.1 Les mixtures de gaussiennes 21
2.2.2 Les modèles de Markov cachés 29
2.2.3 La quantification vectorielle 33
3. Segmentation en locuteurs 36
3.1. Pré-Segmentation acoustique 37
3.1.1 Approche utilisant des paramètres acoustiques autres que les coefficients cepstraux 38
3.1.2 Approche utilisant des paramètres cepstraux 41
3.1.3 Décision inter-classe dans le cas de la pré-segmentation acoustique 41
3.2. Détection de changements de locuteur 43
3.2.1 Détection de changements de locuteur par détection des silences 44
3.2.2 Détection de changements de locuteur par utilisation d'une distance 44
3.2.3 Détection de changements de locuteur par identification de la nature des segments 47
3.3. Regroupement des segments 48
3.3.1 Regroupement par division 49

3.3.2 Regroupement par agglomération 50
3.3.3 Critère d'arrêt du regroupement hiérarchique 52
3.4. Re-segmentation 53
4. Segmentation en locuteurs : format du résultat et évaluation 54
5. Différentes architectures de systèmes de segmentation 58
5.1. L'architecture ascendante 58
5.2. L'architecture évolutive ou "intégrée" 59
5.3. Architecture temps réel 61
5.4. Commentaires sur les trois architectures 64
6. Le suivi du locuteur 65
6.1. Généralités 65
6.2. La tâche de suivi du locuteur dans les évaluations NIST et ESTER 67
7. Conclusion générale 68

Deuxième partie : Segmentation en locuteurs sans informations *a priori* ; système de référence 71
1. Introduction 73
2. Plate-forme expérimentale et corpus de segmentation 73
2.1. Plate-forme ELISA 73
2.2. Corpus expérimental issu des campagnes d'évaluation NIST et ESTER 74
3. Le système de segmentation initial (2001-2002) 75
3.1. La paramétrisation 76
3.2. Détection de changements de locuteur 76
3.3. Le regroupement des segments 80
3.4. Premiers résultats : Campagne d'évaluation NIST SpRec 2002 82
4. Amélioration du système initial (2002-2003) 84
4.1. La paramétrisation 84
4.2. Estimation du nombre de locuteurs 85
5. Ajout des modules LIA dans notre système de référence (2003-2004) 88
5.1. La pré-segmentation acoustique 88
5.2. La re-segmentation 91
6. Résultats du système de référence final : Evaluations RT 2003 et ESTER 92
7. Conclusion 93

Troisième partie : Intégration d'informations *a priori* dans un système de segmentation en locuteurs 95
1. Introduction 97
2. Informations *a priori* obtenues par annotation manuelle (certaines) 98
2.1. Connaissance *a priori* du nombre de locuteurs 99
2.2. Données de référence disponibles pour tous les locuteurs 101
3. Informations *a priori* obtenues pas annotation automatique (incertaines) 105
3.1. Une pré-segmentation acoustique disponible *a priori* 105
3.2. Utilisation *a priori* d'un modèle pour certains locuteurs seulement 108
3.3. Utilisation d'un ou plusieurs résultats de segmentation disponibles *a priori* 111
3.3.1 Deux systèmes cascadés (ELISA hybride) 111
3.3.2 Fusion d'étiquettes (ELISA fusion) 112
3.4. Utilisation de multiples capteurs pour la segmentation en locuteurs (enregistrements de réunions, évaluations RT 04 Meeting) 116
4. Informations *a priori* qui se sont avérées non-utilisables 120
4.1. Utilisation de la probabilité *a priori* du locuteur majoritaire 121
4.2. Utilisation de la durée moyenne d'un tour de parole 122
5. Conclusion 123

Quatrième partie : Application à la recherche d'information multimédia 127
1. Introduction 129
2. Recherche d'information multimédia : généralités et métriques d'évaluation 130
3. Contexte de travail : la campagne d'évaluation TRECVID 132
3.1. Introduction 134
3.2. Corpus vidéo expérimental 136
3.3. Campagne d'évaluation TREC : description des tâches d'extraction 137
3.3.1 Extraction de plans contenant de la parole : critère audio pur pour la recherche d'un document vidéo 138
3.3.2 Extraction de plans Monologue : critère conjoint audio-vidéo pour la recherche d'un document vidéo 138
3.3.3 Extraction de plans Personne X : critère multimodal (audio-video-texte) pour la recherche d'un document vidéo 140

3.4. Premiers résultats sur les évaluations TREC 142
3.5. Enseignements tirés des campagnes TREC pour notre travail de thèse 144
4. Tâches inspirées de la campagne d'évaluation TRECVID 145
4.1. Utilisation de la détection de frontières de plans vidéo dans un système de segmentation en locuteurs (*comment l'information vidéo peut aider au traitement de l'audio* ?) 146
4.2. Identification sans données *a priori* du présentateur d'un journal télévisé sur un document vidéo 147
4.3. Expérimentations et résultats 150
4.3.1 Segmentation audio/vidéo en locuteurs 150
4.3.2 Extraction de plans où le présentateur parle 151
5. Vers une segmentation vidéo multimodale : la segmentation en histoires 153
5.1. Un premier système de segmentation en histoires proposé au CLIPS 153
5.2. Premiers résultats expérimentaux 154
6. Conclusions 155

Conclusions et perspectives 157
1. Conclusions 159
1.1. Evaluation d'un système sur le long terme et sur plusieurs types de données. 159
1.2. Utilisation d'informations *a priori* pour la segmentation en locuteurs 160
1.3. Application à la recherche d'informations multimédia 162
2. Perspectives 163

Annexes 167
Annexe A Démonstration de la formule de vraisemblance généralisé 169
Annexe B Bibliographie personnelle 172
Annexe C Article Informations a priori conférence A Speaker Odyssey 2004 174
Annexe D Article segmentation en locuteurs a priori conférence A Speaker Odyssey 2004 181
Liste des abréviations 187
Bibliographie 189

Liste des figures

Figure 1 Architecture d'un moteur de transcription enrichie 3
Figure 2 Evolution du domaine en traitement automatique de la parole 4

Figure I. 1 Correspondance mels / hertz 18
Figure I. 2 Banc de filtres MEL 19
Figure I. 3 Calcul des coefficients MFCC 19
Figure I. 4 Adaptation d'un modèle du monde (GMM avec quatre gaussiennes) sur un locuteur particulier. Deux gaussiennes ont changé de moyenne 26
Figure I. 5 Un modèle HMM ergodique (toutes les transitions sont possibles) 30
Figure I. 6 Quantification vectorielle pour un système de télécommunication 34
Figure I. 7 Résultat d'un classificateur de plus proches voisins 34
Figure I. 8 Architecture générale d'un système de segmentation 36
Figure I. 9 Décision hiérarchique pour la pré-segmentation acoustique 42
Figure I. 10 Utilisation de deux fenêtres adjacentes pour le calcul d'une distance 45
Figure I. 11 Le regroupement hiérarchique 49
Figure I. 12 Mise à jour d'une matrice de distance pour le regroupement hiérarchique 52
Figure I. 13 Evaluation NIST de la segmentation en locuteur 58
Figure I. 14 Architecture évolutive pour la segmentation en locuteurs 60
Figure I. 15 Architecture de segmentation en locuteurs temps réel 63
Figure I. 16 Différences entre la sortie d'un système de suivi du locuteur et la sortie d'un système de segmentation en locuteurs 66

Figure II. 1 Système de segmentation initial 76
Figure II. 2 Les fenêtres glissantes pour la détection des changements de locuteur 77
Figure II. 3 La recherche de maxima locaux sur la courbe BIC 80
Figure II. 4 Le regroupement hiérarchique de notre système de segmentation 81
Figure II. 5 Détecteurs de changements de locuteur : BIC et Taille Fixe 82
Figure II. 6 Pourcentage d'erreur de segmentation en fonction du nombre de filtres utilisés pour le calcul de coefficients MFCC sur le corpus BN 2002 85
Figure II. 7 La pré-segmentation acoustique 89
Figure II. 8 Segmentation en locuteurs appliquée sur chaque classe acoustique 90

Figure II. 9 Re-segmentation à partir d'un HMM ergodique, un état par locuteur.....91

Figure III. 1 Système de segmentation quand le nombre de locuteurs est connu.....99
Figure III. 2 Système de segmentation quand des données de référence sont disponibles pour tous les locuteurs 102
Figure III. 3 Pourcentage d'erreur de segmentation en fonction de la quantité de données a priori disponible pour chaque locuteur pour le corpus RT 2003 Dev 103
Figure III. 4 Pourcentage d'erreur de segmentation en fonction de la quantité de données a priori disponible pour chaque locuteur pour le corpus RT 2003 Eva.....103
Figure III. 5 Utilisation de données appartenant à un locuteur comme information *a priori* pour la segmentation en locuteurs 109
Figure III. 6 Taux d'erreur utilisés dans les résultats concernant les expérimentations de la section 3.2.....110
Figure III. 7 Utilisation d'un résultat de segmentation comme information *a priori* pour un autre système de segmentation 112
Figure III. 8 Utilisation de plusieurs résultats de segmentation comme information *a priori* pour un autre système de segmentation.....114
Figure III. 9 Exemple de résultat final de segmentation après l'utilisation de la fusion d'étiquettes 114
Figure III. 10 L'algorithme de "Stratégie de Fusion sur Microphones Individuels" (SFMI) .118
Figure III. 11 Pré-segmentation acoustique parole/non-parole pour un enregistrement d'une réunion avec plusieurs capteurs.....119

Figure IV. 1 La recherche d'information interactive 135
Figure IV. 2 La détection de plans vidéo contenant du monologue.....139
Figure IV. 3 Recherche de plans vidéo contenant Madleine Albright dans une collection de documents vidéo.....141
Figure IV. 4 Le système audio/vidéo de segmentation en locuteurs.....146
Figure IV. 5 Differents detecteurs de changements de locuteur.....147
Figure IV. 6 Sortie d'un système de segmentation en locuteurs sur un journal français radiodiffusé de 20 minutes. Le présentateur est le locuteur L0 (premiere ligne).....148
Figure IV. 7 Sortie d'un système de segmentation en locuteurs sur un journal anglais télévisé de 30 minutes. Le présentateur est le locuteur MS4 (deuxieme ligne).....149

Figure IV. 8 Calcul de l'erreur de détection du présentateur 152

Liste des tableaux

Tableau 1 Moteurs de recherche multimédia pour le grand public 5
Tableau 2 Problèmes spécifiques selon le type de documents à segmenter 9

Tableau II. 1 Les différents corpus utilisés dans nos expérimentations 75
Tableau II. 2 Résultats officiels NIST SpRec 2002 ; pourcentage d'erreur de segmentation et temps d'exécution pour les systèmes CLIPS 82
Tableau II. 3 Erreur de segmentation en fixant ou en estimant le nombre de locuteurs pour le corpus BN 2002 88
Tableau II. 4 Pourcentage d'erreur de segmentation de notre système de référence pour les corpus RT 2003 et ESTER sans aucune information *a priori* et sans re-segmentation LIA 92

Tableau III. 1 Différences entre documents audio (données SpRec 2002) 98
Tableau III. 2 Pourcentage d'erreur de segmentation pour différents critères d'arrêt sur le nombre de locuteurs présents dans un enregistrement 100
Tableau III 3 Nombre total de locuteurs utilisé comme critère d'arrêt dans les différents corpus (cumulé sur tous les fichiers d'un même corpus) 100
Tableau III. 4 Pourcentage d'erreur de segmentation quand des données a priori sont disponibles pour tous les locuteurs 104
Tableau III. 5 Erreurs de pré-segmentation acoustique sur les corpus RT 2003 Dev et Eva . 106
Tableau III. 6 Erreurs de segmentation en locuteurs sur les corpus RT 2003 Dev et Eva pour différents niveaux de pré-segmentation acoustique utilisés comme information *a priori* 107
Tableau III. 7 Erreur de segmentation initiale et erreur de segmentation due au locuteur connu pour le corpus ESTER Dev 110
Tableau III. 8 Utilisation de données de référence disponibles pour un locuteur (présentateur), résultats sur le corpus ESTER Eva 111
Tableau III. 9 Exemple de fusion d'étiquettes pour quatre systèmes de segmentation 113
Tableau III. 10 Un ou plusieurs résultats de segmentation disponibles *a priori* : Expérimentations sur le corpus RT 2003 Eva 115

Tableau III. 11 Pourcentage d'erreur de segmentation pour le corpus Meet 2004 (RT 04 Meeting) 119
Tableau III. 12 Comparaison entre la probabilité *a priori* et la probabilité *a posteriori* (obtenue automatiquement par notre système) du locuteur majoritaire 122
Tableau III. 13 Comparaison entre la durée moyenne *a priori* et *a posteriori* (obtenue par le système) d'un tour de parole 123
Tableau III. 14 Synthèse des résultats concernant les différentes informations *a priori* utilisées et selon différents types de données (journaux télévisés et réunions). Amélioration (valeur négative) ou dégradation (valeur positive) en terme de pourcentage d'erreur de segmentation 124

Tableau IV. 1 Les tâches proposées dans la campagne TREC ou inspirées de la même campagne sur lesquelles nous avons travaillé 133
Tableau IV. 2 Corpus expérimental TREC Vidéo 136
Tableau IV. 3 Résultats officiels TREC : extraction de plans vidéo contenant un trait (TREC 2002 : parole, monologue et TREC 2003 : personne X). Les mesures de performance utilisées sont : précision (P 1000), rappel (R 1000), F ratio (F 1000) sur 1000 plans et la précision moyenne (PM) 143
Tableau IV. 4 Performance du système de détection de visages : nombre de plans pertinents sur premiers 100 (D 100), nombre de plans pertinents sur premiers 1000 (D 1000) et précision moyenne 143
Tableau IV. 5 L'erreur de segmentation sur le corpus TREC 4F 2003 en utilisant différentes stratégies de détection de changements de locuteur 150
Tableau IV. 6 L'erreur de détection du présentateur sur le corpus TREC 4F 2003 151
Tableau IV. 7 Taux de rappel, précision et F ratio pour l'extraction de plans où le présentateur parle, sur le corpus TREC 4F 2003 151
Tableau IV. 8 Segmentation en histoires, taux de rappel, taux de précision et F Ratio (évaluation réalisé sur 50 heures de signal vidéo du corpus TREC 2003) 155

Tableau 1 Evolution des taux d'erreur de notre système sur 3 ans et sur différentes données. Synthèse de résultats sur les corpus NIST SpRec 2002, NIST RT 2003 et NIST RT 2004 Meeting 160

Introduction

Cette partie est une introduction générale sur le domaine de l'annotation automatique de documents audio. Une courte vue d'ensemble sur l'évolution du domaine est d'abord présentée. Celle-ci nous permettra de positionner notre sujet dans son contexte, qui est la *transcription enrichie* de documents audio et la recherche d'information dans une grande collection de documents vidéo. La problématique du sujet et la méthodologie utilisées seront ensuite présentées et nous finirons par présenter l'organisation du manuscrit.

1. Contexte

Les technologies existantes permettent depuis longtemps de numériser l'information avec beaucoup d'avantages : transfert rapide, reproduction illimitée, stockage facile. Pratiquement toute information multimédia se trouve aujourd'hui sous format numérique. Avec la facilité de stockage due en partie à des algorithmes de compression très efficaces, les corpus de documents audio et vidéo ne cessent de croître. La recherche de documents multimédia basée sur le contenu est devenue ainsi, une tâche très importante.

Dès l'apparition du numérique dans la représentation de données multimédia, l'effort principal a été consacré à la mise en place d'algorithmes efficaces pour la transmission et le stockage, sans envisager la possibilité de manipuler l'information selon le contenu. Cet aspect à tendance à être corrigé avec l'apparition de nouveaux standards comme par exemple le MPEG-7 et MPEG-21.

Pour pouvoir envisager un accès selon le contenu dans de grandes bases de données multimédia, ces données doivent être annotées. Nous nous intéressons dans cette thèse plus particulièrement aux documents audio ou à la partie audio d'un document vidéo. Pour ces signaux, une simple transcription manuelle ou automatique issue d'un moteur de reconnaissance de la parole n'est pas suffisante. Le signal audio, et en particulier le signal de parole est très *riche* en informations. Il contient notamment des informations de haut niveau que nous pouvons appeler meta-données.

Les meta-données sont des informations de haut niveau sémantique issues de l'analyse de la parole. Des exemples de telles informations sont : les hésitations, les frontières de phrases, les locuteurs, etc. Le moteur de reconnaissance de la parole qui fournit normalement une transcription simple (ce qui est dit) peut être complété par le moteur d'extraction de meta-données (Figure 1) pour fournir comme résultat final ce qu'on peut appeler une "Transcription Riche".

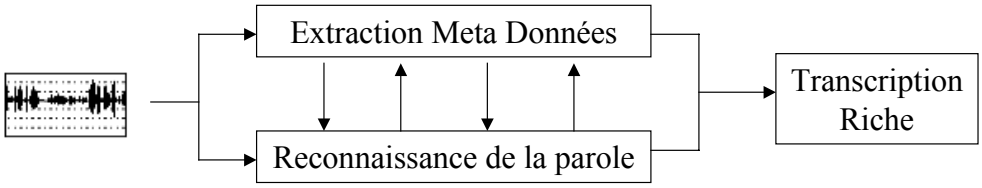


Figure 1 Architecture d'un moteur de transcription enrichie

Ce concept de transcription enrichie est lié fortement à l'évolution du domaine de la reconnaissance automatique de la parole. Avant les années 90, la recherche dans le domaine s'intéressait principalement aux systèmes de dictée vocale et à la transcription d'enregistrements téléphoniques. Vers la fin des années 90, l'intérêt de la recherche s'est porté vers des données plus riches en information et de moins en moins contrôlées comme les journaux télévisés et radiodiffusés, les documents issus de réunions enregistrées dans des "environnements perceptifs" équipés de nombreux capteurs.

Cette récente évolution amène de nouvelles difficultés dans le traitement automatique de la parole illustrées par la Figure 2. Par exemple, dans les systèmes de dictée, le locuteur contrôlait lui même le début et la fin de l'enregistrement. Dans les journaux télévisés, le flux audio est continu et hétérogène et c'est le système qui doit segmenter de façon automatique le signal. Une autre nouvelle problématique est la nécessité de traiter dans le cas d'enregistrements de réunions par exemple, du signal audio provenant de plusieurs capteurs différents (plusieurs microphones, plusieurs cameras vidéo, détecteurs de mouvements, etc.).

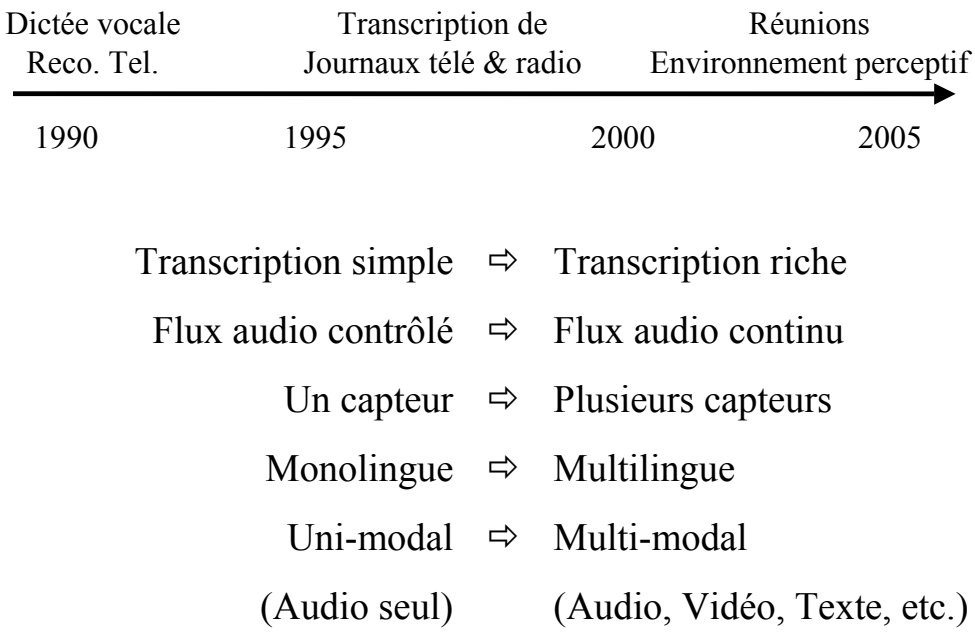


Figure 2 Evolution du domaine en traitement automatique de la parole

La difficulté des nouvelles tâches apparues est due aussi à la grande quantité de données à traiter. A titre d'exemple la Smart Room NIST [Stanford 2003] qui permet d'enregistrer des réunions par exemple, génère environ 1 Go de données par minute issues de

multiples capteurs (pas seulement de la parole). La nécessité d'indexer une telle quantité de données est plus qu'évidente.

Avant d'aller plus loin nous allons présenter très brièvement les outils existants pour le grand public. La plupart des moteurs de recherche sur Internet offrent la possibilité de rechercher des pages web et des images selon le contenu. Par contre il y a quelques moteurs dont le plus connu est le *SpeechBot* de *Compaq* qui offrent la possibilité au grand public de rechercher des vidéos et des fichiers audio selon une requête. Voilà quelques exemples :

Tableau 1 Moteurs de recherche multimédia pour le grand public

Moteur de recherche	Image	Audio	Vidéo
http://speechbot.research.compaq.com/	Non	Oui	Oui
http://fr.altavista.com/video/default	Oui	Oui	Oui
http://www.alltheweb.com/	Oui	Oui	Oui
http://www.singingfish.com/	Non	Oui	Oui
http://disney.ctr.columbia.edu/webseek/	Oui	Non	Oui
http://www.tf1.fr/video/news/lesjt/	Non	Non	Oui

Les documents multimédia dans ce cas sont généralement indexés à partir du contexte (le document vidéo est référencé dans une page web sur tel ou tel sujet) plutôt qu'à partir des caractéristiques extraites du signal brut. Ceci est dû d'une part aux erreurs faites par les extracteurs de caractéristiques mais aussi aux contraintes liées au temps d'exécution et d'accès, un document textuel comme un page web étant bien sûr, accédé et traité beaucoup plus vite qu'un document audio ou vidéo.

2. Définition de la tâche

Nous avons vu que le but de la transcription enrichie est de produire un document structuré et très riche en informations à partir d'un signal audio brut. Dans ce contexte la *segmentation en locuteurs*, qui fait l'objet de cette thèse, est une première étape vers la création d'un document structuré et indexé. Segmenter un document en locuteurs veut dire découper le document en segments homogènes appartenant à un seul locuteur et assembler ces segments selon chaque locuteur. Selon le résultat de la segmentation en locuteurs un index du document est généré. Nous devons bien préciser dès maintenant que par hypothèse aucune information *a priori* n'est disponible sur le document. En conséquence l'identité des locuteurs

n'est pas connue à l'avance. Un locuteur est représenté dans l'index du document par une étiquette dont le code est choisi de façon arbitraire.

Selon la définition du dictionnaire, un index est une liste d'objets d'un ensemble accompagnée des adresses de ces objets afin de pouvoir les retrouver rapidement. Un document indexé est donc un document pour lequel un index a été calculé en fonction d'une clé pour pouvoir permettre de retrouver rapidement une information. Dans notre cas, la clé de l'indexation est le locuteur et les objets sont des parties de signal audio auxquelles nous avons attribué un locuteur. L'ensemble est représenté par une grande collection de documents multimédia (audio pur ou vidéo).

Dans le contexte de la transcription enrichie, le problème *locuteur* consiste à répondre de façon automatique à la question suivante : "Qui parle et quand sur un document audio?". La réponse à cette question peut être obtenue en plusieurs étapes :

- couper un document audio en segments homogènes appartenant à seulement un locuteur ;
- grouper les segments appartenant à un seul locuteur sur un même document ou sur une grande collection de documents audio ;
- identifier le locuteur auquel appartient une partie d'un document audio ;
- utiliser et évaluer ce type d'information pour une tache de type recherche d'information.

La première étape est appelée *détection de changements de locuteur* tandis que la deuxième étape est appelée *regroupement selon le locuteur*. Si nous restons dans le cas où il y a un seul document à traiter, l'ensemble des deux premières étapes définit la tâche de *segmentation en locuteurs*. Si le regroupement est fait sur une grande collection de documents audio nous pouvons alors parler *d'appariement selon le locuteur* [Meignier 2002]. La troisième étape est appelée *suivi du locuteur* et dans ce cas certains locuteurs appelés locuteurs cibles sont connus à l'avance par le système. Le système dispose alors d'un modèle *a priori* pour chacun d'entre eux.

Cette thèse est essentiellement consacrée à la *segmentation en locuteurs*, mais la dernière étape sera également abordée dans une des parties de cette thèse où nous essayerons de répondre à la question "Quelles sont les performances et l'apport d'un système de segmentation en locuteurs pour une tâche de type recherche d'information?".

En effet, l'information *locuteur* sur un document audio peut apporter des réponses à différentes questions selon le type de document. Pour un journal télévisé, nous pouvons

essayer de retrouver toutes les interventions d'une personnalité connue. Pour un enregistrement de réunion, nous pouvons calculer le temps de parole de chaque intervenant et aussi obtenir automatiquement un document structuré qui sera une aide importante à la rédaction du compte rendu ou des minutes.

3. Le domaine de la reconnaissance du locuteur

Le problème *locuteur* n'est pas spécifique à la transcription enrichie. Dans le domaine du traitement de la parole, depuis les années 70 de nombreuses recherches ont été faites en reconnaissance du locuteur avec comme but des applications de type judiciaire, serrure vocale ou authentification pour des transactions téléphoniques. Les premières tâches dans ce domaine ont été *l'identification du locuteur* et la *vérification du locuteur*.

L'identification du locuteur doit établir à partir d'un échantillon de voix l'identité d'une personne parmi N personnes connues à l'avance par le système. Le système peut décider alors que l'échantillon de voix appartient à une des N personnes connues ou qu'il appartient à une personne inconnue.

La vérification du locuteur doit établir si un échantillon de voix correspond à une identité proclamée. La réponse du système dans ce cas est binaire : l'identité proclamée est l'identité réelle du locuteur ou non.

Dans ces deux tâches, le flux audio est contrôlé : le signal audio contient donc presque uniquement de la parole qui appartient à un seul locuteur et l'environnement varie peu pour un même enregistrement (niveau de bruit, type de microphone utilisé, etc.). La première évolution du domaine vers un flux audio continu est la tâche de *suivi du locuteur*.

Le suivi du locuteur a comme but de rechercher à partir d'un enregistrement où il y a plusieurs locuteurs qui parlent, les segments appartenant à un certain locuteur connu par le système. Cette tâche est assez proche de la segmentation en locuteurs mais il existe quelques différences : au niveau de l'évaluation, le suivi du locuteur s'intéresse uniquement à certains locuteurs et pas à tous les locuteurs, au niveau de la mise en œuvre ces locuteurs sont connus à l'avance par le système (le système dispose de modèles *a priori* pour les locuteurs concernés appelés locuteurs cible).

Pour résoudre le problème de la segmentation en locuteurs nous pouvons nous inspirer des techniques existantes en reconnaissance du locuteur. Il existe un certain nombre d'outils théoriques communs aux deux tâches comme la paramétrisation du signal de parole et la

modélisation statistique du locuteur. La principale différence par rapport à une tâche classique d'identification ou de vérification du locuteur vient, comme nous l'avons déjà précisé, de la nécessité de traiter un flux audio continu et hétérogène mais aussi de l'absence d'un modèle de locuteur *a priori*. Le signal audio ne contient pas seulement de la parole, plusieurs locuteurs (jusqu'à plusieurs dizaines sur les enregistrements de journaux télévisés) peuvent être présents, la conversation entre différents locuteurs peut être également très spontanée (les locuteurs se coupent la parole, durée très variable des interventions de chaque locuteur). D'un autre côté, il existe également des problèmes spécifiques à la reconnaissance du locuteur qui ne se retrouvent pas dans la segmentation en locuteurs, comme par exemple la variabilité intra-locuteur à long terme liée à la vieillesse et à l'état de santé du locuteur.

Nous réutiliserons donc dans cette thèse certains outils spécifiques à la reconnaissance du locuteur mais ils seront adaptés à la tâche de segmentation en locuteurs.

4. Problématique et méthodologie

Nous avons déjà précisé que dans la segmentation en locuteurs aucun locuteur n'est connu à l'avance par le système et aucune information *a priori* comme la présence ou l'absence d'un genre ou d'une qualité de parole (bande large, bande étroite, parole propre, parole bruitée, etc.) n'est censée être disponible concernant les documents de test.

L'absence de toute information *a priori* disponible concernant le documents de test conduit à un certain nombre de problèmes qui doivent être résolus :

p1. le système doit être capable d'estimer de façon automatique le nombre de locuteurs présents dans un document audio ;

p2. le système doit être capable de traiter des documents contenant également de la musique, du silence, du bruit et de la parole de qualités variables ;

p3. le système doit être capable de traiter de la parole spontanée avec des locuteurs qui parlent en même temps, avec des hésitations et des interventions de longueur très variable.

Selon le type de documents à traiter (téléphoniques, journaux, réunions) un ou plusieurs de ces problèmes doivent être résolus, ce qui change considérablement la difficulté de la tâche.

Tableau 2 Problèmes spécifiques selon le type de documents à segmenter

Documents	Téléphoniques	Journaux	Réunions
Problèmes	p3	p1+p2	p1+p2+p3

Le Tableau 2 montre la difficulté croissante des tâches due à l'accumulation des problèmes décrits au paragraphe précédent. Nous verrons plus loin dans ce manuscrit que nos expérimentations sur les trois types de données confirmerons cette difficulté croissante des tâches.

La première étape dans nos travaux de recherche sera de se doter d'une plate-forme expérimentale en réalisant un système de segmentation en locuteurs le plus proche possible de l'état de l'art en terme de performances. Notre système sera évalué et comparé à d'autres systèmes selon un protocole expérimental bien défini et reconnu dans la communauté scientifique du domaine. Ceci sera fait dans le cadre des évaluations américaines NIST (National Insitute of Standard and Technologies) sur la reconnaissance du locuteur et la transcription enrichie, et des évaluations ESTER (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) organisées au niveau national.

Les évaluations NIST [Martin 2001] sont proposées chaque année pour encourager la recherche dans le domaine de la reconnaissance de la parole et du locuteur. Les organisateurs proposent un corpus commun et aussi des métriques d'évaluation communes à tous les participants. Les données proposées en quantité suffisamment importante couvrent aussi tous les types d'enregistrements : enregistrements téléphoniques, enregistrements de journaux télévisés et enregistrements de réunions. Ceci permettra d'étudier les limites d'un système de segmentation confronté à tous les problèmes énoncés au début de cette section.

Une fois notre plate-forme expérimentale validée (elle sera remise en cause chaque année puisque nous avons participé aux évaluations NIST 2002, 2003 et 2004), nous pourrons aborder une partie plus exploratoire. Dans la segmentation en locuteurs, par hypothèse générale, aucune information *a priori* n'est disponible sur les documents analysés. En fonction des documents traités, cette contrainte n'est pas toujours aussi stricte. Pour des documents de type réunion, nous pouvons connaître par exemple le nombre de participants à partir d'une liste de présence. Dans le cas de journaux télévisés, nous pouvons avoir des données de référence pour certains locuteurs comme le présentateur du journal, les reporters et certaines personnalités très connues. Une annotation manuelle partielle du document à traiter peut également être disponible dans certains cas.

En fonction des résultats obtenus par notre système qui n'utilise aucune information *a priori*, nous allons donc nous intéresser à l'apport de ces informations *a priori* sur le résultat de la segmentation. Pour pouvoir utiliser de telles informations, nous devons connaître les types de données à traiter mais aussi les points forts et les points faibles de notre système. Si par exemple, notre système détecte déjà bien le nombre de locuteurs présents dans un enregistrement une telle information connue *a priori* lui sera inutile.

La deuxième partie exploratoire de ce travail concernera l'évaluation de l'apport d'un système de segmentation en locuteurs dans un contexte de recherche d'information sur une collection de documents vidéo. Nous nous intéresserons également à l'interaction entre les informations audio et vidéo et leur exploitation pour la segmentation en locuteurs sur un document vidéo. Ainsi, nous essayerons de montrer que l'utilisation du flux vidéo peut aider à améliorer un système de segmentation en locuteurs initialement basé sur le canal audio seul.

Finalement, pour aller encore plus loin, nous évaluerons notre système de segmentation en locuteur qui extrait ses informations uniquement de la piste audio, dans une tâche multi-modale de segmentation en histoires de documents vidéo.

5. Le manuscrit

Au vu de la problématique et de notre méthodologie, nous proposons l'organisation de manuscrit suivante.

Une première partie sera dédiée à l'état de l'art. Dans un premier temps, les principales techniques existantes en reconnaissance du locuteur seront détaillées. Ensuite, nous présenterons les principales étapes d'un système de segmentation en locuteurs. La fin de cette partie sera dédiée à la présentation des différentes architectures de systèmes de segmentation en locuteurs existantes dans la littérature.

La deuxième partie sera dédiée à la présentation de notre système de référence. Ce système sera ensuite amélioré au niveau de la paramétrisation mais aussi en proposant une méthode d'estimation du nombre de locuteurs qui à donné des très bons résultats lors des évaluations NIST RT 2003 et RT 2004. Les premiers résultats officiels pour les évaluations NIST et ESTER seront également présentés.

La troisième partie, plus exploratoire, concernera l'utilisation d'informations *a priori* pour améliorer un système de segmentation en locuteurs. Nous allons étudier l'influence sur notre système des informations issues d'une annotation manuelle comme le nombre de

locuteurs ou une courte annotation incomplète. Nous chercherons aussi à déterminer l'apport d'informations issues d'une annotation automatique comme une pré-segmentation en classes acoustiques (parole/musique/genre etc.) ou un résultat de segmentation en locuteurs issu d'un autre système.

La quatrième partie sera dédiée à l'utilisation de la segmentation en locuteurs, qui est un traitement uniquement sur l'audio, pour l'indexation d'une grande collection de documents vidéo issus des évaluations TREC. Dans un premier temps, nous nous intéressons à la recherche de documents vidéo selon des critères fondés sur l'audio pur mais aussi conjointement sur l'audio et la vidéo. Dans un second temps, nous essayerons d'évaluer l'apport de la vidéo pour la segmentation en locuteurs. Par exemple, nous proposerons une méthode originale d'extraction, sans modèle du locuteur *a priori*, de plans vidéo contenant le présentateur d'un journal télévisé. Nous finirons avec l'intégration de notre système de segmentation en locuteurs dans une tâche plus générale de segmentation en histoires d'un document vidéo, en utilisant simultanément des données audio, vidéo et texte. Le document se terminera par la conclusion et les perspectives, suivies par les annexes.

Première partie :

Etat de l'art

Dans cette partie nous présentons d'abord les principaux outils de traitement de la parole utilisés en segmentation en locuteurs. L'architecture la plus utilisée en segmentation en locuteurs a été proposée par H. Gish de la société américaine BBN dans le début des années 90. Les principaux modules ce cette architecture sont présentés avec leurs différentes implémentations existantes. Des variantes de cette architecture apparues récemment sont également présentées dans cette chapitre. La fin de ce chapitre est dédiée à une courte présentation de la tâche de suivi du locuteur qui permet d'enrichir le résultat d'un système de segmentation en locuteurs.

1. Architecture générale d'un système de segmentation

Le principal but de la segmentation dans le domaine du traitement de la parole est de décrire un signal audio donné. La segmentation peut être faite en thèmes [Chaisorn 2003], en qualité (bande large/bande étroite, bruité/non bruité) [Lamy 2003], en parole/non parole [Parlangeau 2003], en locuteurs, etc. Dans le cas de la segmentation en locuteurs, le but est de découper le flux audio en segments précisant pour chacun le début, la fin ainsi que l'étiquette du locuteur auquel il correspond.

Les travaux sur la segmentation automatique des documents audio sont pour la plupart fondés sur des méthodes de traitement du signal et sur des approches statistiques.

Les premiers travaux sur la segmentation en locuteurs ont été faits par Herbert Gish dans les années 1990 [Gish 1991]. Avant l'apparition de la nécessité d'utiliser la segmentation pour l'accès à des grandes bases de données multimédia, ce problème a été abordé par la société américaine BBN pour une application liée au trafic aérien. Le but de l'application était d'améliorer le trafic aérien dans un aéroport. Celle-ci consistait à segmenter les conversations enregistrées entre les pilotes et les contrôleurs de vols dans une tour de contrôle de l'aéroport. En effet, un contrôleur peut utiliser une même fréquence pour parler à plusieurs pilotes. Il est donc possible que le contrôleur parle à plusieurs pilotes dans une même conversation. Les résultats attendus de l'application étaient la détermination des ordres donnés par le contrôleur à chaque pilote.

Herbert Gish propose dans [Gish 1991] une architecture reprise ensuite par la plupart des systèmes de segmentation [Reynolds 2003, Nguyen 2003a]. Cette architecture se compose de plusieurs étapes :

- Paramétrisation du signal de parole (calcul de paramètres acoustiques) ;
- Pré-segmentation acoustique (l'enregistrement est découpé en segments de parole, silence, musique, etc.) ;
- Détection de changements de locuteur ;
- Regroupement des segments (les segments trouvés pendant la phase précédente sont regroupés et le nombre de locuteurs est estimé).

Dans certain cas, les chercheurs ont ajouté à la fin une phase de re-segmentation [Reynolds 2002, Meignier 2001] pour augmenter la précision au niveau des frontières des segments.

Dans cette architecture les hypothèses de travail précisées dans le chapitre antérieur sont respectées, notamment la segmentation en locuteurs n'utilise aucune information sur les locuteurs et leur nombre qui est inconnu doit être déterminé de façon automatique. Cette architecture constitue la base des systèmes proposés dans la littérature toutefois, nous allons voir dans la suite qu'il existe des variations de cette architecture. Certaines parties peuvent être supprimées ou remplacées, d'autres comme la paramétrisation sont indispensables.

2. Outils pour le traitement de la parole utilisés en segmentation

Cette section décrit les outils utilisés dans le domaine du traitement de la parole qui vont être employés ensuite pour la segmentation en locuteurs. Nous allons décrire les paramètres acoustiques, la modélisation du locuteur et la théorie de la décision appliquée à la segmentation en locuteurs.

2.1. Paramétrisation

Comme nous avons vu, la première étape du processus de segmentation en locuteurs est la paramétrisation ou le calcul des paramètres acoustiques à partir du signal de parole. Le signal de parole n'est pas directement utilisable à cause de sa grande complexité (grande diversité d'information) et de son caractère redondant. Le but de la paramétrisation est d'extraire l'information pertinente pour la tâche proposée.

La première étape de la paramétrisation acoustique consiste à découper le signal de parole en fenêtres de taille fixe (variable de 20 ms à environ 40 ms) reparties de façon uniforme le long du signal (toutes les 10 ms). La taille des fenêtres est choisie en considérant que les propriétés du conduit vocal peuvent être considérées comme invariables sur une petite durée égale à la taille de la fenêtre. Le signal audio est donc considéré comme stationnaire sur la durée de la fenêtre.

Il existe de nombreux paramètres utilisés dans la littérature. Plusieurs classifications peuvent être envisagées mais trois catégories principales peuvent être considérées :

- les paramètres spectraux : transformée de Fourier, bancs de filtres;
- les paramètres temporels : le taux de passage par zéro, le débit d'élocution;
- les paramètres cepstraux : LPCC, LFCC, MFCC.

L'énergie est aussi un paramètre important dans le traitement de la parole. Celle-ci peut être considérée, selon la façon dont elle est calculée, comme un paramètre spectral ou

temporel (le *théorème de Parceval* nous indique qu'un calcul d'énergie peut aussi bien être fait dans le domaine spectral que temporel).

Les paramètres acoustiques les plus utilisés pour la caractérisation du locuteur sont les coefficients cepstraux, notamment les coefficients MFCC. Parmi les systèmes présentés pendant les évaluations NIST [NIST 2002] en reconnaissance de locuteur et en segmentation, quasiment tous ont utilisé des coefficients cepstraux et une grande majorité [Liu 2002, Navratil 2002, Reynolds 2002] a utilisé les coefficients MFCC.

Les paramètres cepstraux [Calliope 1989, Rabiner 1993] sont utilisés pour séparer l'influence de la source d'excitation vocale et celle du conduit vocal, cette dernière étant généralement la seule à être utilisée dans le domaine de la reconnaissance du locuteur. L'influence de la source est considérée [Atal 1976] comme n'étant pas assez fiable pour être utilisée dans un système de reconnaissance du locuteur. Cependant, elle peut être associée à d'autres paramètres pour améliorer le résultat du système de reconnaissance [Furui 1994].

Les coefficients cepstraux [Furui 1981] sont utilisés dans le traitement du signal et sont des coefficients d'énergie calculés dans des bancs de filtres. Selon la distribution des filtres dans la bande utile du signal l'un des deux cas suivants peut être considéré :

- les filtres sont uniformément distribués dans la bande utile du signal. Dans ce cas les coefficients calculés sont appelés LFCC (Linear Frequency Cepstral Coefficients) ;
- les filtres sont distribués selon une échelle non-linéaire, plus proche de la perception humaine appelée échelle Mel. Dans ce cas les coefficients calculés sont les MFCC (Mel Frequency Cepstral Coefficients).

Etant donné que les coefficients LFCC sont similaires aux MFCC, sauf pour la distribution de filtres dans la bande utile du signal, nous allons décrire plus en détail ces derniers seulement.

2.1.1 Les coefficients MFCC

De nombreuses études ont montré que la perception humaine des sons ne suit pas une échelle linéaire [Calliope 1989], d'où l'idée de définir pour chaque valeur de fréquence f (en Hz) une hauteur subjective qui est mesurée sur une échelle "mel" (en mels). Comme point de référence la fréquence de 1 kHz correspond à 1000 mels.

Pour la correspondance entre la fréquence mel et la fréquence réelle en Hz la fonction la plus utilisée est :

$$F_{MEL} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \qquad \text{(I. 1)}$$

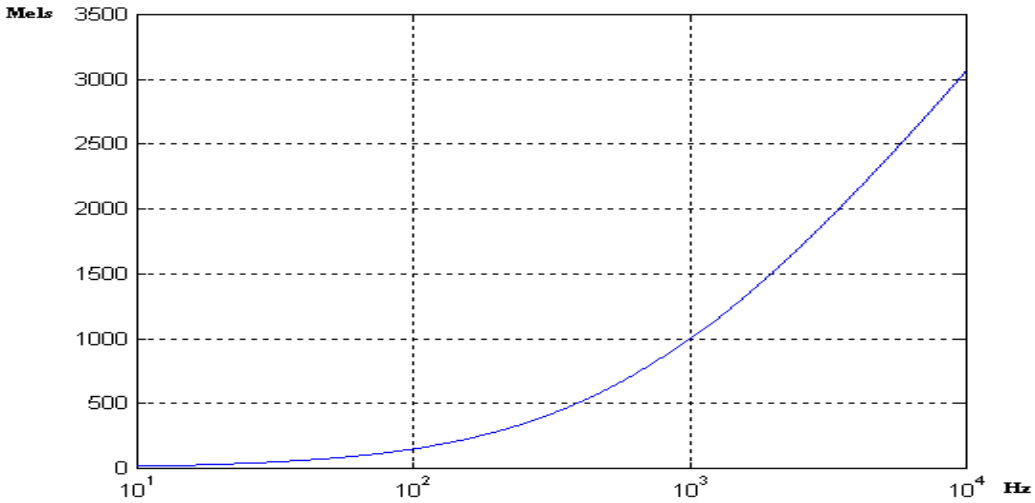


Figure I. 1 Correspondance mels / hertz

La Figure I. 1 montre la correspondance entre la fréquence en mels et la fréquence réelle en Hz.

Pour des raisons pratiques, le spectre mel est simulé en utilisant un banc de filtres triangulaires uniformément répartis sur l'échelle mel. La distance entre deux triangles est d'environ 150 mels et la largeur d'un triangle est d'environ 300 mels. Cependant, il existe des approches où ces deux paramètres ne sont pas respectés. Dans ce cas, les filtres sont uniformément distribués dans la bande passante du signal mesurée en Mels et pas en Hertz. Par exemple pour une bande passante de 8000 Hz ce qui fait environ 2840 Mels et pour 24 filtres la distance entre ces filtres est d'environ 120 Mels.

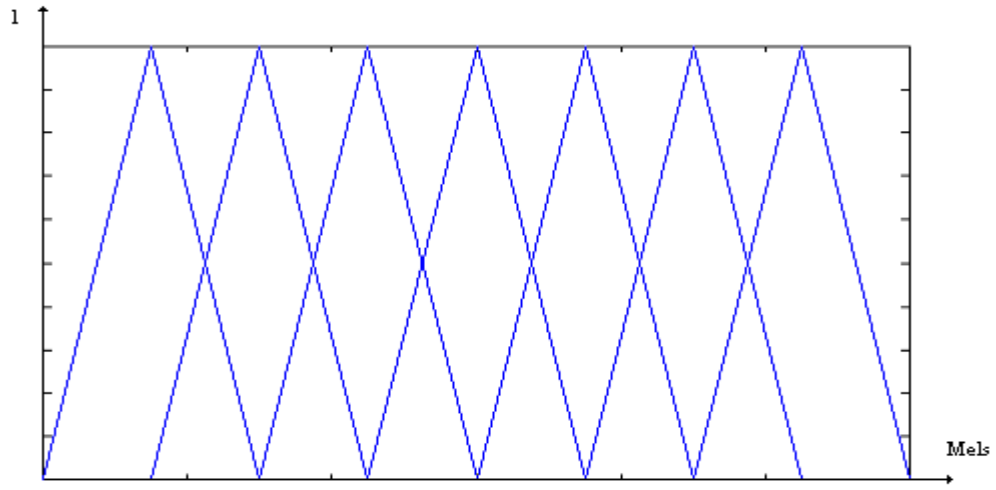


Figure I. 2 Banc de filtres MEL

Selon [Soong 1988] l'utilisation d'une analyse cepstrale à la place de la simple utilisation des amplitudes à la sortie des filtres produit des coefficients peu corrélés. Ceci est souhaitable si les coefficients sont ensuite utilisés par un système de reconnaissance basé sur des distributions gaussiennes avec des matrices de covariance diagonales.

Les coefficients MFCC [Rabiner 1993] d'une trame de parole sont calculés de la façon suivante :

Figure I. 3 Calcul des coefficients MFCC

1. Apres le filtrage de pré-accentuation, le signal de parole est d'abord découpé en fenêtres de taille fixe reparties de façon uniforme le long du signal.
2. La FFT (Fast Fourier Transform) de la trame est calculée. Ensuite, l'énergie est calculée en élevant au carré la valeur de la FFT. L'énergie est passée ensuite à travers chaque filtre Mel. Soit S_k l'énergie du signal à la sortie du filtre K, nous avons maintenant m_p (le nombre de filtres) paramètres S_k.
3. Le logarithme de S_k est calculé.

4. Finalement les coefficients sont calculés en utilisant la DCT (Discrete Cosine Transform)

$$c_i = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m_p} \log\left(S_k\right) \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{m_p} \right] \right\} \text{ pour } i = 1 \ldots N \qquad \text{(I. 2)}$$

où N est le nombre de coefficients MFCC que l'on souhaite obtenir.

2.1.2 La moyenne de coefficients cepstraux

L'une des informations contenues dans les coefficients cepstraux est l'information concernant le canal de transmission. Cette information est contenue dans la moyenne des vecteurs cepstraux [Atal 1974, Furui 1981] et peut poser des problèmes en particulier dans les tâches d'identification/vérification du locuteur.

La méthode la plus utilisée pour l'élimination de l'influence du canal de transmission est la suppression de la moyenne cepstrale (*Cepstral Mean Substraction* - CMS) [Furui 1994]. Pour la vérification du locuteur, la moyenne globale calculée sur tout le signal de test, est extraite. Dans le cas où il existe plusieurs locuteurs et plusieurs canaux de transmission sur le même signal de test (suivi du locuteur, etc.), la solution consiste à extraire une moyenne locale calculée à l'aide d'une fenêtre glissante [Reynolds 1996].

Cependant, l'information du canal de transmission peut s'avérer très utile dans une tâche de segmentation où la présence de deux canaux de transmission différents peut être synonyme de la présence de deux locuteurs différents.

2.1.3 Coefficients différentiels et énergie

Certains systèmes de reconnaissance du locuteur utilisent en plus des coefficients cepstraux appelés aussi statiques, des coefficients différentiels qui sont les dérivées du premier et du deuxième ordre des coefficients statiques. Les coefficients différentiels ne sont généralement pas utilisés dans les systèmes de segmentation mais plutôt utilisés dans les systèmes de reconnaissance du locuteur et de la parole. L'utilisation des coefficients différentiels, comme l'utilisation de la CMS, augmente la robustesse du système aux perturbations dues au canal de transmission [Reynolds 1996, Soong 1988]. Comme nous l'avons vu dans le paragraphe précédent, les perturbations dues au canal de transmission

peuvent au contraire être exploitées avantageusement pour un système de segmentation en locuteurs. Les coefficients différentiels sont donc peu utilisés en segmentation.

L'énergie du signal, par contre, est un paramètre souvent utilisé dans la segmentation. L'énergie contient l'information liée au niveau acoustique moyen du signal audio et dans une conversation, il est parfaitement possible qu'un locuteur parle plus fort qu'un autre, ou soit plus éloigné du micro qu'un autre par exemple.

2.2. La modélisation du locuteur

Les modèles utilisés pour la caractérisation du locuteur sont issus des modèles statistiques employés en reconnaissance de formes. Les modèles les plus utilisés dans la reconnaissance du locuteur sont les mixtures de gaussiennes GMM (Gaussian Mixture Models) qui ont été introduits en 1992 par Reynolds [Reynolds 1992].

Les deuxièmes modèles utilisés sont les modèles de Markov cachés HMM (Hidden Markov Models) qui sont largement employés en reconnaissance de la parole et qui ont été introduits en 1975 par [Backer 1975] et [Jelinek 1976]. Ces modèles sont généralement utilisés en reconnaissance du locuteur dans des conditions dépendantes du texte tandis que les modèles GMM sont utilisés dans des conditions indépendantes du texte.

Les GMM et les HMM apportent les meilleures performances en reconnaissance automatique du locuteur depuis au moins 1996 [Martin 2001, NIST 2002, NIST 2003]. Il existe bien d'autres modèles comme ceux issus de l'analyse prédictive, les réseaux de neurones ou les machines à support de vecteurs (SVM), cependant ils ne feront pas l'objet d'une description plus détaillée dans cette thèse.

2.2.1 Les mixtures de gaussiennes

Les GMMs modélisent les séquences de vecteurs acoustiques, correspondant dans notre cas à un locuteur donné, par une somme pondérée de distributions gaussiennes multidimensionnelles.

Une distribution gaussienne multidimensionnelle est définie par la formule suivante :

$$P(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right) \qquad \text{(I. 3)}$$

où d est la dimension d'un vecteur de paramètres noté x, μ le vecteur moyen et Σ la matrice de covariance estimée à partir de données $X = \{x_1, x_2, \ldots, x_N\}$ selon les formules suivantes :

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \Sigma = (X - \overline{X})(X - \overline{X})^T \tag{I. 4}$$

Un GMM est défini par un ensemble de T distributions gaussiennes et de T poids w_i associés à chaque distribution $P_i(x)$. Le nombre T est appelé l'ordre du modèle.

$$P(x) = \sum_{i=1}^{T} w_i P_i(x) \tag{I. 5}$$

L'utilisation de GMM pour la modélisation du locuteur est motivée par Reynolds [Reynolds 1992] par deux interprétations :

- chaque distribution dans un modèle GMM est capable de représenter la structure spectrale d'une large classe phonétique. Ces classes représentent des configurations du conduit vocal spécifiques au locuteur et donc utiles pour la modélisation du locuteur ;
- une mixture de distributions gaussiennes donne une représentation approximative de la distribution à long terme de vecteurs acoustiques provenant des énoncés du même locuteur.

En ce qui concerne la première interprétation, il faut tenir compte du fait qu'un GMM est entraîné de façon non supervisée, à partir de données non étiquetées du point de vue phonétique. En conséquence, les distributions gaussiennes ne sont pas contraintes directement à correspondre à des classes phonétiques.

Pour la deuxième interprétation, comme un modèle GMM est une somme pondérée de distributions gaussiennes, nous pouvons considérer que c'est un développement en série d'une distribution quelconque.

a. Estimation par maximum de vraisemblance (avec l'algorithme EM)

Un modèle est estimé à partir des données audio d'apprentissage. Pour cela on utilise l'algorithme EM (Expectation Maximization) [Dempster 1977] qui maximise un critère de

maximum de vraisemblance. L'estimation par maximum de vraisemblance (avec l'algorithme EM) est un outil très puissant pour estimer les paramètres d'un processus stochastique à partir d'un ensemble d'observations. L'algorithme EM présente l'avantage de donner une fonction de vraisemblance croissante après chaque itération. On doit trouver le modèle $\hat{M}$ le plus probable d'avoir produit un ensemble d'observations données X, soit le modèle qui maximise la probabilité d'observations X étant donné le modèle M :

$$\hat{M} = \arg\max_{M} P(X / M) \qquad \text{(I. 6)}$$

Dans le cas d'un GMM, les poids, les moyennes et les variances sont appris. L'algorithme EM est itératif, mais d'après Reynolds [Reynolds 1992], le maximum de vraisemblance est presque atteint après une quinzaine d'itérations.

Pour de bonnes performances, les conditions d'enregistrement des données d'apprentissage doivent être comparables aux conditions d'enregistrement des données de test (exemple : même bande passante, même bruit ambiant, etc.).

L'algorithme itératif EM estime à partir d'un modèle existant M_n (à l'itération n) un nouveau modèle M_{n+1} (à l'itération n+1) tel que les vraisemblances après chaque itération augmentent : $P(X / M_n) \geq P(X / M_{n+1})$. Le processus continue jusqu'à ce qu'un certain critère de convergence soit atteint (exemple : gain de vraisemblance inférieur à un seuil).

Si nous voulons estimer les paramètres d'un modèle GMM [Bilmes 1998, Reynolds 1992], nous ne pouvons pas utiliser directement l'équation I. 6, car celle-ci selon les deux auteurs ne garantit pas l'augmentation de vraisemblance à chaque itération. Pour cette raison les auteurs proposent d'utiliser ce qu'on appelle la fonction auxiliaire de Baum [Baum 1970], qui peut être vue comme une entropie de données d'entraînement étant donné le modèle. Le lecteur pourra trouver une démonstration mathématique détaillée concernant toutes les formules d'estimation de l'algorithme EM dans [Bilmes 1998]. Nous allons présenter ici seulement un court résumé de ces formules.

Soit M_{n+1} et M_n les modèles respectivement aux itérations n+1 et n. La fonction de Baum est définie par :

$$Q(M_{n+1}, M_n) = \sum_{l=1}^{T} \sum_{i=1}^{N} \log\left(w_{l,n+1} p\left(x_i / m_{l,n+1}\right)\right) p\left(l / x_i, M_n\right) \qquad \text{(I. 7)}$$

$$p\left(l / x_i, M_n\right) = \frac{w_{l,n} p\left(x_i / m_{l,n}\right)}{p\left(x_i / M_n\right)} \tag{I. 8}$$

où $m_{l,n}$ et $m_{l,n+1}$ sont les l composantes gaussiennes de chaque modèle et $p\left(l / x_i, M_n\right)$ est la probabilité *a posteriori* de la composante l et $p\left(l / m_{l,n}\right)$ est la probabilité du vecteur x_i par rapport à la composante mono-gaussienne l, soit :

$$p\left(x_i / m_{l,n}\right) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_{l,n}\right|}} \exp\left(-\frac{1}{2}\left(x_i - \mu_{l,n}\right)^T \Sigma_{l,n}^{-1}\left(x_i - \mu_{l,n}\right)\right) \tag{I. 9}$$

Pour déterminer w_l le poids de chaque distribution dans le modèle on doit rajouter la contrainte suivante : $\sum_{l=1}^{T} w_l = 1$. En conséquence, pour trouver les paramètres w_l il faut dériver (en tenant comte du multiplicateur de Lagrange) :

$$\frac{\partial}{\partial w_l}\left(Q\left(M_{n+1}, M_n\right) + \lambda\left(\sum_{l=1}^{T} w_l - 1\right)\right) = 0 \tag{I. 10}$$

Pour les autres paramètres, il suffit de dériver directement la fonction de Baum de l'équation I. 7. Finalement, après les calculs, les paramètres du modèle M_{n+1} en fonction des paramètres du modèle M_n sont donnés par les équations suivantes :

$$w_{l,n+1} = \frac{1}{N}\sum_{i=1}^{N} p\left(l / x_i, M_n\right) \tag{I. 11}$$

$$\mu_{l,n+1} = \frac{\sum_{i=1}^{N} x_i p\left(l / x_i, M_n\right)}{\sum_{i=1}^{N} p\left(l / x_i, M_n\right)} \tag{I. 12}$$

$$\Sigma_{l,n+1} = \frac{\sum_{i=1}^{N} p\left(l / x_i, M_n\right) x_i x_i^T}{\sum_{i=1}^{N} p\left(l / x_i, M_n\right)} - \mu_{l,n+1}\mu_{l,n+1}^T \tag{I. 13}$$

Nous constatons que dans les équations précédentes un rôle important est joué par les probabilités *a posteriori* $p(l / x_i, M_n)$. Ces probabilités sont utilisées par exemple si on regarde l'équation I. 8 pour régler l'influence de chaque vecteur d'entraînement sur chaque composante du modèle.

Dû au fait que l'algorithme EM est assez ancien et bien décrit du point de vue théorique, il existe peu d'améliorations récentes sur l'estimation des paramètres. Les essais d'amélioration sont concentrés sur l'une des options suivantes : faire une pré-sélection de données d'entraînement (exemple : ne pas utiliser les trames de faible énergie [Besacier 1998]), imposer des contraintes sur les différents paramètres (exemple : limitation de la variance), ou bien choix du modèle d'initialisation. Le choix du modèle d'initialisation est important parce que la fonction de Baum utilisée dans l'algorithme atteint généralement en pratique un maximum local et non pas un maximum absolu [Reynolds 1992].

b. Adaptation par Maximum A Posteriori

L'estimation par maximum de vraisemblance donne des bons résultats à condition que la quantité de données soit suffisamment grande pour donner une estimation robuste des paramètres du modèle [Gauvain 1994]. Dans le cas où la quantité de données n'est pas suffisante pour apprendre un modèle GMM directement par l'algorithme EM les méthodes d'adaptation d'un modèle sont utilisées. Cette approche consiste à adapter un modèle générique appelé "modèle du monde" à la séquence d'apprentissage. Le modèle du monde est généralement appris par l'algorithme EM sur une grande quantité de données appartenant à plusieurs locuteurs. Il peut être uni-condition ou multi-condition. Dans le premier cas les données d'apprentissage contiennent une seule et unique condition acoustique (e. g. un modèle de femme, une condition de conversation téléphonique). Si la condition du modèle du monde et celle des données de test sont les mêmes, les résultats sont meilleurs en utilisant un modèle uni-condition. Ce modèle permet au système de reconnaissance de mieux s'adapter aux données de test.

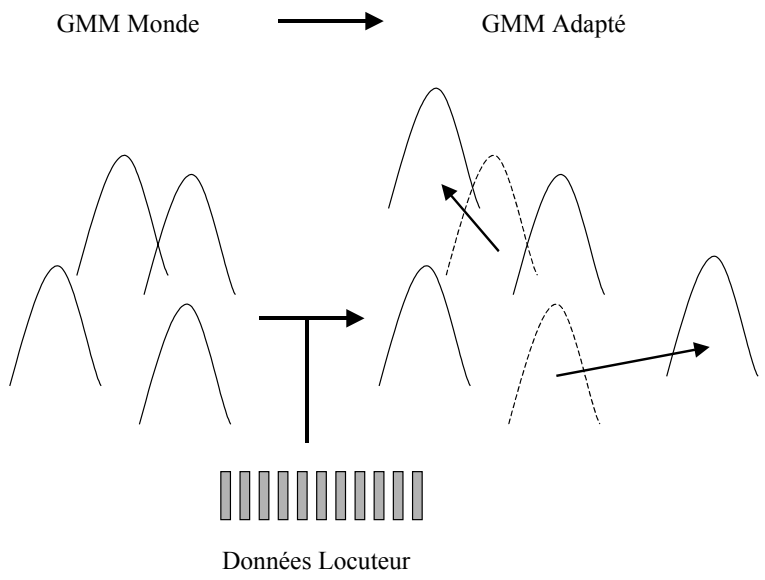


Figure I. 4 Adaptation d'un modèle du monde (GMM avec quatre gaussiennes) sur un locuteur particulier. Deux gaussiennes ont changé de moyenne

L'adaptation est faite généralement soit par une méthode MAP (Maximum A Posteriori) [Gauvain 1994] soit par une méthode MLLR (Maximum Likelihood Linear Regression) [Leggetter 1995], la différence entre les deux méthodes étant au niveau de l'estimateur.

La méthode d'adaptation la plus utilisée en reconnaissance du locuteur est la méthode MAP avec certaines variantes [Gauvain 1994, Reynolds 2000]. L'adaptation par la méthode MAP est un moyen d'introduire des informations *a priori* dans le processus d'entraînement, particulièrement utile dans le cas de faible quantité de donnés d'entraînement. La différence principale entre les méthodes MAP et EM réside dans l'hypothèse d'une distribution *a priori* sur les paramètres du modèle estimé. Contrairement à la formule I. 6 cette fois nous voulons maximiser :

$$\hat{M} = \arg\max_{M} P(M / X) = \arg\max_{M} P(X / M) P(M) \qquad \text{(I. 14)}$$

En considérant qu'aucune information *a priori* sur le modèle n'est disponible, c'est-à-dire que la densité de probabilité $P(M)$ est une constante, on retrouve dans ce cas l'estimateur utilisé dans l'algorithme EM.

L'utilisation de l'adaptation MAP pour l'adaptation du modèle du monde sur un locuteur particulier revient à dire que les paramètres *a priori* du modèle adapté sont les

paramètres du modèle du monde. Dans le cas d'une adaptation, les poids, les moyennes ou l'ensemble des paramètres peuvent être adaptés.

Contrairement à l'algorithme EM, les paramètres du nouveau modèle adapté à partir du modèle du monde sont combinés avec les paramètres du modèle du monde. En pratique [Reynolds 2000] pour simplifier les calculs on utilise un coefficient global α (inférieur à 1) pour donner plus ou moins d'importance aux données d'adaptation par rapport aux données du monde :

$$w_{l,A} = \left[(1-\alpha) w_{l,M} + \alpha \ \frac{1}{N} \sum_{i=1}^{N} p\left(l / x_i, M_n\right) \right] \gamma \qquad \text{(I. 15)}$$

$$\mu_{l,A} = (1-\alpha) \mu_{l,M} + \alpha \frac{\sum_{i=1}^{N} x_i p\left(l / x_i, M_n\right)}{\sum_{i=1}^{N} p\left(l / x_i, M_n\right)} \qquad \text{(I. 16)}$$

$$\alpha = \frac{\sum_{i=1}^{N} p\left(l / x_i, M_n\right)}{r + \sum_{i=1}^{N} p\left(l / x_i, M_n\right)} \qquad \text{(I. 17)}$$

Les deux équations donnent les valeurs des paramètres m_A et μ_A (moyenne et poids) pour le modèle adapté. Le paramètre γ est calculé sur tous les poids pour s'assurer que la somme de tous les poids reste égale à 1. Le paramètre α est calculé pour chaque composante du modèle GMM. Le paramètre r est appelé facteur de relevance et donne la quantité de données nécessaire pour adapter une gaussienne. La première partie de chaque équation représente l'influence du modèle du monde et la deuxième partie l'influence des données du locuteur. Nous constatons que si une gaussienne l n'attire pas des données, alors les probabilités $p\left(l / x_i, M_n\right)$ sont très faibles et en conséquence α peut avoir des valeurs proches de 0. Il existe le cas contraire où une gaussienne attire trop des données et α peut avoir des valeurs proches de 1. Les deux cas peuvent avoir des conséquences négatives sur le système de reconnaissance soit par le fait que le modèle est trop proche du modèle du monde, donc il ne représente pas le locuteur cible soit il se spécialise trop sur certaines données et il est incapable de généralisation.

Nous avons donné seulement les équations pour la moyenne et pour le poids parce que l'adaptation de la variance n'est pratiquement jamais utilisée. Selon [Reynolds 2000] l'adaptation de la variance n'améliore pas les résultats.

L'adaptation la plus rapide est évidement l'adaptation des poids seuls, mais selon [Reynolds 2000] l'adaptation de moyennes donne de meilleurs résultats.

c. Calcul de probabilité

Pour prendre une décision concernant une séquence de vecteurs acoustiques $X = \{x_1, x_2, \ldots, x_N\}$ par rapport à un locuteur, la probabilité que la séquence donnée ait été prononcée ou non par le locuteur concerné doit être calculée. En fait, seulement la vraisemblance de chaque observation $P(x_i / M)$, donnée par la formule suivante dans le cas des GMMs, est calculée :

$$P(x_i / M) = \sum_{j=1}^{T} w_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\left(-\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right) \qquad \text{(I. 18)}$$

$$P(X / M) = \sqrt[N]{\prod_{i=1}^{N} P(x_i / M)} \qquad \text{(I. 19)}$$

La vraisemblance de la séquence considérée est alors donnée par la moyenne géométrique des vraisemblances des observations (équation I. 18). L'utilisation de la moyenne géométrique est nécessaire pour réduire l'influence de la durée du segment de test sur la valeur de la vraisemblance $P(X / M)$. Le calcul de cette moyenne repose aussi sur l'hypothèse que les vecteurs acoustiques sont indépendants entre eux [Reynolds 2000]. Vu la fréquence de calcul de vecteurs acoustiques (environ tous les 10 ms) cette hypothèse n'est jamais vérifiée mais elle est utilisée tout de même pour faciliter les calculs.

d. Décision : Test d'hypothèse de Bayes

Dans le cas de la reconnaissance automatique du locuteur, le système doit répondre à la question : est-ce que la séquence $X = \{x_1, x_2, \ldots, x_N\}$ a été prononcée par le locuteur M. Dans le cas de la vérification du locuteur, le locuteur cible M est donné par l'utilisateur. Le système doit alors répondre à la question concernant un seul modèle. Dans le cas de l'identification du locuteur, le locuteur cible doit être trouvé parmi les locuteurs connus par le système. Le système choisi un locuteur parmi l'ensemble des modèles connus.

La réponse à la question est donnée par le test d'hypothèse de Bayes. Soit deux hypothèses possibles :

- H_0 : la séquence d'observations X a été prononcée par le locuteur M
- H_1 : la séquence d'observations X n'a pas été prononcée par le locuteur M

L'hypothèse H_0 est vraie si et seulement si le rapport de vraisemblance $lr(X)$ est supérieur à un seuil :

$$lr(X) = \frac{P(H_0)}{P(H_1)} = \frac{P(X/M)}{P(X/\overline{M})} > a \qquad \text{(I. 20)}$$

Le paramètre α est le seuil de décision et il dépend du coût du risque de prendre une mauvaise décision.

Le calcul de l'hypothèse H_0 est assez simple et il est équivalent à calculer $P(X/M)$. Par contre le calcul de l'hypothèse inverse H_1 n'est pas évident. En fait, il faut calculer $P(X/\overline{M})$. Le modèle $\overline{M}$ est un modèle appris sur des données qui n'appartiennent pas au locuteur cible ou il est remplacé par un ensemble de modèles de locuteurs différents du locuteur cible. Dans le cas de l'utilisation d'un ensemble de modèles de locuteurs $M_0, M_1, \ldots, M_J$ différents du locuteur cible on parle de *cohorte* de locuteurs, tandis que le cas où un seul modèle appris sur les données de plusieurs locuteurs est appelé *modèle du monde*.

Dans le cas d'une cohorte, la formule I. 19 devient :

$$lr(X) = \frac{P(H_0)}{P(H_1)} = \frac{P(X/M)}{f\left(P(X/M_0), P(X/M_1), \ldots, P(X/M_J)\right)} > a \qquad \text{(I. 21)}$$

où f est une fonction de vraisemblance de la séquence étant donné chaque locuteur de la cohorte (exemple : moyenne, maximum, etc.).

2.2.2 Les modèles de Markov cachés

Les modèles de Markov cachés (Hidden Markov Models ou HMMs) ont été introduits par Baum et ses collaborateurs dans les années 70 [Baum 1970] et ont connu un succès important en reconnaissance de la parole où ils se sont imposés comme l'un des modèles de

référence. Dans la reconnaissance du locuteur ils sont utilisés surtout en reconnaissance dépendante du texte où il existe une forte connaissance *a priori* sur l'énoncé du test.

Ces modèles sont issus de la théorie des automates probabilistes. Un automate probabiliste est défini par une structure composée d'états, de transitions et par un ensemble de distributions de probabilités sur chaque état.

Un HMM peut être défini comme un quadruplet :

- un ensemble d'états s_i avec π_i les probabilités initiales d'états ;
- un ensemble de probabilités de transitions entre les états $a_{i,j} = P(i/j)$ avec la contrainte que la somme de probabilités sortant d'un état est égale à 1 :

$$\sum_j a_{i,j} = 1 \qquad \text{(I. 22)}$$

- un alphabet de symboles (pas forcement fini) ;
- un ensemble de densités de probabilité (exemple : GMM) pour l'émission de symboles associés à chaque état $b_i(x)$.

Deux états supplémentaires, début et fin, peuvent être ajoutés pour permettre d'interconnecter des HMMs ou d'imposer des contraintes sur la séquence d'états. Ces deux états n'ont pas de densité de probabilité pour l'émission de symboles. Pour l'état début dans ce cas, les probabilités de transitions correspondent aux probabilités initiales π_i des autres états.

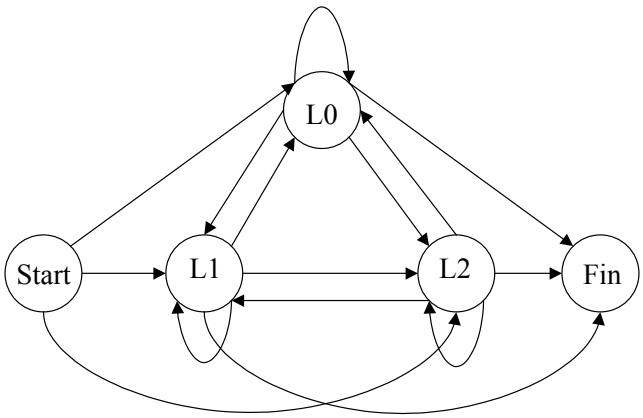


Figure I. 5 Un modèle HMM ergodique (toutes les transitions sont possibles)

Un des premiers domaines dans lequel les HMMs ont été utilisés est le domaine du traitement de la parole. Une majeure partie des techniques d'utilisation et d'implémentation

des HMMs a été développée dans le cadre de ces applications. Ces techniques furent ensuite appliquées avec succès à d'autres domaines comme la reconnaissance de caractères où l'analyse de séquences biologiques.

Le premier problème qui nous intéresse est l'entraînement d'un HMM à partir de données d'apprentissage. Comme dans le cas de modèles de mixtures de gaussiennes, où le nombre de gaussiennes est fixé au départ, la structure du modèle HMM doit être fixée (nombre d'états, ergodique ou non, etc.). Dans ce cas, on utilise l'algorithme de Baum-Welch, la version de l'algorithme EM pour les modèles de Markov [Rabiner 1989]. Si la structure du HMM est inconnue le modèle peut être construit de façon itérative [Meignier 2000, Freitag 1999]. Comme proposé dans [Meignier 2000] l'apprentissage commence avec un HMM à un état et *évolue* en générant à chaque itération un nouvel état. Ensuite l'algorithme de Baum-Welch est appliqué sur la structure générée à chaque itération.

Les formules d'apprentissage des distributions GMM associées à chaque état sont proches de celles pour les GMM à une différence près [Bilmes 1998]. Il faut remplacer les probabilités $p(l / x_i, M_n)$ par $p(l / x_i, M_n, s_k)$ qui sont conditionnées aussi par le fait de se trouver dans l'état s_k. Les probabilités initiales π_i et les probabilités de transitions $a_{i,j}$ sont estimées selon les formules suivantes, où s_t est l'état au moment de la génération du symbole x_t :

$$\pi_i = \frac{P(X, s_0 = i / H)}{P(X / H)} \tag{I. 23}$$

$$a_{i,j} = \frac{\sum_{t=1}^{N} P(X, s_{t-1} = j, s_t = i / H)}{\sum_{t=1}^{N} P(X, s_{t-1} = i / H)} \tag{I. 24}$$

Les principales applications des HMMs peuvent être classées en deux catégories. La première traite des problèmes de reconnaissance ou de classification comme par exemple la reconnaissance d'un mot parmi un ensemble de mots possibles à partir d'un signal audio (reconnaissance de la parole). Pour ces applications, un HMM différent est construit pour chaque classe à reconnaître par exemple un HMM pour chaque mot dans la reconnaissance de la parole. La reconnaissance d'une séquence consiste alors à calculer la probabilité de génération de la séquence par chacun des HMMs, et à assigner à cette séquence la classe la

plus probable. Ce type d'application nécessite donc un algorithme efficace pour le calcul de la probabilité de génération d'une séquence.

Le second type d'application des HMMs est constitué de problèmes de segmentation c'est-à-dire de découpage d'une séquence en sous-séquences de différents types comme par exemple découpage d'un signal audio en zones contenant une certaine caractéristique (parole/silence). Pour ces applications, un HMM dont les états sont typés est généralement utilisé. La procédure de segmentation d'une séquence de symboles consiste alors à calculer à l'intérieur du HMM le chemin (la suite d'états) qui a la probabilité maximale de générer cette séquence. Ensuite, à chaque symbole de la séquence est associé son type, en fonction du type de l'état correspondant dans le chemin calculé.

L'approche directe pour calculer la probabilité de générer une séquence de symboles $X = \{x_1, x_2, \ldots, x_N\}$ par un HMM H, consiste à calculer la probabilité de génération pour chaque chemin possible et à faire la somme de ces probabilités, c'est l'algorithme forward-backward [Rabiner 1993, Brehélin 2000].

$$P(X/M) = \sum_{<S>} P(X/S) \qquad \text{(I. 25)}$$

La probabilité de générer la séquence X suivant le chemin $S = s_{start} s_1 \ldots s_N s_{fin}$ est :

$$P(X/M) = P(s_1/s_{debut}) P(x_1/s_1) \ldots P(s_N/s_{N-1}) P(x_N/s_N) P(s_{fin}/s_N) \qquad \text{(I. 26)}$$

$$P(X/M) = \pi_1 b_1(x_1) \ldots a_{N,N-1} b_N(x_N) a_{fin,N} \qquad \text{(I. 27)}$$

Généralement la séquence d'états est inconnue. Le problème qu'on se pose ici pour une application de type décodage, est de trouver, étant donné une séquence de symboles $X = \{x_1, x_2, \ldots, x_N\}$ et un HMM H, la séquence d'états qui à la probabilité maximale de générer X, c'est l'algorithme de décodage de Viterbi [Viterbi 1967]. Ce n'est pas la probabilité maximale qui nous intéresse mais le chemin appelé *chemin de Viterbi* qui permet de générer la séquence de probabilité maximale. L'approche directe consiste à calculer la probabilité de génération suivant tous les chemins possibles et choisir le chemin qui présente la probabilité la plus élevée.

$$P(X/M) = \max_S P(X/S) \qquad \text{(I. 28)}$$

Dans les deux cas d'applications, les probabilités désirées ne sont pas directement calculables et il existe des algorithmes (l'algorithme forward-backward et l'algorithme de Viterbi) [Rabiner 1993, Bilmes 1998] qui font les calculs nécessaires de façon bien plus efficace.

Nous allons voir dans la section 5.2 que le décodage de l'algorithme Viterbi peut être utilisé pour faire de la segmentation en locuteurs en utilisant un HMM avec un locuteur par état.

2.2.3 La quantification vectorielle

La quantification scalaire consiste à représenter une valeur d'un échantillon de signal pas forcement audio avec une précision réduite, par exemple la représenter avec une valeur appartenant à un ensemble plus petit que l'ensemble original. C'est le cas typique de la conversion analogique/digitale.

Lorsque ce principe est appliqué par bloc d'échantillons (*vecteurs*), on peut parler de quantification vectorielle. La quantification vectorielle est alors une généralisation de la quantification scalaire. Mais, pendant que la quantification scalaire est dans sa forme la plus simple juste une conversion analogique/digitale, la quantification vectorielle est une méthode de codage/compression puissante.

La quantification vectorielle est souvent utilisée dans les télécommunications pour le codage de la source, ou dans la compression des données notamment dans la compression des images. Elle est aussi un puissant outil de classification.

La quantification vectorielle est définie par un doublet : un ensemble de vecteurs représentatifs appelés mots $C = c_1 c_2 \ldots c_M$, qui forme un dictionnaire (*codebook* en anglais) et un critère de distorsion $d(.,.)$. Un système de quantification vectorielle est présenté dans la Figure I. 6. Le codeur réalise une correspondance entre un vecteur d'entrée et un mot du dictionnaire en minimisant la distorsion donnée par le critère considéré $d(.,.)$. Le mot du dictionnaire est transmis à travers le canal de transmission et le décodeur réalise l'opération inverse du décodeur, c'est-à-dire obtient une approximation du vecteur initial.

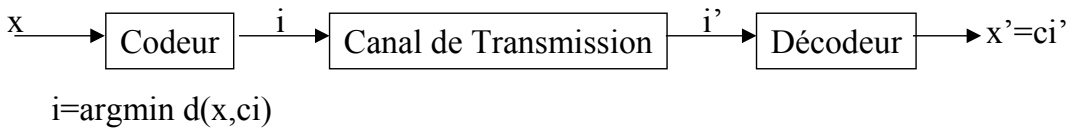


Figure I. 6 Quantification vectorielle pour un système de télécommunication

Il existe une classe particulière de quantificateurs vectoriels qui sont complètement définis par l'ensemble dictionnaire/mesure de distorsion appelés quantificateurs des plus proches voisins. Un quantificateur vectoriel de ce type est souvent utilisé comme classificateur. Dans ce cas, l'espace des vecteurs est divisé en cellules (classes) S_K définies par :

$$S_K = \{x \mid d(x, c_k) < d(x, c_j),\ j = 1 \ldots M,\ j \neq k\} \qquad \text{(I. 29)}$$

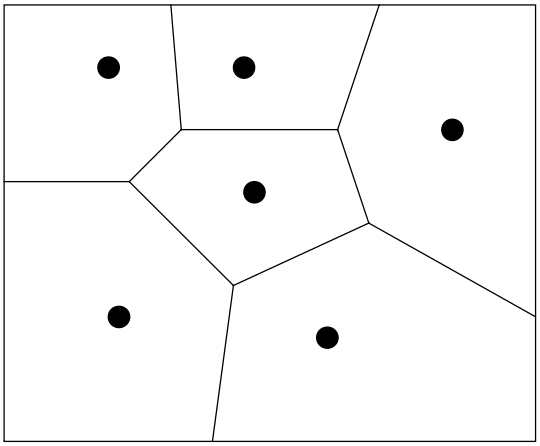


Figure I. 7 Résultat d'un classificateur de plus proches voisins

Dans le cas de données audio, le résultat de la paramétrisation du signal de parole (LFCC, MFCC, Banc des Filtres) est une série temporelle de vecteurs. La quantification vectorielle est basée sur le fait que la représentation finale d'une classe acoustique (locuteur, genre, phonème, etc.) peut être réduite à seulement un seul ou quelques vecteurs caractéristiques. Les vecteurs choisis pour représenter une classe acoustique, appelés mots ou "centroides" constituent le dictionnaire.

Les principaux avantages de la quantification vectorielle sont : espace réduit nécessaire pour le stockage du dictionnaire (élément non négligeable pour certaines implémentations pratiques) et puissance de calcul réduite (le calcul de similarité est réduit à seulement un calcul de distance entre les vecteurs analysés et le dictionnaire).

Le principal désavantage est une distorsion spectrale inhérente dans la représentation des vecteurs d'analyse. Comme le nombre de mots est fini, le processus consistant à choisir le meilleur mot pour représenter un vecteur introduit, par définition, une certaine erreur de quantification.

Pour construire un dictionnaire, l'algorithme le plus connu est l'algorithme "Binary Split" [Linde 1980] et il est implémenté par la procédure suivante :

- faire un dictionnaire avec un seul mot. Ce mot représente tous les vecteurs d'entraînement ;
- doubler la taille du dictionnaire par division de chaque mot suivant la règle :

$$\begin{aligned} y_n^+ &= y_n(1+\varepsilon) \\ y_n^- &= y_n(1-\varepsilon) \end{aligned} \qquad \text{(I. 30)}$$

où n varie de 1 à la taille du dictionnaire et ε est un paramètre généralement largement inférieur à 1 ;

- utiliser un algorithme (par exemple le plus proche voisin) pour attribuer chaque vecteur d'entraînement à un mot et modifier les mots en conséquence ;
- réitérer les deux dernières étapes jusqu'à atteindre la taille désirée du dictionnaire.

La classification dans le cas de la quantification vectorielle est une recherche dans le dictionnaire pour trouver le meilleur mot. Soit un vecteur d'analyse x, alors l'index m du meilleur mot pour quantifier x est le mot le plus proche et il est donné par :

$$m = \arg\min_j d(x, c_j) \qquad \text{(I. 31)}$$

Selon [Thyes 2000] les performances d'un système de reconnaissance du locuteur qui utilise des modèles GMM ou HMM se détériorent de façon drastique dans le cas où il y a peu de données disponibles pour l'entraînement. Dans ce cas, la quantification vectorielle peut être une solution alternative et peu coûteuse en termes de temps de calcul.

En ce qui concerne la segmentation, parmi les participants à la tache segmentation en locuteurs dans le cadre des évaluations NIST il existe un seul système présenté qui utilise la quantification vectorielle : le système Panasonic [Nguyen 2003a]. La quantification vectorielle est utilisée dans le calcul d'une normalisation géométrique différentielle (SWAMP Sweeping Metric Parameterization) de paramètres acoustiques MFCC [Nguyen 2003b]. A

part la paramétrisation originale, les autres étapes de ce système (pré-segmentation acoustique, etc.), sont les étapes classiques d'un système de segmentation.

3. Segmentation en locuteurs

La segmentation en locuteurs d'un document audio a comme but de couper un flux audio donné en interventions pas forcement continues contenant chacune la voix d'un seul et même locuteur. Il faut répondre à la question qui parle et quand sur un document audio. Dans les évaluations NIST la tâche de segmentation en locuteurs est appelée aussi "Who Spoke When" (Qui parle et quand) ou "Speaker Diarization" (Journalisation en locuteurs) [Tranter 2004].

A part l'étape de paramétrisation qui a été décrite dans le paragraphe 2.1, un système de segmentation contient trois étapes comme il a été précisé au début :

- pré-segmentation acoustique : c'est la partie qui vise à détecter les segments contenant uniquement de la parole et à éliminer les segments qui contiennent du silence, de la musique, du bruit, etc. ;
- détection de changements de locuteur : c'est une pré-segmentation en segments *uni-locuteur* ;
- regroupement : c'est l'étape qui vise à regrouper les segments issus de la détection de ruptures par locuteurs.

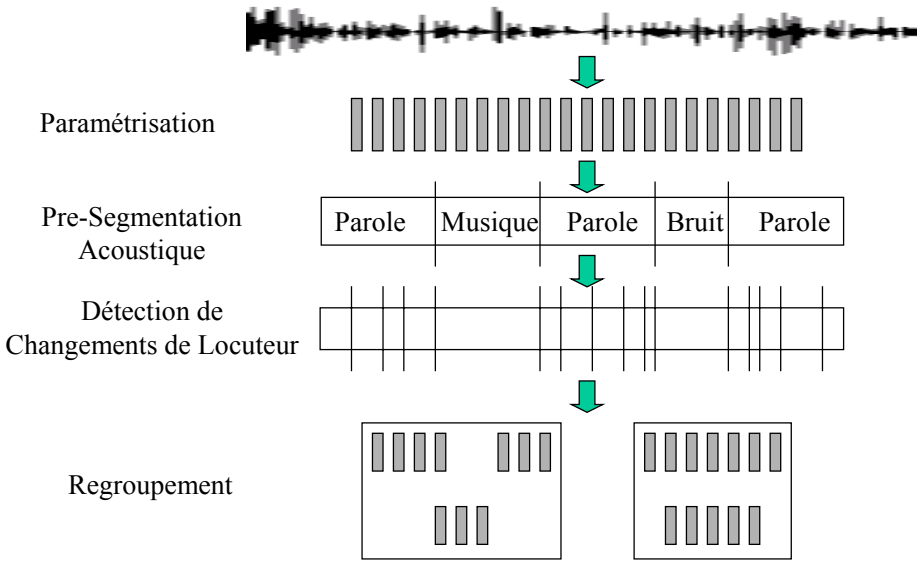


Figure I. 8 Architecture générale d'un système de segmentation

L'étape de pré-segmentation acoustique n'est pas toujours nécessaire notamment dans le cas de documents où la présence d'événements acoustiques autre que la parole est peu probable comme certaines conversations téléphoniques et réunions.

Les trois étapes présentées ne sont pas toujours évidentes à discerner dans l'architecture du système. Par exemple, le système de segmentation de l'Université de Cambridge [Tranter 2003a] n'utilise pas de détection de changements de locuteur. Ils font l'hypothèse qu'un changement de conditions acoustiques (au moins l'apparition d'un segment de silence) est suffisante pour détecter un changement de locuteur. [Hain 1998] précise que le nombre de détections manquées reste assez faible avec une telle stratégie mais les expériences reportées ne concernent que des documents issus de journaux télévisés (parole relativement préparée et contrôlée).

Le système du Laboratoire d'Informatique d'Avignon (LIA) [Meignier 2001] n'utilise pas de regroupement. A la place, un modèle HMM avec un état par locuteur est utilisé pour appliquer l'algorithme de décodage de Viterbi.

3.1. Pré-Segmentation acoustique

Dans cette section nous allons décrire les paramètres et les méthodes de classification concernant la pré-segmentation acoustique. Pour décrire un document en terme de locuteurs qui parlent il faut d'abord sélectionner les parties du document qui contiennent uniquement de la parole.

La plus simple pré-segmentation acoustique est la pré-segmentation parole/silence. Cependant elle n'est pas suffisante notamment pour les enregistrements de journaux télévisés qui contiennent entre autres de la musique et du bruit.

Plusieurs méthodes de pré-segmentation acoustique, notamment en ce qui concerne la discrimination parole/musique, ont été décrites dans la littérature. Elles peuvent se classer en deux grands groupes [Pinquier 2003b] :

- d'une part dans la communauté spécialiste de la musique, l'accent porte sur les paramètres permettant de séparer au mieux la musique du reste. C'est ce qu'on appelle la séparation musique/non musique. Les exemples de paramètres utilisés dans ce cas sont : le taux de passage par zéro [Saunders 1996], le centroïde spectral [Scheirer 1997], etc. ;

- d'autre part dans la communauté spécialiste du traitement automatique de la parole, les paramètres cepstraux sont utilisés pour séparer au mieux la parole du reste. C'est ce qu'on appelle la séparation parole/non parole. Ces paramètres [Scheirer 1997, Carey 1999] sont généralement les mêmes que ceux utilisés dans la reconnaissance de la parole et du locuteur, comme les coefficients cepstraux, l'énergie, la fréquence fondamentale, etc.

Il existe aussi des approches combinées [Pinquier 2003a] qui essayent d'exploiter les avantages de chaque approche pour arriver à un résultat optimal.

En ce qui concerne la classification, trois approches sont souvent utilisées: les modèles de mélanges de lois de gaussiennes, les k plus proches voisins et les modèles de Markov cachés.

Une bonne étude en ce qui concerne les différents paramètres peut être trouvée dans [Carey 1999]. Comme les paramètres de type cepstraux ont déjà été décrits, nous décrivons dans la section suivante les paramètres utilisés en pré-segmentation acoustique par la communauté spécialiste de la musique.

3.1.1 Approche utilisant des paramètres acoustiques autres que les coefficients cepstraux

Dans cette partie, nous présentons d'abord les paramètres classiques comme la fréquence fondamentale f0, le taux de passage par zéro, l'énergie mais aussi des paramètres inhabituels : la modulation de l'énergie à 4 Hz, la modulation de l'entropie, etc.

Les sons de la parole dits voisés (voyelles et certaines consonnes) sont produits avec vibration des cordes vocales ou vibration laryngienne [Calliope 1989]. La fréquence fondamentale f0, au sens de l'analyse de Fourier, est définie comme la fréquence de vibration laryngienne. En ce qui concerne la parole, la fréquence fondamentale moyenne apporte une mesure globale de la hauteur de la voix (aiguë, grave, voix de femme, voix d'homme, etc). Ses variations lentes caractérisent aussi la prosodie.

En ce qui concerne l'utilisation de la fréquence fondamentale pour la discrimination parole/musique certains travaux ont montré [Carey 1999] des différences au niveau de la variation de la première dérivée de la courbe f0. La musique présente une forte concentration aux alentours du zéro de la première dérivée due aux notes de musique stables et des variations importantes de fréquence fondamentale qui correspondent aux passages d'une note

de musique à une autre. La parole, par contre, montre une distribution de changements de fréquence fondamentale moins régulière due à l'alternance de parties voisées et non-voisées.

Le taux de passage par zéro a été utilisé par [Saunders 1996] comme une solution de détection parole/musique peu coûteuse en terme de calcul. Dans le cas de l'utilisation du taux de passage par zéro, la musique à tendance à avoir un taux plus élevé par rapport à la parole. Ceci est dû à la succession de notes de musique qui génère une structure plus périodique donc avec plus de passages par zéro.

Toujours parmi les paramètres non-cepstraux, il existe une approche originale basée sur des paramètres qui essayent de caractériser indépendamment la musique et la parole [Pinquier 2003a]. Ces paramètres se révèlent tout aussi performants que les paramètres classiques [Scheirer 1997]. Les paramètres utilisés sont :

- le nombre de segments par seconde ;
- la durée des segments ;
- la modulation de l'énergie à 4 Hz ;
- la modulation de l'entropie.

En ce qui concerne les deux premiers paramètres, l'algorithme de découpage en segments est issu de l'algorithme de "Divergence Forward-Backward" [André-Obrecht 1988]. En faisant l'hypothèse que le signal de parole est décrit par une suite de zones quasi-stationnaires, un modèle auto-régressif gaussien peut être utilisé :

$$\begin{cases} y_n = \sum_i a_i y_{n-i} + e_n \\ \operatorname{var}(e_n) = \sigma_n^2 \end{cases} \qquad \text{(I. 32)}$$

où y_n est le signal de parole et e_n est un bruit blanc gaussien.

La méthode consiste à détecter les changements de modèles auto-régressifs à travers les erreurs de prédiction calculées sur deux fenêtres d'analyse. A l'issue de cette détection, des segments, où les paramètres de modèles peuvent être considérés comme constants, sont obtenus.

A partir de ces segments deux paramètres sont calculés : le nombre de segments par seconde et la durée moyenne des segments. Le nombre de segments par unité de temps est beaucoup plus important pour la parole que pour la musique, dû à l'alternance voisé/non-voisé. En conséquence la durée moyenne des segments est beaucoup plus grande pour la musique que pour la parole.

Par ailleurs, le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique de 4 Hz. Pour extraire et modéliser ce paramètre la procédure suivante est appliquée :

- toutes les 16 ms, 40 coefficients spectraux sont extraits suivant l'échelle Mel et correspondent à l'énergie des 40 bandes de fréquence ;
- pour chaque bande, cette énergie est filtrée grâce à un filtre passe-bande de fréquence centrale à 4 Hz ;
- la somme des énergies filtrées est effectuée sur l'ensemble des canaux et est normalisée par l'énergie moyenne ;
- la modulation est obtenue en calculant la variance de l'énergie filtrée sur une seconde de signal.

La musique ne contient pas de modulation de l'énergie à 4 Hz par rapport à la parole qui contient une modulation très importante. Un débit verbal moyen donne environ 4 syllabes par seconde, d'où la présence de la composante à 4 Hz sur le spectre d'énergie.

Finalement, en ce qui concerne l'entropie, les observations menées sur le signal ainsi que sur le spectre du signal font apparaître une structure plus "ordonnée" du signal de la musique par rapport au signal de parole. Pour mesurer le degré de "désordre" un paramètre qui peut être utilisé est l'entropie du signal :

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i \qquad \text{(I. 33)}$$

avec p_i la probabilité de l'événement i et k le nombre d'événements possibles.

Une procédure similaire à celle employée pour la modulation de l'énergie à 4 Hz est utilisée. L'entropie est estimée grâce à un estimateur non biaisé [Pinquier 2003a] et ensuite la modulation est obtenue en calculant la variation de l'entropie sur une seconde de signal.

La modulation de l'entropie est beaucoup plus élevée pour la parole que pour la musique. Les meilleurs résultats en utilisant ce type de paramètres sont autour de 90 % de taux d'identifications correctes.

3.1.2 Approche utilisant des paramètres cepstraux

Les paramètres de type cepstraux sont très utilisés par la communauté spécialiste de traitement de la parole [Hain 1998, Gauvain 2002]. L'utilisation de ce type de paramètres est beaucoup plus coûteuse en temps de calcul par rapport aux paramètres non cepstraux.

Cependant le temps de calcul ne constitue pas un problème à partir du moment où on utilise les mêmes paramètres pour faire la tâche finale, la reconnaissance de la parole par exemple.

Par contre la paramétrisation utilisée lors de la pré-segmentation acoustique peut être différente de la paramétrisation utilisée pour la tâche finale [Meignier 2004].

Pour synthétiser, selon [Carey 1999] qui a fait une étude comparative de plusieurs paramètres, les meilleurs résultats en pré-segmentation acoustique semblent être obtenus par les paramètres cepstraux.

3.1.3 Décision inter-classe dans le cas de la pré-segmentation acoustique

Une fois effectué le choix des paramètres à utiliser pour la pré-segmentation acoustique, la question sur la stratégie de reconnaissance de différentes classes acoustiques se pose. On peut diviser les stratégies de décision en trois grandes catégories :

- stratégies fondées sur la détection de discontinuités (présence d'un pic, changement brusque de valeur de la courbe, etc.) sur la courbe temporelle des paramètres utilisés ;
- stratégies fondées sur le calcul d'une distance par rapport à une classe acoustique ;
- stratégies fondées sur le calcul d'une probabilité d'appartenance à une classe acoustique.

La première stratégie est la plus simple et il s'agit en fait de calculer une ou plusieurs courbes temporelles de paramètres et de regarder leur valeur le long du signal de parole analysé. Il faut établir une correspondance entre la valeur de la courbe et les caractéristiques acoustiques recherchées (exemple : une faible valeur sur la courbe d'énergie indique la présence d'une partie de silence). L'avantage de cette méthode est la simplicité, le désavantage est un certain manque de précision, la nécessité de fixer des seuils et aussi le fait qu'elle n'est pas toujours facile à appliquer lors de l'utilisation de plusieurs paramètres (nous pouvons être amenés à prendre une décision à partir de plusieurs courbes de paramètres qui peuvent se contredire).

Le calcul d'une distance par rapport à une classe acoustique est la méthode utilisée dans le cas de classificateurs issus de la quantification vectorielle. L'algorithme des plus proches voisins [Carbonell 1999] par exemple permet de calculer une distance entre un segment de test et des mots d'un dictionnaire qui représentent les classes acoustiques. La classe la plus proche est attribuée au segment analysé.

Le calcul d'une probabilité d'appartenance à une classe acoustique est la méthode utilisée dans le cas de classificateurs issus de la théorie statistique telle que les modèles GMM ou HMM. La décision dans ce cas est prise en utilisant le critère du risque minimal de Bayes. Etant donné un segment de test, la probabilité d'appartenance à chaque classe est calculée et le segment est attribué à la classe la plus probable.

Cependant, au moment de la décision, on a le choix entre mettre en compétition toutes les classes [Reynolds 2003, Gauvain 2003, Tranter 2003a] au même niveau ou faire une décision hiérarchique. La décision hiérarchique peut être utile [Lamy 2004] notamment dans le cas de données de type journaux télévisées ou radiophoniques. Ce type de données contient un grand nombre de caractéristiques différentes : parole studio propre, parole téléphonique, musique, bruits de toutes sortes, etc. Un exemple de classification hiérarchique est donné dans la figure suivante :

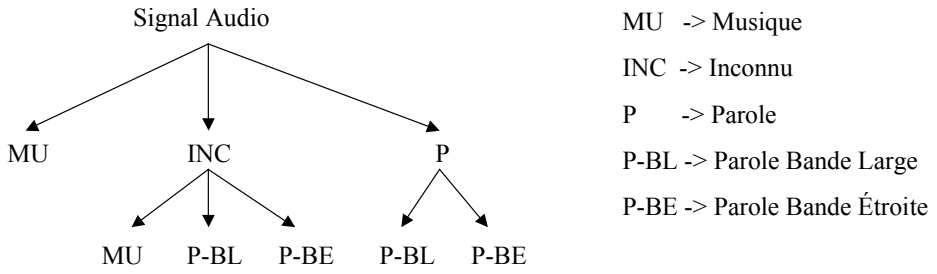


Figure I. 9 Décision hiérarchique pour la pré-segmentation acoustique

Dans cette exemple, il existe trois classes acoustiques à déterminer : musique seule (MU), parole bande large (P-BL), parole bande étroite (P-BE). La classe inconnue (INC) qui apparaît au premier niveau de la classification hiérarchique est composée de tout signal qui n'est pas suffisamment proche (au sens de la distance ou de la probabilité utilisée) des deux autres classes (P et MU) selon un intervalle de confiance. La décision hiérarchique permet de séparer dans ce cas les parties de signal audio difficilement étiquetables (INC), et d'utiliser dans leur cas des traitements différents (exemple : des modèles GMM avec plus de gaussiennes).

Dans la deuxième partie de ce manuscrit, nous présenterons une méthode de classification acoustique utilisée par notre système de segmentation.

3.2. Détection de changements de locuteur

La deuxième étape d'un système de segmentation en locuteurs est la détection de changements de locuteur. Le but de la détection de changements de locuteur est d'obtenir comme résultat des segments contenant de la parole appartenant à un seul locuteur. Idéalement, une détection de changements de locuteur parfaite fournit tous les changements de locuteur et ne rajoute aucun changement inexistant. Les segments obtenus sont les plus longs possibles.

A cet instant, les segments ne sont pas encore étiquetés en terme de locuteur. C'est l'étape suivante de regroupement qui rassemblera les segments appartenant à un même locuteur et fournira une étiquette par segment. Lors de la détection de changements de locuteur, il n'est pas très grave de *sur-segmenter* c'est-à-dire de couper en plusieurs parties un segment appartenant au même locuteur. L'étape de regroupement pourra très bien ensuite grouper dans une même classe deux segments consécutifs et corriger ce type de faute.

A la limite, on peut se demander s'il ne serait pas tout simplement plus facile de remplacer cette étape par une procédure consistant à couper les segments issus de la pré-segmentation acoustique en un grand nombre de segments de très petite taille (quelques trames par segment). Le regroupement serait ensuite appliqué pour obtenir le résultat final. La première raison pour ne pas procéder ainsi réside dans le fait qu'en pratique, les distances et les probabilités calculées sur des séquences plus longues sont généralement plus fiables et donc génèrent moins d'erreurs. La deuxième raison provient des algorithmes de regroupement classiques [Fukunaga 1990] qui ont une complexité de calcul qui croît de façon quadratique par rapport au nombre de segments en entrée.

Les techniques de détection de changements de locuteur peuvent être regroupées en trois classes [Kemp 2000] :

- détection de changements de locuteur par détection de silences ;
- détection de changements de locuteur par utilisation d'une distance ;
- détection de changements de locuteur par identification de la nature des segments.

3.2.1 Détection de changements de locuteur par détection des silences

Dans cette partie nous allons décrire la détection de changements de locuteur la plus simple, c'est-à-dire la détection des silences. L'idée est que les changements de locuteur sont souvent séparés par des segments de silence. Cette méthode suppose que les locuteurs ne se coupent pas la parole.

Les segments de silence sont généralement caractérisés par un faible niveau d'énergie, donc les algorithmes de détection de silences sont basés sur le calcul d'énergie du signal. En pratique, à part le calcul d'une courbe d'énergie pour le signal analysé, une durée minimale d'un silence est définie. Cette durée est nécessaire à cause de la forte variabilité de la valeur de l'énergie le long du signal. Par ailleurs, il ne faut pas confondre les courtes pauses inter-phonémiques qui durent quelques centaines de millisecondes avec les silences inter-locuteur qui durent au moins une demie-seconde.

Une fois calculée, l'énergie est comparée à un seuil. Si elle est inférieure à ce seuil on considère qu'on a du silence sinon on a de l'activité audio (parole, bruit, etc.). Le seuil est généralement très dépendant du document analysé notamment très dépendant du niveau du bruit de fond. Il existe des méthodes basées sur l'histogramme de l'énergie [Montacie 1997] qui sont capables de déterminer le seuil de façon automatique.

Le deuxième paramètre, la taille minimale d'un segment de silence détecté, est aussi très dépendant du type de document. Nous allons voir dans la troisième partie de la thèse que ce paramètre peut varier beaucoup comme par exemple entre un enregistrement téléphonique et un enregistrement de journal télévisé.

3.2.2 Détection de changements de locuteur par utilisation d'une distance

La méthode est fondée sur le fait que deux segments appartenant à deux locuteurs différents peuvent être différenciés par le calcul d'une mesure de similarité entre eux.

La méthode générale consiste à utiliser un couple de fenêtres adjacentes qui se déplace le long du signal analysé. Une distance est ensuite calculée à la frontière commune entre les deux fenêtres. A partir des valeurs de cette distance un ensemble de points de changement potentiels est sélectionné.

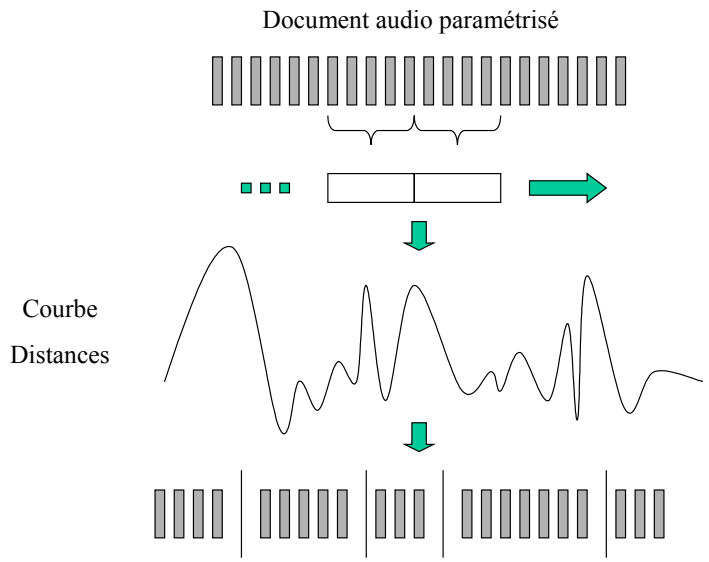


Figure I. 10 Utilisation de deux fenêtres adjacentes pour le calcul d'une distance

Une telle méthode est caractérisée par les paramètres suivants :

- la taille des fenêtres : nombre de vecteurs acoustiques par fenêtre ;
- la vitesse de déplacement des fenêtres le long du signal (exemple : la distance est calculée tous les 10 vecteurs) ;
- la mesure de distance entre les fenêtres ;
- la procédure de décision pour sélectionner les points de changement potentiels.

Le choix de la taille des fenêtres est un des paramètres qui donne la résolution de la sortie de détection. Les fenêtres doivent être suffisamment grandes pour fournir des valeurs de distance fiables. Les fenêtres doivent être suffisamment petites pour ne pas contenir la parole de plusieurs locuteurs. La parole spontanée (conversations téléphoniques, réunions) est composée d'interventions beaucoup plus courtes par rapport à la parole préparée (journaux télévisés). Généralement, pour un bon choix, il sera nécessaire d'avoir des informations *a priori* sur le document analysé, connaître au moins le type de document (téléphonique, journaux, etc.). Il existe des travaux [Delacourt 2000b, Wilcox 1994a] où pour compenser l'absence de ce type d'information, une deuxième étape de pré-segmentation avec une taille de fenêtre variable est utilisée.

La vitesse de déplacement des deux fenêtres le long du signal est importante seulement si l'algorithme est très coûteux en terme de temps de calcul sinon il est préférable de calculer la distance à chaque instant par trame, donc d'avoir une vitesse de déplacement

minimale. L'augmentation de la vitesse de déplacement correspond seulement à un sur-échantillonnage de la courbe de distances.

Le choix de la distance entre les fenêtres X et Y est généralement le résultat d'un processus d'optimisation expérimentale et de réglage. Il existe de nombreuses distances comme la distance euclidienne, la divergence gaussienne [Hain 1998], la distance de KullBach-Liebler (KL, eq. I. 33) [Siegler 1997] ; mais la distance la plus utilisée est la distance BIC (Bayesian Information Criterion, eq. I. 34) [Chen 1998]. Cette distance va être décrite plus en détail dans la deuxième partie de la thèse.

$$d_{KL}(X,Y) = E_X\left(\log P(X/M_X) - \log P(X/M_Y)\right) + \\ + E_Y\left(\log P(Y/M_Y) - \log P(Y/M_X)\right) \quad \text{(I. 34)}$$

$$d_{BIC}(X,Y) = \frac{P(X/M_X)\cdot P(Y/M_Y)}{P(XY/M_{XY})} \quad \text{(I. 35)}$$

La procédure de décision se résume à la sélection de points de changement de locuteur les plus probables. Généralement, nous avons besoin de définir et de régler un seuil pour la décision. Dans la deuxième partie de la thèse nous allons présenter une méthode simple qui fusionne la détection de maximums sur la courbe de distance avec l'utilisation d'un seuil estimé directement sur les signaux de test.

Généralement la méthode de détection de changements de locuteur par l'utilisation d'une distance fournit uniquement l'information de la présence d'un changement mais aucune information sur la nature des segments. L'exception se trouve par exemple dans [Seck 2001] qui propose l'utilisation d'un indice de rupture dit "polarisé". La propriété de polarisation permet dans le cas de la présence uniquement de deux locuteurs L1 et L2, de distinguer les deux cas possibles de ruptures : le passage du locuteur L1 vers L2 et inversement de L2 vers L1. Un type de passage correspond au maximum de l'indice et l'autre au minimum de l'indice.

Les performances de la méthode de détection de changements de locuteur sont fortement corrélées avec la collection de données analysées. Selon [Reynolds 2002] l'utilisation d'une distance par rapport à un découpage uniforme en segments de taille fixe, n'améliore pas le résultat final de la segmentation en locuteurs. Cependant, dans [Kemp 2000] les meilleurs résultats sont obtenus avec une distance BIC. Les segments dans les documents utilisés dans cette article ont une durée moyenne d'interventions d'environ 25 secondes, qui est beaucoup plus grande par rapport aux données utilisées dans les évaluations NIST [NIST

2002] qui ont une durée moyenne d'interventions d'environ 6 secondes. La conclusion générale semble être que, pour une durée moyenne d'interventions grande, l'utilisation d'une distance est fortement conseillée conjointement avec des fenêtres de grande taille, qui fournissent des résultats plus fiables.

3.2.3 Détection de changements de locuteur par identification de la nature des segments

Dans ce cas, contrairement aux deux méthodes précédentes, non seulement la présence d'un changement de locuteur est détectée, mais la caractéristique acoustique des segments peut être aussi identifiée.

Dans 3.2.1 nous avons fait l'hypothèse que tout passage d'une zone de parole vers une zone de silence, ou inversement, est équivalente à un début ou à une fin d'intervention d'un locuteur. La détection de changements de locuteur par identification de la nature des segments est une extension de cette hypothèse. Dans ce cas, tout passage d'une caractéristique acoustique (parole, silence, musique, jingle, etc.) à une autre est équivalente à un début ou à une fin d'intervention d'un locuteur. L'hypothèse est tout à fait plausible dans le cas de journaux télévisés. En regardant les transcriptions de référence des données NIST [NIST 2002, NIST 2003] il existe toujours un intervalle de non-parole (silence ou musique) entre deux interventions appartenant à deux locuteurs différents.

Pour identifier la nature des segments il suffit d'appliquer l'une des méthodes de pré-segmentation acoustique présentées dans 3.1. Les données d'entraînement dans le cas idéal proviennent de la même source que les données de test. Par exemple dans le cas de journaux télévisés les données proviennent des enregistrements de journaux de même type ; pour les enregistrements de réunions les données proviennent des réunions précédentes.

Le système de segmentation de l'université de Cambridge [Tranter 2003a] utilise uniquement une stratégie de ce type, basée sur des modèles GMM (silence, parole, musique, etc.) pour détecter aussi les changements de locuteur. [Hain 1998] affirme sans préciser un taux d'erreur exact que presque tous les changements de locuteur sont détectés ainsi.

Malgré la vérité terrain issue des transcriptions de référence, il ne faut pas oublier que la pré-segmentation acoustique est un processus automatique, donc générateur d'erreurs. Un des désavantages de cette méthode est l'impossibilité de sur-segmenter le signal. Comme nous l'avons affirmé au début, le découpage en plusieurs parties d'un segment appartenant au même locuteur peut très bien ensuite être corrigé pendant l'étape de regroupement.

3.3. Regroupement des segments

Une fois les étapes de pré-segmentation acoustique et de détection de changements de locuteur appliquées, nous obtenons idéalement des segments de signal audio contenant la parole d'un seul locuteur. Le but du regroupement est de classer alors selon le locuteur tous ces segments issus de la phase précédente.

Cette phase de regroupement n'est pas spécifique à la segmentation. Elle peut être utilisée par d'autres applications. Par exemple elle pourrait être employée pour extraire les messages téléphoniques déposés par un certain locuteur sur un répondeur téléphonique. Pour cet exemple, la phase préalable de détection de changements de locuteur n'est pas vraiment nécessaire parce que les messages sont séparés par de longs silences ou par un bip sonore.

Le concept de regroupement est plutôt universel, connu sous des noms différents et utilisé dans beaucoup de domaines. Par exemple, en anglais il est désigné, en dehors du terme correct de "clustering", par les termes "grouping", "clumping", ou même "numerical taxonomy". Il est utilisé en biologie, psychologie, sociologie, informatique, recherche en marketing, etc.

Le regroupement dans notre cas est un problème de classification non-supervisé. Il est rappelé que par hypothèse le nombre de locuteurs n'est pas connu dans la tâche de segmentation en locuteurs et aucune information *a priori* sur les locuteurs n'est disponible. A la fin du regroupement chaque classe doit idéalement contenir la parole d'un seul locuteur.

La méthode la plus utilisée dans la littérature [Fukunaga 1990] est le regroupement hiérarchique. Il existe deux grandes approches pour le regroupement hiérarchique de données. Ces deux approches sont : l'approche ascendante ou par agglomération (buttom-up clustering) et l'approche descendante ou par division (top-down clustering) :

- le regroupement par agglomération considère qu'au début chaque segment est dans sa propre classe et à chaque itération, les classes (segments) les plus proches sont réunies jusqu'à un critère d'arrêt ;
- le regroupement par division est l'opération inverse du regroupement par agglomération. Au début la réunion de tous les segments est considérée comme étant une seule classe. A chaque itération une nouvelle classe est créée par *division* d'une classe existante. De même que dans le cas précédent, le regroupement continue jusqu'à ce qu'un critère d'arrêt soit atteint.

Dans tout les cas, le résultat est présenté comme un arbre qu'on appelle dendrogramme qui décrit les opérations (agglomérations/divisions) faites à chaque itération. Le sens de

déplacement sur le dendrogramme pendant l'algorithme donne le type de regroupement par division (de haut en bas) ou par agglomération (de bas en haut).

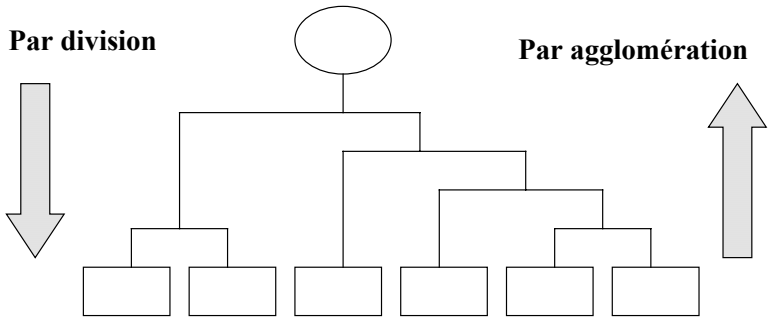


Figure I. 11 Le regroupement hiérarchique

Il existe deux facteurs qui sont très importants pour un algorithme de regroupement: le critère d'agglomération/division et le critère d'arrêt.

3.3.1 Regroupement par division

Le regroupement par division est peu utilisé dans le cas de la segmentation en locuteurs. A la dernière campagne d'évaluation NIST (RT 2003) sur la segmentation en locuteurs seulement un système utilisait ce type de regroupement : le système de l'Université de Cambridge [Tranter 2003a].

Le système de l'Université Cambridge n'utilise pas de détection de changements de locuteur. Comme nous avons vu dans 3.2.3, seulement une pré-segmentation acoustique est utilisée. A l'issue de cette phase on obtient quatre classes de parole différentes en fonction de la bande passante du signal (bande large/bande étroite), et en fonction du genre du locuteur (homme/femme). Deux classes non-parole sont obtenues : silence et musique. Les segments de silence et musique sont éliminés et le regroupement est appliqué ensuite de façon indépendante sur chaque classe acoustique de type parole.

Pour la phase de regroupement, chaque segment est modélisé par une matrice de covariance de paramètres statiques et des premiers paramètres dérivés. La stratégie utilisée est une stratégie de regroupement par division de haut en bas :

- à chaque itération, chaque nœud du dendrogramme est divisé en un maximum de quatre nœuds en optimisant une fonction de coût ;
- ensuite les segments de parole sont attribués aux nœuds les plus proches en utilisant une distance de moyenne harmonique sphérique [Bimbot 1993] ;

- la procédure de division continue jusqu'à ce qu'il ne reste aucun nœud à diviser tout en respectant un critère de densité minimale ;
- à la fin, tout nœud trop petit pour générer une matrice de covariance est attribué au nœud le plus proche.

La méthode est comparée avec une méthode ascendante et a donné de bons résultats pendant les campagnes d'évaluation NIST [Tranter 2003a] et DARPA [Hain 1998].

3.3.2 Regroupement par agglomération

Le regroupement par agglomération est la méthode la plus utilisée en segmentation en locuteurs [Gish 1991, Reynolds 2003, Nguyen 2003a, Ajmera 2003]. Dans les évaluations NIST elle est utilisée par la quasi totalité des systèmes de segmentation.

A l'issue de la détection de changements de locuteur, une série de segments est obtenue. Chaque segment est considéré comme une classe de départ. A chaque itération, les deux classes les plus proches sont groupées en une nouvelle classe selon une distance calculée entre tous les couples de classes. L'algorithme s'arrête selon un critère d'arrêt ou quand toutes les classes ont été regroupées en une seule classe, le but dans ce dernier cas étant d'obtenir le dendrogramme complet.

La répartition finale des données dans les classes est faite en sélectionnant un ensemble de classes parmi les nœuds (classes) du dendrogramme. Le regroupement hiérarchique est entièrement défini par la distance entre les classes et par le critère d'arrêt.

Différentes distances sont proposées dans la littérature. Ces mesures sont généralement issues, soit de la théorie de la quantification vectorielle comme la distance par paire moyenne, la distance par paire minimale, etc. ; soit de la théorie des modèles probabilistes.

Dans cette dernière catégorie, les mesures le plus utilisées sont :

- la distance BIC (Bayesian Information Criterion, BIC [Chen 1998, Delacourt 2000a]) ;
- la distance de KullBach-Liebler ([Siegler 1997]) ;
- le rapport de vraisemblance croisé (Cross Likelihood Ratio, CLR [Reynolds 1998]).

Les deux premières distances ont été définies dans 3.2.2. Elles sont similaires à celle utilisées en détection de changement de locuteur, mais elles s'appliquent sur des segments de

parole de taille variable plutôt que sur des fenêtres adjacentes de signal. Le rapport de vraisemblance croisé est donné quant à lui par :

$$d_{CLR}(X,Y) = \frac{P(X/M_X)}{P(X/M_Y)} \cdot \frac{P(Y/M_Y)}{P(Y/M_X)} \qquad \text{(I. 36)}$$

La distance BIC est la plus coûteuse en temps de calcul, notamment dans le cas de l'utilisation de modèles de mélanges de gaussiennes. Cependant, elle donne les meilleurs résultats. Les deux dernières méthodes sont moins coûteuses en temps de calcul car elles ne nécessitent pas l'estimation du modèle représentant la réunion de chaque couple de classes. Cependant, elles semblent donner des résultats moins performants.

Etant données N classes (ou segments) au départ, une matrice symétrique de distances de taille N^2 est calculée en utilisant la distance choisie. Les éléments de la matrice sont les distances calculées entre toutes les paires de classes (c_i, c_j).

A chaque itération, les classes qui génèrent la plus petite distance sont fusionnées en une nouvelle classe. La matrice devient une matrice de taille $(N-1)^2$ et toutes les distances entre la nouvelle classe et les classes restantes sont re-estimées. Il faut dans ce cas re-estimer $N-2$ distances.

Pour calculer toutes les distances entre la nouvelle classe et les classes restantes, deux stratégies sont envisageables :

- apprendre le modèle de la nouvelle classe et ensuite calculer normalement toutes les distances entre la nouvelle classe et les anciennes classes ;
- calculer les distances entre la nouvelle classe et les anciennes classes à partir des distances déjà existantes dans la matrice avant l'apparition de la nouvelle classe. Ceci est possible par exemple pour les distances statistiques utilisant des modèles mono-gaussiens.

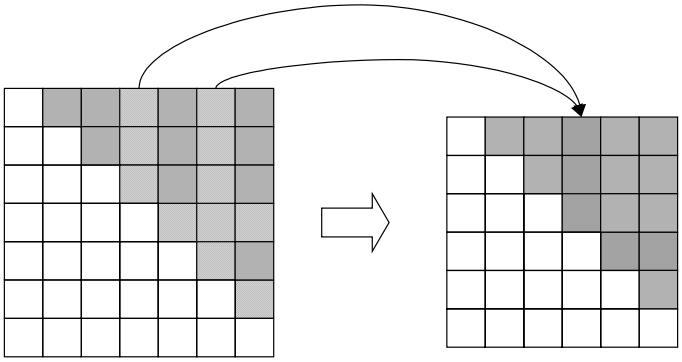


Figure I. 12 Mise à jour d'une matrice de distance pour le regroupement hiérarchique

Il est évident que la dernière option est mois coûteuse en temps de calcul mais il n'est pas toujours possible de faire le calcul envisagé.

La complexité du calcul du regroupement est de l'ordre N^2 avec N le nombre de segments fournis par la détection de changements du locuteur. L'utilisation d'un découpage en nombreux segments de taille fixe, à la place d'une détection de changements de locuteur augmente donc de façon considérable le temps d'exécution de l'étape du regroupement.

3.3.3 Critère d'arrêt du regroupement hiérarchique

Le deuxième élément dans un algorithme de regroupement hiérarchique est le critère d'arrêt qui donne la possibilité de sélectionner le nombre final de classes.

Le principe général [Everitt 1993] est d'essayer de maximiser (ou minimiser) une probabilité (ou une distance). Comme dans cette thèse, l'approche de segmentation est basée uniquement sur des modèles statistiques, nous sommes intéressés uniquement à la maximisation d'une probabilité.

L'idée est la suivante :

- soit une répartition des données selon un nombre k de classes à un moment donné ;
- soit un ensemble de modèles statistiques M_i (GMM ou autre) de chaque classe ;
- on calcule la probabilité des données X étant donné les k classes c'est à dire :

$$P\left(X / M_1 M_2 \ldots M_k\right) \qquad \text{(I. 37)}$$

- on veut donc trouver le nombre de classes n qui satisfait :

$$n = \arg\max_k P\left(X / M_1 M_2 \ldots M_k\right) \qquad \text{(I. 38)}$$

Cependant, le problème n'est pas aussi simple que cela. La probabilité considérée dans l'équation I. 37 augmente avec le nombre k de classes. La solution choisie dans la littérature est donc de pénaliser la probabilité par le nombre et la complexité des modèles. C'est le cas des différents critères proposés dans les équations I. 38 à I. 40: le Critère du Descripteur Minimal (MDL) [Rissanen 1989], le Critère d'Information d'Akaike (AIC) [Kieseppa 1997] et, plus utilisé, le Critère d'Information Bayesien (BIC) [Schwarz 1978].

$$MDL(M) = \log P\left(X / M_1 M_2 \ldots M_k\right) - \\ - \frac{k \cdot m}{2} \log N_X - \left(\frac{k \cdot m}{2} + 1\right) \log\left(\frac{k \cdot m}{2} + 2\right) \qquad \text{(I. 39)}$$

$$AIC(M) = \log P\left(X / M_1 M_2 \ldots M_k\right) - k \cdot m \qquad \text{(I. 40)}$$

$$BIC(M) = \log P\left(X / M_1 M_2 \ldots M_k\right) - \frac{k \cdot m}{2} \log N_X \qquad \text{(I. 41)}$$

Dans les formules précédentes, m représente la complexité du modèle M_i et N_X la quantité de donnés X utilisée, exprimée en nombre de trames. Les principales différences entre les trois critères proviennent de la façon de calculer le facteur de pénalisation de la complexité des modèles.

Dans la deuxième partie de la thèse, nous présenterons notre propre façon d'utiliser le critère BIC pour estimer le nombre de locuteurs présents dans un document audio.

3.4. Re-segmentation

Une fois l'étape de regroupement finie, le résultat est une répartition des données entre les classes, c'est-à-dire une répartition de la parole d'un document audio entre les locuteurs qui ont été détectés. Cependant, dans tout le processus décrit pour l'instant, les frontières détectées lors de la pré-segmentation acoustique et de la détection de changements de locuteur n'ont jamais encore vraiment été remises en cause.

Le but de la re-segmentation est de remettre en question ces frontières [Reynolds 2002, Adami 2002, Gauvain 2003], et donc de faire un raffinement de la sortie d'un système de segmentation en locuteurs.

Pour cela, à partir du résultat de la segmentation, un modèle statistique peut être entraîné pour chaque locuteur. Chaque vecteur acoustique est ensuite attribué au locuteur le plus probable. En pratique, les vraisemblances de deux trames consécutives peuvent avoir une grande variabilité, alors que ces trames ont une forte probabilité *a priori* d'appartenir au même locuteur. La solution la plus simple pour corriger ce problème est d'utiliser une fenêtre glissante centrée sur la trame en question et de calculer la vraisemblance moyenne de cette fenêtre selon chaque locuteur.

La re-segmentation doit être appliquée avec précaution. Il n'est pas toujours évident que la re-segmentation améliore les résultats du regroupement, notamment dans le cas de très bons résultats à l'issue du regroupement. La re-segmentation peut évidemment être appliquée de façon itérative. A partir du résultat de re-segmentation, il est possible d'entraîner à nouveau des modèles pour chaque locuteur et de refaire une re-segmentation. Cependant, les avis sont partagés en ce qui concerne le résultat final. Reynolds constate une dégradation du résultat si la re-segmentation est appliquée de façon itérative [Reynolds 2002]. Meignier utilise la re-segmentation de façon itérative dans son système de segmentation et constate une amélioration [Meignier 2003]. Cette dernière re-segmentation, issue du LIA, sera décrite en détail dans la deuxième partie de la thèse.

4. Segmentation en locuteurs : format du résultat et évaluation

La segmentation en locuteurs d'un document audio a pour but de décrire un flux audio donné en terme de locuteurs qui parlent. Une première question qu'on peut alors se poser est comment représenter la sortie d'un système de segmentation, c'est-à-dire comment décrire le flux audio proprement dit.

Très simplement, toute information concernant les locuteurs peut être décrite par un format de type début, fin, étiquette locuteur. Par exemple, si on admet que les marques de temps (début et fin) sont exprimées en secondes, une ligne du type 17,23 32,67 Loc1 veut dire que le locuteur Loc1 a parlé à partir de l'instant 17,23 secondes jusqu'à l'instant 32,67 secondes. C'est exactement le format NIST d'une hypothèse de segmentation, format qui a été utilisé pendant les évaluations NIST SpRec 2001 et 2002 [NIST 2001, NIST 2002]. Comme

les évaluations SpRec organisées par NIST sont dédiées à la reconnaissance du locuteur uniquement, le format était suffisamment riche pour contenir toute information désirée.

Cependant, à partir de l'année 2003, les évaluations concernant la tâche de segmentation en locuteurs sont passées dans le cadre de NIST RT (Rich Transcription) [NIST 2003]. Les évaluations RT s'intéressent à la transcription enrichie des documents audio. Donc, en plus de l'aspect locuteur, on s'intéresse à la reconnaissance de la parole et aux informations de haut niveau contenues dans un document audio. Ces informations sont appelées meta-données et sont par exemple : hésitations, genre du locuteur, nature de la parole (parole propre, parole bruitée, parole plus musique, etc.). D'où le besoin d'un format plus riche qui peut contenir ces informations. Ce format est le format MDTM (Meta Data Time Marks). Ce format offre, en plus du format précédent, la possibilité de travailler sur des document multi-voix (stéréo ou plus), de préciser le genre du locuteur mais aussi de préciser que sur un certain segment il y a un autre type de signal que de la parole (musique, bruit). Le dernier format proposé est le format RTTM (Rich Transcription Time Markers). Du point de vue "locuteur" les formats RTTM et MDTM [NIST 2003] contiennent la même information mais le format RTTM nous offre la possibilité d'utiliser le même format pour la tâche de reconnaissance de la parole et pour la segmentation.

En ce qui concerne l'évaluation, plusieurs façons d'évaluer ont été proposées. La première approche consiste a évaluer séparément la détection de changements de locuteur et le regroupement de segments par locuteurs.

La détection de changements de locuteur seule, peut être évaluée en terme de fausses alarmes (TFA) et de détections manquées (TDM) :

$$TDM = 100 \times \frac{nombre\ de\ DM}{nombre\ de\ changements\ réels} \qquad \text{(I. 42)}$$

$$TFA = 100 \times \frac{nombre\ de\ FA}{nombre\ de\ changements\ réels\ +\ nombre\ de\ FA} \qquad \text{(I. 43)}$$

Le regroupement seul peut être évalué en terme de *pureté de classes* [Solomonoff 1998]. La pureté d'une classe p_i est définie comme la probabilité que deux trames prises au hasard dans une classe appartiennent effectivement au même locuteur. Soit N_S le nombre total de locuteurs réellement présents, N_C le nombre total de classes obtenues par le système automatique et N_U le nombre total de trames. Le nombre total de classes peut être différent

ou non du nombre de locuteurs (cela dépend de la connaissance ou non *a priori* du nombre de locuteurs). Soit n_{ij} le nombre de trames de la classe i qui ont été prononcées par le locuteur j. Nous avons donc $n_{.j} = \sum_{i=1}^{N_C} n_{ij}$ le nombre total des trames qui ont été prononcées par le locuteur j et $n_{i.} = \sum_{j=1}^{N_S} n_{ij}$ la taille de la classe i.

La pureté de la classe *i* est alors définie par :

$$p_i = \sum_j \frac{n_{ij}^2}{n_{i.}^2} \tag{I. 44}$$

A partir de la pureté calculée pour toutes les classes, il existe deux indices pour caractériser la qualité du résultat d'un regroupement, c'est-à-dire d'une partition : l'indice BBN et l'indice Rand :

$$I_{BBN} = \sum_i n_{i.} p_i - Q \cdot N_C \tag{I. 45}$$

$$I_{Rand} = \sum_i n_{i.}^2 \cdot \left(\frac{1}{2} - p_i \right) + \frac{1}{2} \cdot \sum_j n_{.j}^2 \tag{I. 46}$$

L'indice Rand représente le nombre de paires de trames qui ont été prononcées par le même locuteur mais qui sont dans des classes différentes ou qui sont dans la même classe mais ont été prononcées par des locuteurs différents. Une bonne partition donne un indice Rand petit.

A l'inverse, pour une bonne partition, l'indice BBN doit être grand. Dans l'indice BBN, *Q* est un paramètre qui favorise ou non des partitions grandes avec le risque de mélanger la parole de plusieurs locuteurs. L'indice BBN représente en fait la pureté moyenne d'une partition.

L'avantage de cette façon d'évaluer par rapport aux autres qu'on va présenter tout de suite est qu'elle offre la possibilité d'évaluer séparément la détection de changements de locuteur et le regroupement. Cela peut être utile en fonction de l'utilisation du résultat de segmentation que l'on fera ultérieurement [Tranter 2003b, Tranter 2004].

Si le but final est la reconnaissance de la parole, après l'étape de segmentation en locuteurs, on est plus intéressé par une bonne détection des changements de locuteur qui peut être équivalente sous certaines conditions à une détection de frontières de phrases. Dans ce cas le résultat du regroupement n'est pas très important si on ne souhaite pas faire d'adaptation non-supervisée des modèles acoustiques par exemple.

Par contre, si on veut faire du suivi du locuteur, le résultat du regroupement est important parce qu'une décision prise sur plus de signal appartenant au même locuteur est plus fiable [Bonastre 2000]. Les vraisemblance selon les modèles du locuteur cible sont calculées dans ce cas par classe, donc sur beaucoup plus de données par rapport au cas du calcul par trame ou par fenêtre glissante.

Le désavantage de cette évaluation séparée est qu'il n'y a pas une mesure globale du résultat.

La méthode d'évaluation qui a tendance à s'imposer dans la communauté est la méthode utilisée par NIST pendant les dernières évaluations. En principe, cette méthode consiste à faire d'abord une correspondance entre les locuteurs donnés par la transcription de référence et les locuteurs donnés par le système, et de mesurer la quantité de signal mal segmenté par rapport à la quantité totale de signal, le résultat final étant donc alors une erreur de segmentation.

L'erreur finale est en fait la somme de 5 erreurs différentes :

- Temps de fausse alarme parole (False Alarm Speech) : le temps sur lequel le système a répondu qu'il y avait de la parole alors qu'il n'y en a pas ;
- Temps de parole manquée (Missed Speech) : le temps sur lequel le système a répondu qu'il n'y avait pas de la parole alors qu'en réalité il y en a ;
- Temps de fausse alarme locuteur (False Alarm Speaker Time) : le temps attribué par le système à un locuteur qui n'a pas de correspondance parmi les locuteurs de référence ; ceci arrive si le nombre de locuteurs évalué par le système est supérieur au nombre de locuteurs réel ;
- Temps de locuteur manqué (Missed Speaker Time) : le temps attribué à un locuteur de référence qui n'a pas de correspondance parmi les locuteurs détectés par le système ; ceci arrive si le nombre de locuteurs évalué par le système est inférieur au nombre de locuteurs réel ;
- Temps d'erreur locuteur (Speaker Error Time) : c'est le temps qui n'est pas attribué au même locuteur entre la transcription de référence et la sortie du système.

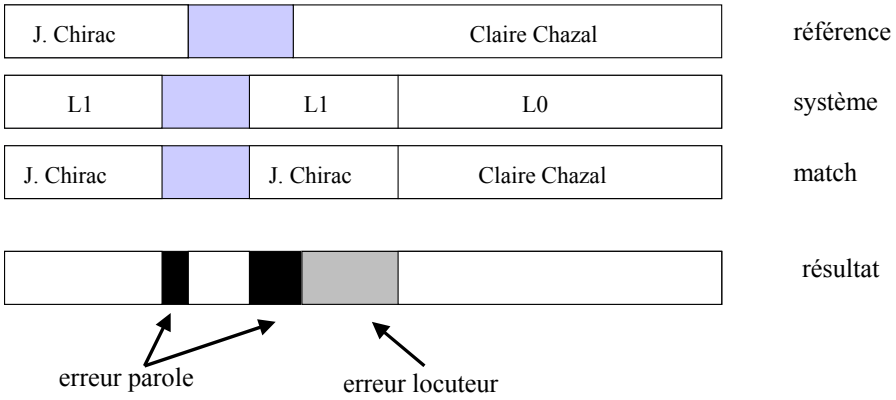


Figure I. 13 Evaluation NIST de la segmentation en locuteur

Les deux premières erreurs représentent l'erreur de détection de parole, les deux erreurs suivantes sont liées à l'erreur d'estimation du nombre de locuteurs dans l'enregistrement et enfin la dernière erreur est l'erreur de classification en locuteur. La Figure I. 13 donne une illustration de ces erreurs.

Dans nos expérimentations, nous utiliserons cette erreur NIST appelée *erreur de segmentation* (ou *diarization error*).

5. Différentes architectures de systèmes de segmentation

Ils existe de nombreux systèmes d'indexation selon le locuteur et pendant la campagne d'évaluation NIST RT 2003 on été présentés au moins six systèmes différents. Toutefois il existe peu d'architectures différentes. Dans cette partie nous présentons les principales architectures présentes dans la littérature :

- l'architecture ascendante ;
- l'architecture évolutive ;
- l'architecture séquentielle ou temps réel.

5.1. L'architecture ascendante

L'architecture ascendante a déjà été décrite dans le début de ce chapitre. C'est l'architecture proposée par Herbert Gish [Gish 1991]. Comme cette architecture à déjà été largement abordée dans ce chapitre, nous allons seulement rappeler ses principales étapes et

donner ensuite quelques détails techniques sur chaque étape, trouvés dans l'article "fondateur" de la méthode :

- la paramétrisation ;
- la pré-segmentation acoustique qui, dans [Gish 1991] est réduite à seulement une détection de silences ;
- la détection de changements de locuteur ;
- le regroupement selon le locuteur.

L'étape de paramétrisation est assez classique : 14 coefficients MFCC sont calculés après un pré-filtrage téléphonique 300-3300 kHz. Les coefficients sont calculés toutes les 10 ms sur des fenêtres de 20 ms.

D'abord, une étape de pré-segmentation acoustique silence/parole basée sur un seuil d'énergie sélectionne uniquement les segments contenant de la parole. Ensuite, une étape de détection de changements de locuteur basée sur la distance BIC, sélectionne uniquement les segments contenant de la parole d'un seul locuteur.

Enfin, les segments obtenus précédemment sont regroupés selon le locuteur. La technique utilisée est un regroupement par agglomération. Le critère d'arrêt dans ce cas est le nombre de classes supposé égal à deux : la classe "pilotes" et la classe "contrôleurs". Par hypothèse *a priori*, la classe la plus grande est considérée comme étant la classe contrôleurs.

Cette architecture a l'avantage de fournir des segments de grande taille et d'une pureté assez bonne en terme de locuteur (un seul locuteur par segment). Par contre, les frontières de segments trouvées lors de la détection de ruptures ne sont jamais remises en cause ce qui fait que généralement, on génère des erreurs à ce niveau là.

5.2. L'architecture évolutive ou "intégrée"

L'architecture évolutive est beaucoup plus récente que l'architecture ascendante. Elle a été proposée par [Meignier 2000].

Dans l'architecture ascendante présentée auparavant, il n'existe aucune interaction entre les différentes étapes de l'algorithme. En conséquence, le résultat d'une étape n'est jamais remis en cause à l'étape suivante. Pour essayer de contourner ce désavantage, Sylvain Meignier a proposé dans [Meignier 2000, Meignier 2001] une méthode reposant sur un modèle de conversation utilisant un HMM évolutif.

Dans la méthode proposée, les états de ce modèle représentent les locuteurs du document et les transitions entre ces états modélisent les changements de locuteur. A chaque

itération du processus un nouveau locuteur est ajouté, c'est-à-dire un nouvel état dans le HMM. Le document analysé est ensuite segmenté en utilisant l'algorithme de Viterbi, en tenant compte de la nouvelle configuration du modèle HHM.

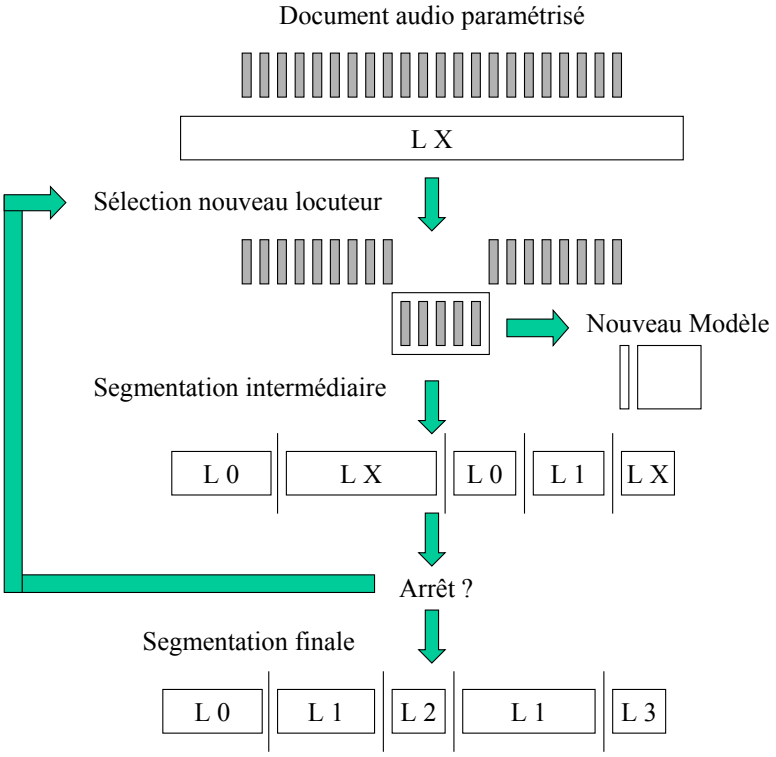


Figure I. 14 Architecture évolutive pour la segmentation en locuteurs

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

a. Initialisation :
   - le HMM contient un seul état représentant l'ensemble des locuteurs présents dans l'enregistrement ;
   - le document est composé d'un seul segment appartenant à un seul locuteur ;
b. Etape 1 (ajout d'un nouveau locuteur) :
   - ajout d'un nouveau locuteur par sélection de données ;
   - le HMM contient en conséquence un nouvel état représentant le nouveau locuteur ;
c. Etape 2 (Adaptation de modèles et segmentation intermédiaire):
   - les modèles de chaque locuteur (état) sont adaptés par MAP selon le locuteur ;
   - un processus de segmentation itérative en utilisant l'algorithme de décodage de Viterbi est lancé ;

- le processus s'arrête quand un maximum de vraisemblance au niveau de la segmentation est atteint ; une segmentation intermédiaire est alors obtenue ;

d. Etape 3 (Critère d'arrêt de la segmentation) :

- l'algorithme s'arrête si après l'ajout d'un nouveau locuteur, une augmentation en terme de vraisemblance n'est pas observée a l'étape 3 ;
- si c'est le cas, alors le dernier locuteur ajouté est éliminé.

Dans cette architecture, la segmentation intermédiaire de l'étape 2 est remise en cause à chaque itération. Cette remise en cause est faite à deux niveaux :

- à l'étape 2, lors de la phase d'adaptation de modèles un segment précédemment attribué à un locuteur peut être attribué à un autre locuteur mais, le segment peut être également divisé en plusieurs segments attribués à des locuteurs différents ;
- à l'étape 3, lors de la fin du processus, le dernier locuteur ajouté peut être supprimé.

Dans cette architecture, la détection de ruptures, le regroupement de segments et la re-segmentation sont faits en même temps. En conséquence, les informations sont disponibles au même moment, aussi bien pour trouver les changements de locuteur que pour détecter des nouveaux locuteurs. On essaye donc d'exploiter et d'améliorer les informations déjà découvertes au cours de la segmentation.

La remise en cause des informations à chaque itération et le respect du critère de maximum de vraisemblance à chaque étape constituent l'originalité de la méthode par rapport au regroupement descendant classique.

5.3. Architecture temps réel

Dans ce paragraphe, nous présentons la dernière et la plus récente des architectures : l'architecture temps réel. L'approche a été proposée dans [Liu 2004] par Daben Liu et Francis Kubala du laboratoire BBN.

Cette architecture présente deux avantages principaux : le temps d'exécution réduit et aussi le fait de ne pas avoir besoin de tout le signal à segmenter dès le début. L'indexation du signal est donc faite au fur et à mesure que le signal audio est enregistré. Cet avantage peut être très utile notamment dans le cas des enregistrements de réunions, qui nécessitent par exemple le calcul en ligne du temps de parole ou du nombre d'interventions de chaque locuteur.

Les algorithmes de regroupement hiérarchiques existants qui sont décrits par un dendrogramme nécessitent d'habitude la présence de toutes les données (tout le signal) avant de commencer son traitement. Il est généralement difficile d'insérer de nouvelles données dans le dendrogramme. Un autre avantage du regroupement proposé dans [Liu 2004] par rapport au regroupement hiérarchique, est que la complexité augmente seulement linéairement avec la taille du signal et pas de façon quadratique comme dans le cas hiérarchique.

Trois algorithmes de regroupement en temps réel sont proposés : un regroupement du type "leader-follower" (LFC), un regroupement basé sur la dispersion intra-classe (DSC) et finalement un algorithme de regroupement hybride (HSC).

Le regroupement LFC est inspiré d'une idée de Duda [Duda 2001] qui a utilisé cet algorithme pour le regroupement de type k plus proches voisins. L'idée est de modifier uniquement le modèle (le centroid dans le cas des k plus proches voisins) le plus proche de la nouvelle classe (de la nouvelle partie de signal enregistré, dans notre cas). Si il existe aucune classe proche de la nouvelle partie de signal, alors une nouvelle classe est créée.

Les deux autres algorithmes sont basés sur la même architecture générale présentée dans la Figure I. 15. Pour chaque nouveau segment de parole qui arrive, une distance est calculée par rapport aux classes existantes. Le segment est attribué à une classe existante ou une nouvelle classe est créée. La différence entre les trois algorithmes réside dans le critère de décision de la création d'une nouvelle classe. L'algorithme LFC calcule toutes les distances entre le nouveau segment et les classes existantes et les compare à un seuil fixe, l'algorithme DSC optimise un critère indépendant des données et finalement l'algorithme HSC utilise les deux approches en même temps.

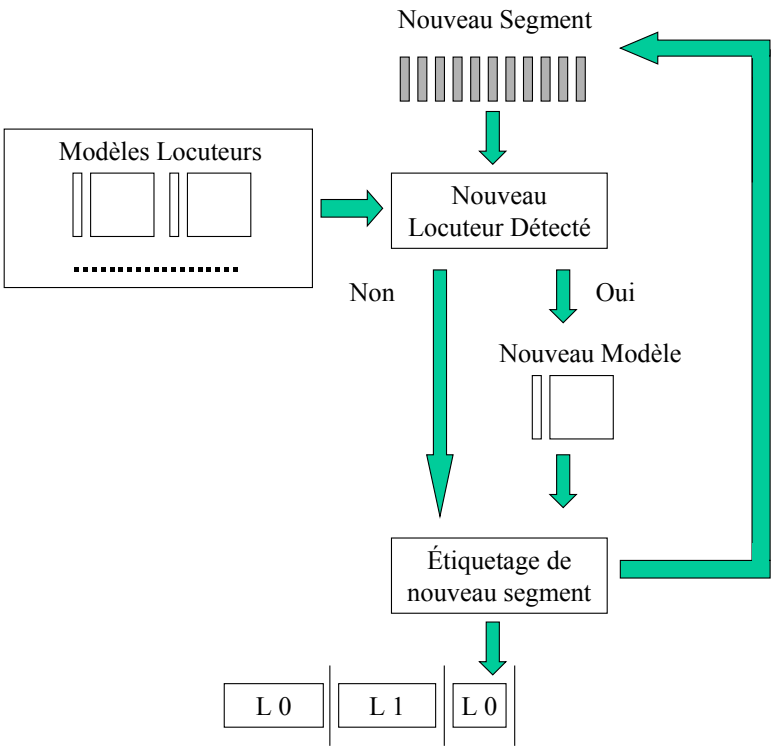


Figure I. 15 Architecture de segmentation en locuteurs temps réel

Les seuils dépendants des données sont généralement peu désirables dans un système de traitement de la parole. L'utilisation du critère de dispersion intra-classe est alors proposée comme alternative (DSC). Le but est de minimiser la quantité exprimée par :

$$G(c) = \left| \sum_{j=1}^{c} N_j \Sigma_j \right| \sqrt{c} \qquad \text{(I. 47)}$$

où c est le nombre total de classes, N_j est la taille de la classe j, Σ_j est la matrice de covariance de la classe j et $|\ |$ dénote le déterminant.

Malgré leurs performances comparables, les algorithmes LFC et HSC ne font pas les mêmes erreurs de classification. D'où l'idée d'utiliser une approche hybride entre les deux méthodes. L'approche inspirée du fait que les distances entre les classes sont largement plus grandes quand les deux classes sont différentes. Cette observation peut donc être utilisée pour améliorer les performances du critère de dispersion intra-classe. Un intervalle d'incertitude pour les distances est défini à partir de deux seuils de décision. A l'intérieur de l'intervalle on

applique le critère de dispersion intra-classe. A l'extérieur de l'intervalle, la décision est prise uniquement sur le seuil de décision.

Le regroupement hiérarchique obtient normalement des performances meilleures que le regroupement en temps réel. Le regroupement hiérarchique dispose de tout le signal audio dès le début du traitement et il a en conséquence la possibilité de trouver un nombre de classe global optimal. Pour le regroupement en temps réel, le signal arrive au fur et à mesure et donc l'optimisation du critère (exemple : critère de dispersion intra-classe) est locale et pas forcement globale.

5.4. Commentaires sur les trois architectures

Trois architectures différentes de segmentation en locuteurs ont été présentées. Chaque architecture a ses avantages et ses inconvénients.

La première architecture, l'architecture ascendante, a comme principal avantage de produire des segments de locuteurs assez longs. Les segments sont aussi de bonne qualité, c'est-à-dire les segments ne mélangent pas la parole de plusieurs locuteurs. Les inconvénients principaux sont le temps d'exécution qui augmente de façon quadratique avec la taille du signal analysé et une certaine imprécision au niveau des frontières de segments notamment sur les zones de parole spontanée. Comme nous l'avons vu dans 3.2.2, la détection de changements de locuteur donne de meilleurs résultats quand la durée moyenne des interventions des locuteurs est grande. Dans les zones de parole spontanée, la durée moyenne des interventions des locuteurs est plus petite. Il existe donc un fort potentiel d'erreur pour la détection de changements de locuteur dans ces cas là. Comme les changements détectés ne sont jamais mis en cause, on retrouve ces erreurs dans le résultat final de la segmentation.

La deuxième architecture, l'architecture évolutive, montre comme principal avantage la puissance de décodage des modèles de Markov cachés. En conséquence, les frontières de segments sont mieux détectées. Le temps d'exécution augmente de façon linéaire avec la taille du signal analysé et il est de l'ordre 3xRT. Les erreurs de segmentation obtenues par ces deux architectures sont généralement comparables. Cependant, les deux architectures ne font pas les mêmes erreurs aux mêmes endroits. En conséquence les deux architectures peuvent se compléter réciproquement. Une technique de combinaison entre les résultats de chaque architecture sera d’ailleurs présentée dans la troisième partie de la thèse.

La dernière architecture, l'architecture temps réel, a deux principaux avantages : le temps d'exécution et aussi le fait de ne pas nécessiter tout le signal de parole au début du traitement. Le résultat est normalement moins bon que dans le deux cas précédents.

6. Le suivi du locuteur

Même si le suivi de locuteur n'est pas l'objet central de cette thèse, nous le détaillons tout de même dans cet état de l'art, car il sera utilisé dans certaines de nos expérimentations. La tâche de suivi du locuteur [Bonastre 2000, Sonmez 1999, Cettolo 2001] est une tâche proche de la segmentation en locuteur. Dans le cas de la segmentation en locuteurs une des hypothèses énoncées au début est : aucune information *a priori* n'est disponible sur les locuteurs, notamment aucun échantillon de la voix des locuteurs n'est disponible. Dans ces conditions il est impossible de connaître l'identité d'aucun locuteur présent dans la conversation.

Dans le cas du suivi du locuteur, des données sont disponibles pour la construction de modèles pour certains locuteurs. Ces locuteurs peuvent être donc identifiés. Cette tâche peut être également vue comme une tâche qui apporte des informations supplémentaires pour enrichir un résultat simple de segmentation en locuteurs.

6.1. Généralités

La tâche de suivi du locuteur consiste à détecter les segments d'un document prononcés par un locuteur spécifique. Il faut répondre à deux questions concernant le locuteur dans cette tâche :

- le locuteur est il présent dans le document analysé ?
- si la réponse à la première question est oui, alors à quel endroit sur le document est il présent ?

Le système de suivi du locuteur possède une référence (un modèle *a priori*) de locuteur considéré, appelé très souvent locuteur cible. Les deux différences essentielles entre le suivi du locuteur et la segmentation en locuteur [Bonastre 2000] sont donc les suivantes :

- le locuteur cible est connu au préalable : le système possède une référence (modèle) du locuteur cible. Les autres locuteurs qui apparaissent dans le document ne sont pas connus par le système ;

- uniquement le locuteur cible est à prendre en compte. Le nombre de locuteurs et les interventions des autres locuteurs ne seront par pris en compte lors de l'évaluation d'un système de suivi.

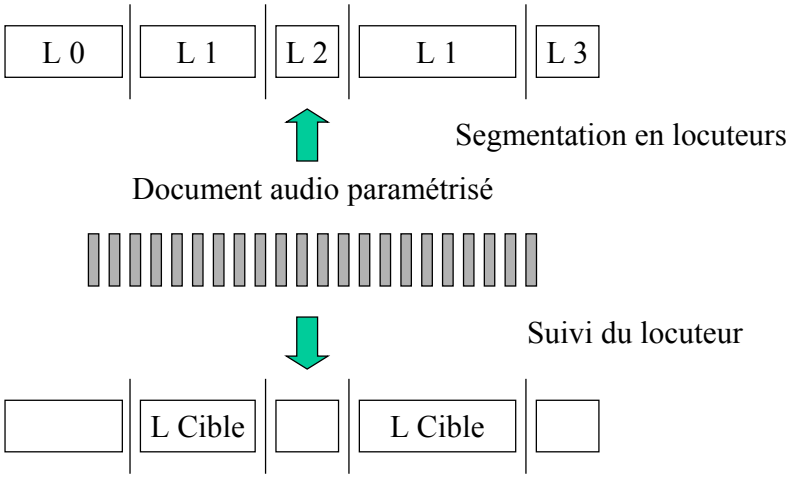


Figure I. 16 Différences entre la sortie d'un système de suivi du locuteur et la sortie d'un système de segmentation en locuteurs

Les différentes méthodes de suivi du locuteur proposées dans la littérature peuvent se découper en deux phases [Seck 2001, Cettolo 2001] :

- d'abord détecter les transitions d'un locuteur à un autre. Cela peut très bien se faire en utilisant les algorithmes de détection de ruptures présentés auparavant ;
- les segments ainsi trouvés, supposés homogènes, sont ensuite classifiés en segments locuteur cible ou autres locuteurs en utilisant par exemple le critère du maximum de vraisemblance.

Deux types d'erreurs peuvent se produire durant la première phase : une détection manquée ou une fausse alarme. Généralement, une détection manquée peut avoir des conséquences beaucoup plus négatives sur le résultat final. La présence d'un changement de locuteur dans un segment peut conduire à une erreur de classification de ce segment. En conséquence, les systèmes de détection de changement de locuteurs sont généralement réglés pour sur-segmenter ou utilisent carrément une segmentation uniforme en segments de taille fixe [Meignier 2000]. Dans ce dernier cas, le système de suivi se résume à l'application d'un critère de maximum de vraisemblance sur des fenêtres glissantes le long du signal.

Comme dans le domaine de l'identification du locuteur, pour chaque locuteur cible, un modèle de distribution de vecteurs acoustiques, généralement un GMM, est estimé à partir des

données d'apprentissage. Pour l'estimation de l'hypothèse inverse (le segment n'appartient pas au locuteur cible) un modèle du monde indépendant du locuteur est également appris.

Dans le cas du suivi du locuteur, la décision est prise sur chaque segment issu de la détection de changements de locuteur. Certains auteurs [Seck 2001] utilisent également un lissage du score de décision en utilisant les segments adjacents.

Les travaux sur le suivi du locuteur sont moins nombreux que les travaux sur l'identification et la vérification du locuteur. Notamment, il existe très peu de travaux sur les données de type journaux télévisés. Une étude préliminaire se trouve dans [Cettolo 2001] qui essaye de retrouver des locuteurs particuliers dans une base de données de journaux télévisés. Différents modèles de classification (GMM, HMM) sont testés, les résultats expérimentaux étant comparables.

6.2. La tâche de suivi du locuteur dans les évaluations NIST et ESTER

Pendant plusieurs années, de 1998 à 2000 [Martin 2001], les évaluations en reconnaissance du locuteur SpRec organisées par NIST ont proposé deux tâches pour le suivi du locuteur :

- *2-speaker detection* : à partir d'un enregistrement contenant deux locuteurs il faut répondre à la question: est-ce que le locuteur cible est présent ou non dans le document de test. C'est la première question à laquelle il faut répondre dans une tâche de suivi du locuteur ;
- *speaker tracking* : cette tâche est la tâche complète de suivi du locuteur telle qu'elle a été définie dans la section précédente.

Dans le premier cas le corpus de test utilisé est issu de la base SwitchBoard 2 [Godfrey 1992] ; il contient uniquement des enregistrements téléphoniques. Dans les signaux de test utilisés pour les évaluations NIST, il existe seulement deux types de combinés téléphoniques : charbon et électret. Chaque document est testé contre une vingtaine de locuteurs cibles dont les deux locuteurs présents dans le document. Dans les conditions d'évaluation, il est toutefois interdit d'utiliser le résultat de plusieurs locuteurs cibles en même temps. Comme dans la base de signaux de tests presque tous les locuteurs sont connus par le système, l'utilisation de vraisemblances de tous les locuteurs cibles pourrait donner une baisse artificielle de l'erreur globale. Cette interdiction crée des conditions de tests sur un corpus de locuteurs ouvert, donc, où tous les locuteurs de tests ne sont pas connus par le système.

La base de test utilisée dans le cas du suivi du locuteur est un sous-ensemble de la tâche *2-speaker detection*. Environ un quart des locuteurs cibles sont recherchés sur la moitié des documents de test. Les locuteurs cibles sont généralement présents dans les documents de test.

L'évaluation française ESTER (Evaluation des Systèmes de Transcription Enrichie des émissions Radiophoniques) [Gravier 2004] est organisée dans le cadre du projet EVALDA sous l'égide scientifique de l'Association Francophone de la Communication Parlée avec le concours de la Délégation Générale de l'Armement et de ELDA. La campagne d'évaluation ESTER vise à l'évaluation des performances des systèmes de transcription d'émissions radiophoniques. Les transcriptions sont enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des entités nommées, etc. Il existe certaines différences entre les évaluations NIST SpRec et ESTER, notamment au niveau des données utilisées pour la tâche de suivi du locuteur : les données utilisées sont des données issues d'émissions radiophoniques et pas d'enregistrements téléphoniques.

Dans ce cadre, les évaluations ESTER sont les premières évaluations à proposer une tâche de suivi du locuteur sur des documents de taille importante de type journaux radiodiffusés. Il y a environ une centaine de locuteurs cibles. Les quantités de données d'entraînement sont très variables : minimum 60 secondes jusqu'à plusieurs dizaines de minutes. La probabilité d'apparition du locuteurs cible est aussi très variable : proche de zéro jusqu'à environ 30 % dans le cas des présentateurs.

Dans la dernière partie de cette thèse, nous allons nous intéresser à la détection d'un locuteur cible particulier dans une grande collection d'enregistrements. La différence par rapport aux évaluations NIST sera le type de données qui seront audiovisuelles (journaux télévisés) ; la différence par rapport aux évaluations ESTER sera la probabilité d'apparition du locuteur cible qui sera extrêmement faible dans un cas (recherche de M.Albright sur une collection de vidéos) et importante dans l'autre cas (recherche du présentateur du journal).

7. Conclusion générale

Dans ce chapitre, nous avons passé en revue les principaux outils de traitement de la parole utilisés en segmentation en locuteurs et aussi les principales architectures de systèmes de segmentation existantes dans la littérature.

L'architecture la plus utilisée est l'architecture ascendante avec ses principaux modules qui sont : la paramétrisation, la pré-segmentation acoustique, la détection de changements de locuteur et le regroupement.

Pour l'étape de paramétrisation, les coefficients cepstraux de type MFCC semblent être les plus utilisés et donnent les meilleurs résultats. L'utilisation de l'énergie est aussi fortement recommandée en segmentation, vu les différences de volume qui peuvent exister entre plusieurs locuteurs présents dans un même enregistrement. Les informations liées aux canaux de transmission doivent être préservées pour la segmentation. Dans ce cadre, la présence de deux canaux de transmission différents (parole studio bande large vs. parole téléphonique) peut correspondre à la présence de deux locuteurs différents. Pour cette raison, les techniques liées à l'élimination de l'effet du canal de transmission utilisées en reconnaissance du locuteur, comme la soustraction de la moyenne cepstrale et les coefficients différentiels, ne sont pas utilisées pour la tâche de segmentation en locuteurs.

Les mêmes coefficients cepstraux semblent donner aussi les meilleurs résultats en ce qui concerne la pré-segmentation acoustique. Pourtant, d'autres paramètres issus du traitement automatique de signaux musicaux, semblent mieux fondés d'un point de vue théorique. Ces paramètres essayent d'utiliser les caractéristiques propre à chaque type de classe (silence, musique, etc.) par rapport aux paramètres cepstraux fondés uniquement sur les différences de spectres entre classes. Dans le cas de la segmentation de données de type journaux télévisés qui contient beaucoup de classes différentes (silence, musique, bruits, parole bande large ou étroite, etc.) une décision de type hiérarchique s'impose. La décision hiérarchique, par opposition à la mise en compétition de toutes les classes au même moment semble donner des résultats légèrement meilleurs.

La détection de changements de locuteur essaie de faire un compromis entre deux aspects : avoir des segments longs pour augmenter la robustesse de la phase de regroupement et en même temps, avoir des segments qui ne contiennent qu'un seul locuteur (pour le regroupement les segments sont considérés uni-locuteur). Pour la détection des changements de locuteur, l'utilisation d'une distance calculée entre deux fenêtres adjacentes le long du signal, donne les meilleurs résultats. Cependant, les résultats sont fortement liés à la nature de données et l'utilisation d'une distance semble moins adaptée pour des conversations de nature très spontanée.

Comme vu dans la partie 3.3, le regroupement le plus utilisé en segmentation est le regroupement hiérarchique ascendant. Cette méthode donne de bons résultats et elle est

fondée sur un cadre théorique robuste. Le principal désavantage du regroupement ascendant est le temps de calcul qui augmente de façon quadratique avec la quantité de données.

Pour résumer, un système de segmentation avec des performances proches de l'état de l'art peut être obtenu avec une paramétrisation qui utilise des coefficients cepstraux, une pré-segmentation acoustique hiérarchique, une détection de changements de locuteur basée sur une distance et finalement une étape de regroupement hiérarchique ascendant. Une étape de re-segmentation peut être optionnellement utilisée à la fin.

L'architecture évolutive à été proposée quant à elle pour essayer de faire interagir les différents modules de l'architecture ascendante. L'architecture repose sur un algorithme itératif et utilise un HMM évolutif avec un locuteur par état. Les résultats sont comparables avec ceux obtenus par l'architecture classique et en plus le temps d'exécution augmente de façon linéaire seulement avec la taille des données à traiter.

L'architecture temps-réel permet de surmonter le défaut du temps d'exécution de l'architecture ascendante classique. Celle-ci présente un temps d'exécution nettement plus faible, mais les résultats obtenus sont moins bons par rapport aux deux autres architectures.

Pour enrichir le résultat de la segmentation, une phase de suivi de locuteur peut être utilisée à condition d'avoir des données d'entraînement pour les locuteurs présents dans les documents analysés.

Deuxième partie :

Segmentation en locuteurs sans informations *a priori* ; système de référence

---

Cette partie présente notre plate-forme expérimentale (système et corpus). Le premier système de segmentation en locuteurs, que nous avons développé, est issu de l'état de l'art et les principales étapes d'un système de segmentation y sont présentes. Ce système nous a permis de participer aux différentes évaluations nationales et internationales avec des résultats très honorables. A partir de ces résultats, nous avons amélioré notre système chaque année, la plus importante amélioration étant le module d'estimation du nombre de locuteurs qui amène un gain significatif en termes d'erreur de segmentation. Les corpus utilisés pour les expérimentations présentées dans le chapitre suivant du manuscrit sont également présentés dans cette partie.

1. Introduction

Cette partie est dédiée à la présentation du système de segmentation en locuteurs du laboratoire CLIPS que nous avons développé. Nous allons présenter dans un premier temps notre système de départ qui va être ensuite amélioré en rajoutant un module pour estimer le nombre de locuteurs présents dans un enregistrement. La paramétrisation a aussi été améliorée pour être mieux adaptée à la tache de segmentation en locuteurs.

Nous rappelons ici que par hypothèse, la segmentation en locuteurs est généralement faite sans aucune information *a priori* disponible au départ. L'utilisation d'informations *a priori* peut cependant aider le système de segmentation à mieux s'adapter au différents types de données audio sur lesquelles il travaille. Cet aspect sera abordé dans la troisième partie de cette thèse.

Tout au long de ce chapitre, des résultats seront présentés dans le contexte des campagnes d'évaluation NIST et ESTER. Nous avons travaillé sur plusieurs types de données : enregistrements téléphoniques, enregistrements de réunions mais la plupart des travaux ont été faits sur des enregistrements de type "journaux télévisés".

2. Plate-forme expérimentale et corpus de segmentation

Avant de décrire le système de segmentation que nous avons conçu et les méthodes proposées, nous allons commencer par décrire la plate-forme expérimentale et le corpus de données utilisé qui nous ont permis de donner des résultats expérimentaux pour illustrer chaque algorithme proposé.

2.1. Plate-forme ELISA

La plate-forme utilisée pour toute implémentation dans cette thèse est la plate-forme commune de développement conçue par le consortium ELISA [Magrin-Chagnolleau 2001]. Le consortium ELISA a été fondé en 1997 par le LIA (Avignon), l'ENST (Paris) et l'IRISA (Rennes), autofinancé par ses participants, avec pour objectif de faciliter les recherches coopératives en reconnaissance du locuteur ainsi que la participation des laboratoires francophones aux campagnes d'évaluations internationales telles que les campagnes NIST. La composition du consortium a varié le long des années et le laboratoire CLIPS a rejoint le

consortium en 2001. Outre la plate-forme ELISA, le consortium organise des réunions régulières et fournit support technique et scientifique à ses membres pour les différentes campagnes d'évaluations. Ce consortium a notamment encouragé des travaux collaboratifs entre plusieurs laboratoires dont certaines expérimentations reportées dans cette thèse sont un exemple.

Les paramètres acoustiques sont calculés à l'aide du module SPRO développé au laboratoire IRISA par Guillaume Gravier. Le module est capable de calculer plusieurs types de paramètres : coefficients FFT, bancs de filtres, coefficients cepstraux y compris MFCC, énergie ; mais également de faire des traitements du signal basiques comme le filtrage passe-bas par exemple.

Le partie modélisation de la plate-forme nous permet d'estimer par EM les paramètres de modèles de mélanges de gaussiennes (GMM) avec des matrices diagonales ou pleines. Plusieurs méthodes d'adaptation MAP sont également disponibles avec cette plate-forme.

Le module de décision permet quant à lui, le calcul de vraisemblances associées à chaque modèle pour le corpus de test considéré.

Actuellement, la plate-forme est en phase de re-écriture sous la forme d'une boîte à outils nommée ALIZE (http://www.lia.univ-avignon.fr/heberges/ALIZE) et une première version est déjà disponible sous forme de logiciel libre. Ensemble avec la plate-forme seront distribués un système de vérification du locuteur et un système de segmentation en locuteurs issus de l'état de l'art.

2.2. Corpus expérimental issu des campagnes d'évaluation NIST et ESTER

Le corpus utilisé est issu des campagnes d'évaluation NIST et ESTER qui ont déjà été décrites dans la partie état de l'art. Les corpus NIST concernent les évaluations NIST SpRec 2002 (Evaluations en reconnaissance du locuteur), NIST RT 2003 et NIST RT 2004 Meeting (Evaluations en transcription enrichie).

Les corpus anglais NIST contiennent trois types de données : téléphoniques (SpRec 2002), journaux télévisés (SpRec 2002, RT 2003), réunions (SpRec 2002, RT 2004 Meeting). Le corpus français ESTER (Phase 1) contient uniquement des journaux radiodiffusés. Tous les corpus sont enregistrés en qualité 16 kHz, 16 bit à l'exception du corpus téléphonique NIST 2002 enregistré en 8 kHz.

Tableau II. 1 Les différents corpus utilisés dans nos expérimentations

Corpus	Evaluation	Fichiers	Taille fichiers (minutes)	Nombre de locuteurs par fichier	Pourcentage de temps multi-locuteur
Phone 2002	SpRec 2002	199	2	2	20 %
BN 2002	SpRec 2002	76	env. 2	2 à 9	0 %
Meet 2002	SpRec 2002	166	2	4 ou 6	41 %
BN 2003 Dev	RT 2003	6	10	6 à 18	1 %
BN 2003 Eva	RT 2003	3	30	10 à 27	0 %
Meet 2004	RT 2004 Meeting	8	10	3 à 10	29 %
ESTER Dev	ESTER	6	20 ou 60	13 à 33	0 %
ESTER Eva	ESTER	6	20 ou 60	16 à 38	0 %

Le Tableau II. 1 rappelle les principales caractéristiques de chaque corpus expérimental. La difficulté des données de type réunion réside dans le caractère totalement spontané de la conversation : grand pourcentage de temps où il y a plusieurs locuteurs qui parlent en même temps, beaucoup d'hésitations, présence de voix de fond (*background voices*). La difficulté des données de type journaux réside dans le grand nombre de locuteurs et aussi dans la présence de plusieurs qualités de signal (studio, téléphonique, bruité, etc.).

En ce qui concerne le calcul d'erreur de segmentation nous allons utiliser systématiquement la métrique officielle des évaluations NIST et ESTER qui à déjà été présentée dans la partie état de l'art. Nous rappelons ici que la métrique en question donne comme résultat un taux d'erreur qui représente le pourcentage de temps total mal classifié, aussi bien en terme de locuteur qu'en terme de détection parole/non-parole.

3. Le système de segmentation initial (2001-2002)

Le système de segmentation initial, présenté dans la Figure II. 1, est un système issu en totalité de l'état de l'art. Il correspond au début des recherches dans le domaine de la segmentation en locuteurs au laboratoire.

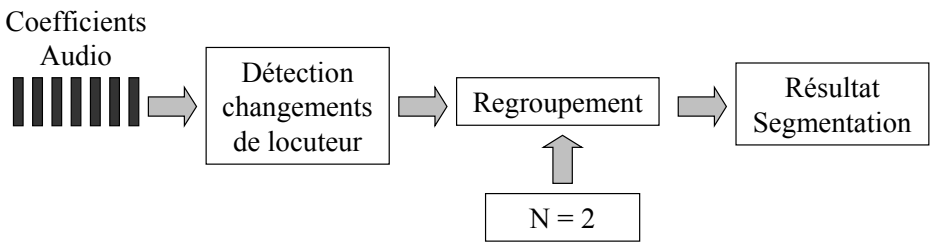


Figure II. 1 Système de segmentation initial

Nous retrouvons ici les principales étapes d'un système de segmentation en locuteurs classique : la paramétrisation, la détection de changements de locuteur et le regroupement. Le critère d'arrêt pour le regroupement était le nombre de locuteurs supposé égal à deux dans un premier temps. Une étape de pré-segmentation acoustique et une étape de re-segmentation seront rajoutées ultérieurement. Au moment des évaluations NIST 2002 sur la segmentation en locuteurs, l'information sur les zones contenant uniquement de la parole était fournie par les organisateurs. L'étape de pré-segmentation acoustique n'était donc pas nécessaire.

3.1. La paramétrisation

Le signal de parole est caractérisé pas 16 coefficients cepstraux repartis selon une échelle mel (MFCC). Les coefficients sont calculés à partir d'un banc de filtres de taille 24, toutes les 10 ms sur des fenêtres de taille 20 ms. Aux 16 coefficients on rajoute un dernier paramètre : c'est le logarithme de l'énergie calculée sur les mêmes fenêtres.

Nous désirons garder les informations liées au canal de transmission qui peuvent être très utiles pour la détection de changements de locuteur. En conséquence, la soustraction de la moyenne cepstrale (CMS) n'est pas appliquée. Par ailleurs, nous n'utilisons pas non plus les coefficients dérivée première et dérivée seconde des coefficients puisque des expérimentations préliminaires ont montré que les coefficients différentiels n'apportaient aucun gain.

3.2. Détection de changements de locuteur

Une fois les segments de parole identifiés dans un document audio paramétrisé, la première étape d'un système de segmentation en locuteurs est la détection de changements de locuteur.

Le but de la détection de changements de locuteur est de découper un document audio en segments contenant la parole d'un seul locuteur. Idéalement, nous devons détecter tous les changements de locuteur en limitant les détections manquées et les fausses alarmes. En pratique une bonne détection de changements de locuteur doit essayer de minimiser le taux de détections manquées tout en gardant un nombre raisonnable de fausses alarmes, ces dernières pouvant être récupérées lors de la phase suivante de regroupement. Cependant, un trop grand taux de fausses alarmes produit des segments de petite taille qui ne sont pas appropriés pour les calculs de probabilité.

La détection de changements de locuteurs de notre système est une détection de changements acoustiques similaire à celle présentée dans la section 3.2.2 de la partie état de l'art. L'idée est de trouver certaines discontinuités dans le signal audio qui pourront nous aider à distinguer deux locuteurs consécutifs. On peut utiliser deux fenêtres adjacentes et une mesure de similarité entre elles. Un exemple de mesure de similarité est la distance BIC (Bayesian Information Criterion) [Delacourt 2000a].

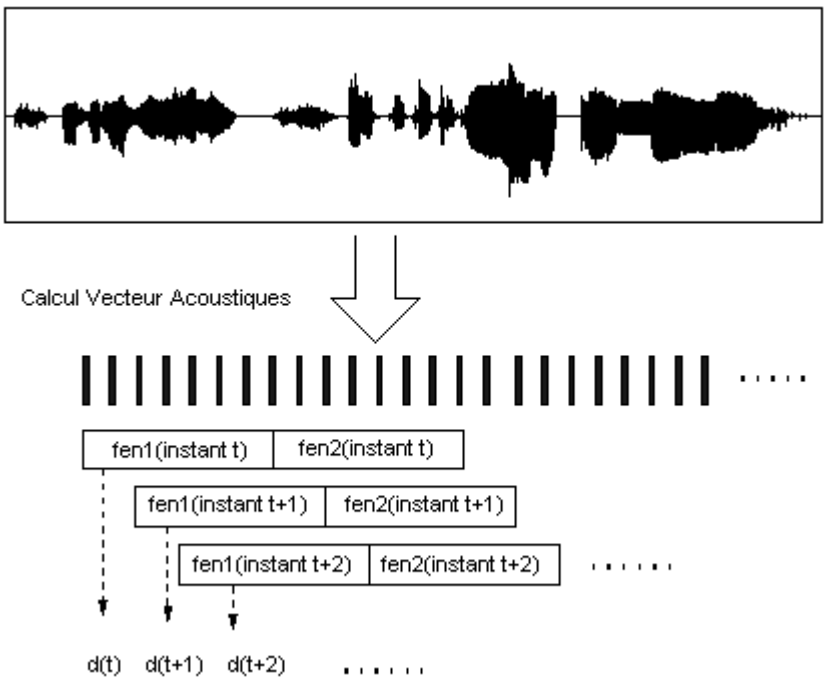


Figure II. 2 Les fenêtres glissantes pour la détection des changements de locuteur

Pour utiliser la distance BIC nous allons considérer que le signal audio paramétrisé est un processus gaussien. L'hypothèse peut être considérée comme vraie si on tient compte du fait que les paramètres MFCC sont une somme pondérée de plusieurs variables aléatoires qui

sont le logarithme des sorties de filtres, ce qui nous approche des conditions du théorème central limite. Cependant l'indépendance de ces variables aléatoires reste discutable.

La détection de changements de locuteur est basée sur la théorie de la décision. Soit deux segments consécutifs de parole, chacun caractérisé par une séquence de paramètres acoustiques spectraux (exemple : MFCC, LPCC, etc.) notés par x_n, $n = 1 \ldots N_1$ et respectivement par y_n, $n = 1 \ldots N_2$. Les vecteurs dans chacune des deux séquences peuvent être modélisés comme provenant d'une distribution gaussienne multidimensionnelle.

La question que nous posons en ce qui concerne les deux séquences est : est-ce qu'elles proviennent ou pas du même modèle fondamental, ou, d'une manière équivalente, est-ce que les segments ont été prononcés par le même locuteur ?

Nous devons tester les deux hypothèses suivantes :

- H_0 : les séquences ont été prononcées par le même locuteur ;
- H_1 : les séquences ont été prononcées par des locuteurs différents.

Si la deuxième hypothèse est vraie, alors un changement de locuteurs est détecté à la frontière commune des deux fenêtres. Nous allons utiliser deux fenêtres glissantes qui se déplacent le long du signal comme dans [Delacourt 2000a].

Nous pouvons évaluer ces deux hypothèses en utilisant le rapport de vraisemblance généralisé [Gish 1991]. Si $L(x;\mu_1,\Sigma_1)$ dénote la probabilité que la séquence x ait été générée par le modèle gaussien caractérisé par le vecteur moyen μ_1 et par la matrice de covariance Σ_1 et si $L(y;\mu_2,\Sigma_2)$ dénote la même probabilité mais pour la séquence y ; alors la probabilité L_1 qui dénote que les deux séquences ont été générées par des modèles différents est :

$$L_1 = L(x;\mu_1,\Sigma_1) \cdot L(y;\mu_2,\Sigma_2) \qquad \text{(II. 1)}$$

La probabilité que les deux séquences aient été générées par le même modèle est donnée par :

$$L_0 = L(z;\mu,\Sigma) \qquad \text{(II. 2)}$$

où z est l'union des deux séquences x et y, et μ et Σ sont les paramètres du modèle estimé à partir de la séquence z.

Si nous considérons que λ dénote le rapport de vraisemblance, alors $\lambda = \frac{L_0}{L_1}$, et nous obtenons :

$$\lambda = \frac{L(z;\mu,\Sigma)}{L(x;\mu_1,\Sigma_1)\cdot L(y;\mu_2,\Sigma_2)} \qquad \text{(II. 3)}$$

Si nous utilisons maintenant des log-vraisemblances alors le rapport de vraisemblance généralisé est donné par :

$$R = -\log\lambda = -\log L(z;\mu,\Sigma) + \log L(x;\mu_1,\Sigma_1) + \log L(y;\mu_2,\Sigma_2) \qquad \text{(II. 4)}$$

Nous pouvons prouver [Delacourt 2000a] que pour des distributions gaussiennes et pour de bonnes estimations des matrices de covariance, le logarithme du rapport de vraisemblance devient (cf. annexe A) :

$$R = \frac{N_1 + N_2}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| \qquad \text{(II. 5)}$$

Nous estimons ce rapport pour toutes les données disponibles et nous obtenons une séquence de valeurs de rapport $R(t)$. L'évaluation du point de changement de locuteur est alors la valeur de i qui maximise $R(t)$:

$$\hat{i} = \arg\max_t R(t) \qquad \text{(II. 6)}$$

L'algorithme qui détermine le changement de locuteur basé sur le rapport de vraisemblance recherche les maxima locaux de la courbe $R(t)$ qu'on va appeler ensuite *la courbe BIC.*

La recherche des maxima locaux de la courbe $R(t)$ peut être faite à partir de la dérivée ou à l'aide d'une fenêtre glissante sur la courbe.

Nous avons choisi de rechercher des maxima sur la courbe BIC à l'aide d'une fenêtre glissante. Le principe est le suivant : pour chaque point de la courbe BIC on prend une fenêtre autour de ce point et on vérifie s'il est bien le maximum de cette fenêtre ou non (Figure II. 3).

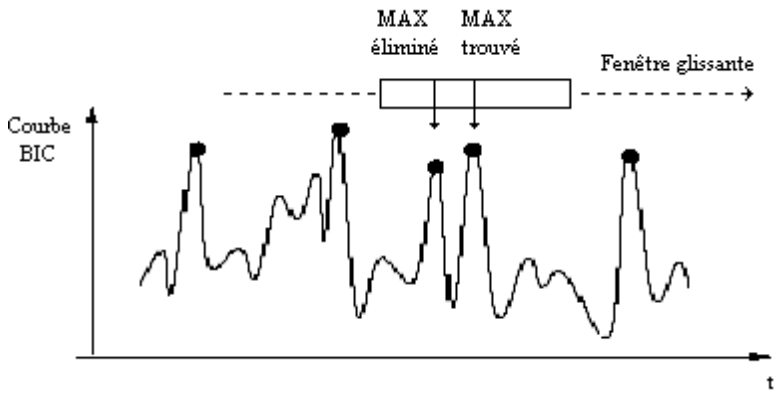


Figure II. 3 La recherche de maxima locaux sur la courbe BIC

Cette méthode a l'avantage d'éliminer un maximum faible trop proche d'un maximum fort. Si deux maxima sont situés dans la même fenêtre, seulement un seul est choisi. Par rapport à l'utilisation de la dérivée, nous avons l'avantage de ne pas détecter tous les maxima locaux. En effet, une dérivée détecterait tous les maxima locaux y compris ceux de très faible amplitude.

Malgré cela, dans notre approche, nous avons toujours un nombre important de maxima détectés. Pour cette raison nous avons besoin d'un critère de sélection de maxima détectés. L'approche la plus simple est l'utilisation d'un seuil de décision. Chaque maximum détecté sera comparé au seuil. Pour assurer une certaine indépendance du seuil face aux données, le seuil est défini comme la moyenne de la courbe BIC multipliée par une constante.

3.3. Le regroupement des segments

Les segments issus de la phase de détection de changements de locuteur doivent être regroupés selon le locuteur. Nous allons utiliser un algorithme de regroupement hiérarchique ascendant. Il y a deux choses importantes pour un tel algorithme : la distance entre les classes et le critère d'arrêt.

Le critère d'arrêt sera pour nous le nombre de locuteurs présents dans l'enregistrement. Quand ce premier système a été réalisé, au moment des évaluations NIST SpRec 2002 sur la segmentation en locuteurs, nous ne savions pas estimer le nombre de locuteurs présents dans

un document audio. Vu la taille des documents à traiter, en moyenne 2 minutes, le besoin d'estimation du nombre de locuteurs n'était pas alors évident. Pour le système CLIPS, le nombre de locuteurs fut donc fixé à 2 pour tous les enregistrements.

La distance entre les classes est toujours la distance BIC. Soit deux classes C_m et C_n, la distance BIC entre ces deux classes est donnée par :

$$d_{BIC}\left(C_m, C_n\right) = \frac{P\left(C_m / M\left(C_m\right)\right) P\left(C_n / M\left(C_n\right)\right)}{P\left(C_{m \cup n} / M\left(C_{m \cup n}\right)\right)} \qquad \text{(II. 7)}$$

ou $M\left(C_m\right)$ est le modèle estimé de la classe C_m.

Les premiers essais utilisant des modèles mono-gaussiens (avec matrices de covariance pleines ou diagonales) n'ayant pas été concluants, nous avons décidé d'utiliser des modèles multi-gaussiens obtenus par adaptation d'un modèle du monde initial appris sur le signal de test. La Figure II. 4 présente une vue générale de notre étape de regroupement hiérarchique.

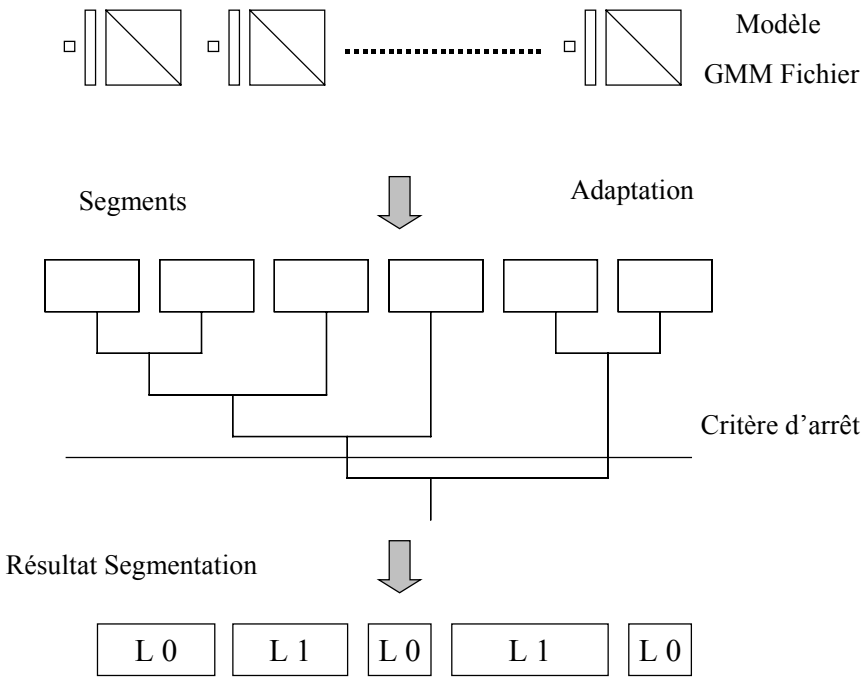


Figure II. 4 Le regroupement hiérarchique de notre système de segmentation

Nous commençons par entraîner d'abord en utilisant l'algorithme EM un modèle de mixture de gaussiennes (GMM) d'ordre 32 avec des matrices de covariance diagonales sur le

signal de test complet. Le modèle sera ensuite adapté par MAP sur chaque segment et à chaque étape du regroupement qui est fait jusqu'à obtenir le dendrogramme complet. Seulement les moyennes sont adaptées. La taille du modèle a été choisie pour être un compromis raisonnable entre la complexité du modèle, la quantité de données disponibles pour l'entraînement (la taille du fichier de test) et le temps total du calcul. Le choix d'adapter uniquement les moyennes est le résultat d'un premier réglage du système.

3.4. Premiers résultats : Campagne d'évaluation NIST SpRec 2002

Les premiers résultats officiels du système de segmentation proposé ont été obtenus lors de la campagne d'évaluation en reconnaissance du locuteur NIST SpRec 2002.

Pour étudier l'influence de la détection de changements de locuteur sur le résultat final de segmentation, nous avons soumis deux systèmes. Le premier est le système que nous venons de décrire. La détection de changements de locuteur est la détection BIC présentée dans la section 3.2. Le deuxième système se différencie par l'étape de détection de changements de locuteur qui est remplacée par une pré-segmentation en segments de taille fixe (0,75 secondes). La durée a été choisie suffisamment faible pour que chaque segment contienne uniquement la parole d'un seul locuteur. Les deux détecteurs de changements de locuteur sont présentés dans la Figure II. 5.

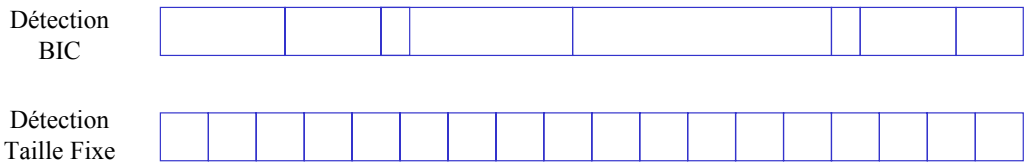


Figure II. 5 Détecteurs de changements de locuteur : BIC et Taille Fixe

Tableau II. 2 Résultats officiels NIST SpRec 2002 ; pourcentage d'erreur de segmentation et temps d'exécution pour les systèmes CLIPS

Changements de locuteur	Phone 2002	BN 2002	Meet 2002	Temps CPU
BIC	16,58 %	30,33 %	50,20 %	37 h
Taille Fixe	17,12 %	30,36 %	50,05 %	120 h

Les résultats présentés dans le Tableau II. 2 montrent que la détection de changements de locuteur BIC améliore légèrement mais pas de façon significative (mois de 1 %) les résultats dans le cas de données téléphoniques et journaux télévisés. Cependant, le temps d'exécution (Temps CPU) est trois fois plus important dans le cas de l'utilisation de la pré-segmentation en segments de taille fixe. La taille totale du corpus est d'environ 14 heures ce qui nous mène à un temps d'exécution d'environ 2,5 fois le temps réel dans le premier cas et plus de 10 fois le temps réel dans le cas des segments de taille fixe. Le nombre de segments issus de la détection de changements de locuteur est beaucoup plus grande dans ce cas et la complexité de l'algorithme de regroupement hiérarchique augmente de façon exponentielle avec le nombre de segments (la taille du dendrogramme présenté dans Figure II. 4 est plus importante).

Les résultats obtenus sont comparables avec les résultats des autres participants et dans le cas des données de type réunion nous avons obtenus le meilleur taux d'erreur parmi quatre participants.

Les résultats obtenus montrent la difficulté croissante des tâches. Les données téléphoniques contiennent seulement deux locuteurs et de la parole de degré de spontanéité acceptable. Les données de type journaux télévisés contiennent un grand nombre de locuteurs et aussi des conditions acoustiques très différentes (bande large, bande étroite, parole propre, parole bruitée, etc.). Par contre, le fait que les données BN contiennent de la parole préparée avec de longues interventions pour chaque locuteur réduit la difficulté de la tâche par rapport aux données de réunions. La difficulté principale de ces dernières réside dans le caractère complètement spontané de la parole. Les locuteurs parlent beaucoup en même temps (Tableau II. 1) et aussi la taille des interventions est très variable (moins d'une seconde jusqu'à plusieurs dizaines de secondes). Les données de réunion Meet 2002 contiennent chaque réunion enregistrée simultanément de deux façons différentes : une fois avec un seul microphone positionné au milieu de la table et une fois avec des microphones cravates pour chaque participant. Dans ce dernier cas, la somme des signaux de chaque microphone a été utilisée comme signal de test. Des test contrastifs réalisés entre ces deux conditions montrent une faible différence (environ 1,5 % en absolu) entre les performances obtenues. Ceci montre que la difficulté des données type réunion n'est pas essentiellement due aux conditions d'enregistrement mais au caractère spontané de la parole qui reste identique dans les deux cas.

4. Amélioration du système initial (2002-2003)

Les améliorations du système de segmentation présenté dans la section précédente concernent deux domaines : la paramétrisation et l'estimation du nombre de locuteurs. Les deux améliorations sont liées aux contraintes dues au changement de type de données dans les évaluations NIST entre les années 2002 et 2003 (enregistrements téléphoniques et journaux de courte durée abandonnés au profit d'enregistrements de journaux de longue durée).

4.1. La paramétrisation

La parametrisation de notre premier système présentée dans la section 3.1 est issue de l'état de l'art, c'est-à-dire des systèmes présentés par les participants à l'évaluation NIST 2001. Elle est normalement adaptée aux données NIST 2001 qui sont des enregistrements téléphoniques issus du corpus SwitchBoard [Godfrey 1992].

Les données téléphoniques SwitchBoard sont enregistrées à une fréquence d'échantillonnage de 8 kHz. Les données de type journaux télévisés BN 2002, qui ont fait leur apparition dans les évaluations à partir de l'année 2002 sont enregistrées à une fréquence d'échantillonnage de 16 kHz.

Les coefficients cepstraux calculés sont uniformément repartis selon une échelle Mel dans la bande passante du signal. La bande passante étant différente pour les deux types de données, pour retrouver la même répartition de coefficients nous devons par exemple augmenter le nombre de filtres utilisés pour le calcul des coefficients cepstraux.

La figure suivante présente le taux d'erreur de segmentation pour différents nombres de filtres utilisés pour le calcul de coefficients MFCC (le système présenté dans la section 3 utilisait 24 filtres).

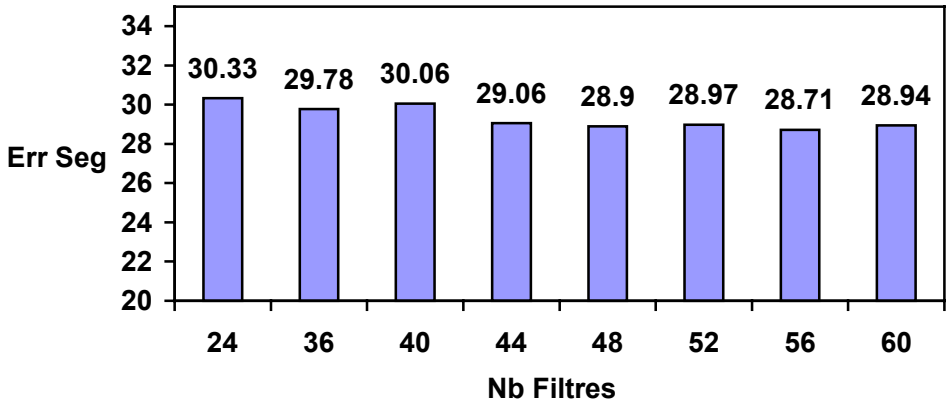


Figure II. 6 Pourcentage d'erreur de segmentation en fonction du nombre de filtres utilisés pour le calcul de coefficients MFCC sur le corpus BN 2002

Etant donné que la bande passante du signal de journaux télévisés est deux fois plus grande que la bande passante de signaux téléphoniques, des améliorations non-négligeables sont obtenues à partir de 48 bancs de filtres, soit deux fois le nombre de filtres utilisés pour les données téléphoniques. Les meilleurs résultats sont obtenus pour 56 bancs de filtres, soit une amélioration de 1,6 % en absolu.

4.2. Estimation du nombre de locuteurs

Le critère d'arrêt pour l'algorithme de regroupement présenté dans la section 3.3 était le nombre de locuteurs supposé égal à deux. L'hypothèse est vraie dans le cas des enregistrements téléphoniques et on peut supposer qu'elle génère des taux d'erreurs raisonnables pour des enregistrements de courte durée où il n'existe généralement pas plus de deux locuteurs majoritaires. Dans le cas de longs (à partir de 10 minutes) enregistrements de journaux télévisés ou de réunions, une estimation ou une connaissance *a priori* du nombre de locuteurs présents est indispensable.

Une possibilité est d'utiliser l'un des critères présentés dans la section 3.3.3 de la partie Etat de l'art. Le plus utilisé semble être le critère d'information bayésien pénalisé par la complexité des modèles, qui selon les études mathématiques, donne en ce qui concerne le modèle sélectionné un meilleur compromis entre les performances et la complexité du modèle. Ce critère a été largement utilisé pour la détection de changements de locuteurs [Gish 1991, Chen 1998, Delacourt 2000b] mais il existe peu de travaux concernant son utilisation pour l'estimation du nombre de locuteurs pour la segmentation en locuteurs [Delacourt 2000a,

Reynolds 2003]. A titre d'exemple, lors des évaluations NIST SpRec 2002, un seul système estimait de façon automatique le nombre de locuteurs présents dans un enregistrement, il s'agissait du système de segmentation en locuteurs du laboratoire LIA. Cependant cette estimation n'utilisait pas le critère BIC.

Nous allons sélectionner alors le nombre de locuteurs N_{SP} qui maximise :

$$BIC(M) = \log P(X/M) - \lambda \frac{m}{2} N_{SP} \log N_X \qquad \text{(II. 8)}$$

où M est le modèle composé par les N_{SP} modèles de locuteurs et N_X est le nombre total de trames utilisées pour le calcul. Le paramètre λ est une constante empirique fixée à 0.6. Le dernier paramètre m représente la complexité de chaque modèle de locuteurs.

Dans le cas d'un modèle de mixture de gaussiennes multidimensionnel GMM de taille p (vecteurs acoustiques de taille p) d'ordre n (n distributions) selon la composition des matrices de covariance, diagonale ou non, la complexité est donnée par l'une des formules suivantes :

$$m_{pleine} = \left(p + \frac{1}{2} p(p+1) \right) \cdot n \qquad \text{(II. 9)}$$

$$m_{diag} = 2 \cdot p \cdot n \qquad \text{(II. 10)}$$

L'originalité de notre approche ne réside pas dans le critère lui même, ni dans le choix des paramètres du calcul mais dans la façon de calculer la première partie de l'équation II. 8, c'est-à-dire la probabilité des données étant donné les modèles.

Soit X_i et M_i les données et respectivement le modèle du locuteur i. Le modèle est obtenu par adaptation MAP à partir du modèle du monde calculé sur le fichier de test, comme dans la section 3.3. Nous pouvons alors écrire :

$$P(X/M) = P\left((X_1, \ldots, X_{N_{SP}}) / (M_1, \ldots, M_{N_{SP}}) \right) \qquad \text{(II. 11)}$$

Si maintenant nous faisons l'hypothèse que les données X_i dépendent uniquement du modèle M_i alors nous avons :

$$P\left(X_i / \left(X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_{N_{SP}}; M_1, \ldots, M_{N_{SP}}\right)\right) = P\left(X_i / M_i\right) \qquad \text{(II. 12)}$$

Cette hypothèse veut dire que les données d'un certain locuteur ne dépendent pas des données des autres locuteurs et aussi ne dépendent pas des modèles des autres locuteurs.

L'hypothèse peut être considérée comme vraie si les locuteurs ne parlent pas en même temps. Si les locuteurs parlent en même temps alors on peut dire que les segments multi-locuteurs dépendent de plusieurs modèles de locuteurs (ceux des locuteurs qui parlent sur le segment en question).

Nous pouvons alors écrire, en utilisant la formule du Bayes sur l'équation II. 11 :

$$P(X/M) = P\left(\left(X_1, \ldots, X_{N_{SP}}\right) / \left(M_1, \ldots, M_{N_{SP}}\right)\right) = \frac{P\left(\left(X_1, \ldots, X_{N_{SP}}\right); \left(M_1, \ldots, M_{N_{SP}}\right)\right)}{P\left(M_1, \ldots, M_{N_{SP}}\right)} =$$

$$= \frac{P\left(X_1 / \left(X_2, \ldots, X_{N_{SP}}\right); \left(M_1, \ldots, M_{N_{SP}}\right)\right) P\left(\left(X_2, \ldots, X_{N_{SP}}\right); \left(M_1, \ldots, M_{N_{SP}}\right)\right)}{P\left(M_1, \ldots, M_{N_{SP}}\right)} =$$

$$= \frac{P\left(X_1 / M_1\right) P\left(\left(X_2, \ldots, X_{N_{SP}}\right); \left(M_1, \ldots, M_{N_{SP}}\right)\right)}{P\left(M_1, \ldots, M_{N_{SP}}\right)} =$$

$$= \ldots = \frac{P\left(X_1 / M_1\right) P\left(X_2 / M_2\right) \ldots P\left(X_{N_{SP}}; \left(M_1, \ldots, M_{N_{SP}}\right)\right)}{P\left(M_1, \ldots, M_{N_{SP}}\right)} \qquad \text{(II. 13)}$$

Alors la probabilité $P(X/M)$ peut s'écrire finalement :

$$P(X/M) = \prod_{i=1}^{N_{SP}} P\left(X_i / M_i\right) \qquad \text{(II. 14)}$$

Les résultats obtenus en estimant ainsi le nombre de locuteurs, pour le corpus BN 2002 sont présentés dans Tableau II. 3. La paramétrisation correspond à la meilleure trouvée dans la section 4.1 (56 bancs de filtres). Nous obtenons une réduction d'erreur de 8 % en absolu, ce qui montre la nécessité de l'estimation du nombre de locuteurs dans la segmentation de documents de type journaux télévisés, même si les extraits sont de courte durée (environ 2 minutes).

Tableau II. 3 Erreur de segmentation en fixant ou en estimant le nombre de locuteurs pour le corpus BN 2002

Nombre de locuteurs	Fixé 2	Estimé
BN 2002	28,71 %	20,74 %

Nous allons faire ici quelques commentaires concernant le résultat de cette estimation. Normalement nous devons obtenir le nombre réel de locuteurs présents dans l'enregistrement. Mais, il ne faut pas oublier que nous utilisons un critère d'arrêt pour le regroupement hiérarchique. Nous verrons plus loin dans nos expériences que le nombre de locuteurs réel ne correspond pas toujours au nombre de locuteurs optimal lorsqu'il s'agit d'avoir les meilleures performances de segmentation après regroupement.

5. Ajout des modules LIA dans notre système de référence (2003-2004)

Les modules présentés dans cette section ont été développés au Laboratoire d'Informatique d'Avignon (LIA). Leur présentation est indispensable pour la compréhension de certains résultats expérimentaux qui seront présentés plus tard. Ces modules ont été ajoutés à notre meilleur système décrit dans le paragraphe précédent.

5.1. La pré-segmentation acoustique

Nous avons vu dans la partie état de l'art qu'une des étapes importantes d'un système de segmentation est la pré-segmentation acoustique qui a comme but principal de sélectionner uniquement les segments contenant de la parole.

Une fois que les segments contenant de la parole ont été sélectionnés, une classification plus fine peut être envisagée. Les segments peuvent être ensuite classés par genre ou par bande passante (bande large, bande étroite). Nous pouvons également identifier les segments contenant de la musique superposée sur la parole. Ce type de segments est spécifique aux enregistrements de journaux télévisés.

Au moment des premières évaluations auxquelles nous avons participé (NIST 2002) l'information sur les segments contenant de la parole était fournie par les organisateurs. L'étape de pré-segmentation acoustique n'était donc pas nécessaire. A partir de l'année 2003,

cette information n'était plus donnée et les erreurs de classification parole/non-parole étaient incluses dans le calcul d'erreur de segmentation totale.

Dans [Meignier 2004], une approche hiérarchique est proposée sur trois niveaux différents :

- parole/non-parole ;
- parole studio (bande large)/parole plus musique/parole téléphonique (bande étroite) ;
- parole homme/parole femme.

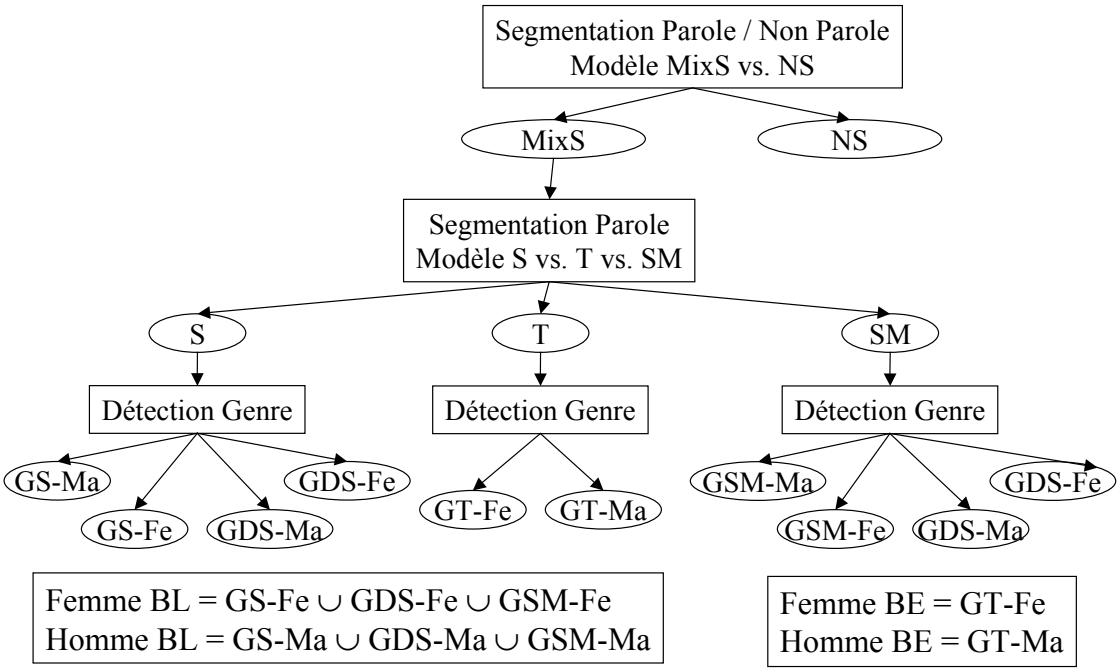


Figure II. 7 La pré-segmentation acoustique

Le premier niveau de classification est réalisé en utilisant deux modèles MixS (parole) et NS (non-parole). Le premier modèle correspond à toutes les conditions parole (homme, femme bande large, bande étroite, etc.) tandis que le deuxième modèle correspond à un mélange de conditions non-parole (musique, silence, bruit). La segmentation se résume ensuite à la sélection du meilleur modèle pour chaque trame de signal.

La deuxième étape de pré-segmentation utilise trois modèles : parole propre (S), parole mélangée à de la musique (SM) et parole téléphonique ou bande étroite (T) pour différencier les segments étiquetés comme parole à l'étape précédente.

Le dernier niveau est composé de l'étape de détection de genre. En fonction du résultat du niveau précédent, chaque segment sera identifié comme étant prononcé par un homme ou par une femme selon les modèles G(S/SM/T)-Ma et G(S/SM/T)-Fe.

Deux autres modèles GDS-Fe et GDS-Ma qui représentent respectivement la parole dégradée femme et homme enregistrée dans des conditions de moins bonne qualité sont utilisés aussi pour éliminer l'influence de certains bruits sur le résultat de la classification.

La paramétrisation utilisée pour la pré-segmentation acoustique est complètement différente par rapport à la paramétrisation utilisée pour la segmentation en locuteurs. Un nombre de 39 (12 MFCC plus énergie plus les paramètres dérivés du premier et second ordre) coefficients sont calculés toutes les 10 ms sur des fenêtres Hamming de taille 25 ms. Tous les modèles utilisés sont des GMMs avec des matrices de covariances diagonales. A l'exception des modèles NS et MixS caractérisés respectivement par 1 et 512 composantes, tous les autres modèles sont de 1024 composantes gaussiennes.

Quatre classes acoustiques sont obtenues à la fin selon les règles présentées dans la Figure II. 7 : femme bande large, homme bande large, femme bande étroite, homme bande étroite.

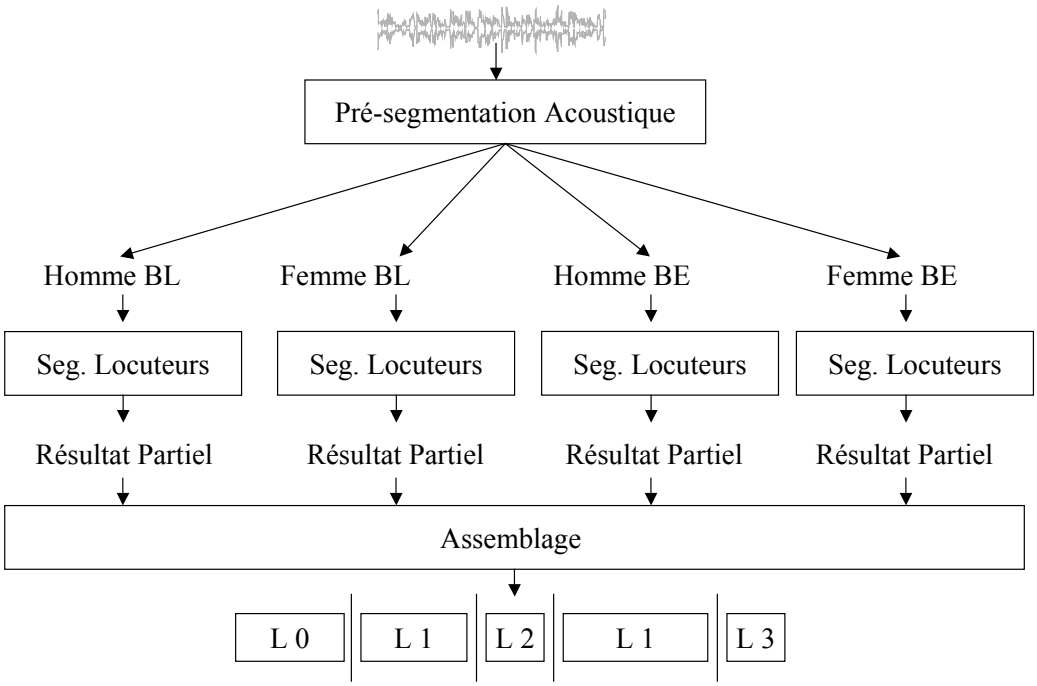


Figure II. 8 Segmentation en locuteurs appliquée sur chaque classe acoustique

Une fois les segments appartenant à chaque classe acoustique identifiés, le système de segmentation est ensuite appliqué indépendamment sur chaque classe. A la fin, les résultats

partiaux obtenus sur chaque classe sont assemblés pour obtenir le résultat final de segmentation (Figure II. 8).

Nous avons fait ici l'hypothèse que chaque locuteur ne pourra être présent que dans une seule des quatre classes acoustiques. En ce qui concerne le genre, l'hypothèse est évidement vraie sauf erreur de pré-segmentation. En ce qui concerne la bande passante du signal (large ou étroite), nous faisons l'hypothèse que dans un journal télévisé, le locuteur est par exemple soit dans le studio soit au téléphone mais jamais dans les deux conditions lors d'une même émission.

5.2. La re-segmentation

Nous avons vu dans la partie état de l'art que certains chercheurs ont complété la segmentation en locuteurs par une phase de re-segmentation. Sylvain Meignier propose dans [Meignier 2000] un système de segmentation en locuteurs basé sur un modèle de Markov (HMM) évolutif qui sera ensuite complété par une étape de re-segmentation.

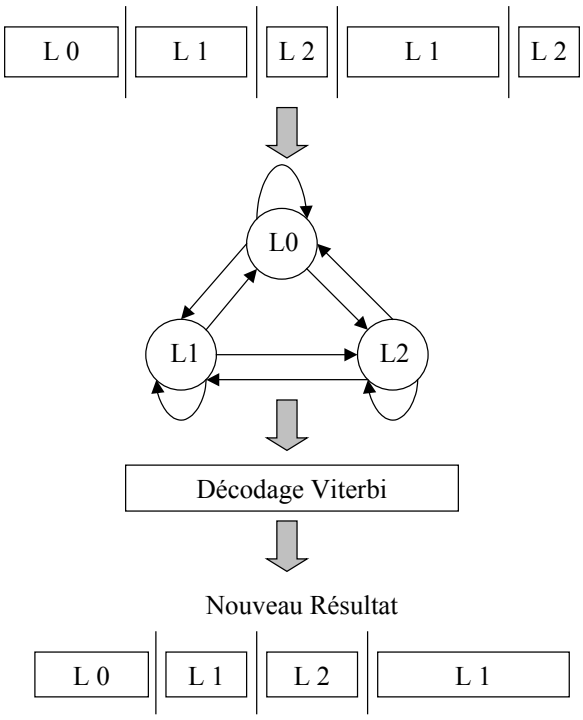


Figure II. 9 Re-segmentation à partir d'un HMM ergodique, un état par locuteur

A partir d'un résultat de segmentation en locuteurs, un HMM ergodique peut être appris. Chaque état correspond à un locuteur identifié par le système. Les probabilités de

transitions sont fixées et seulement les distributions GMM de chaque état sont apprises à partir du résultat de segmentation.

Les distributions GMM correspondant à chaque état du HMM sont apprises par adaptation MAP (moyennes seulement) à partir d'un modèle du monde. Le modèle du monde est un GMM diagonal d'ordre 128 appris sur un sous-ensemble du corpus Hub-4 1996.

Le rôle principal de la re-segmentation est de redistribuer les données de test entre les différents locuteurs présents. Aussi comme la re-segmentation est appliquée après l'assemblage des résultats obtenus séparément sur chaque classe acoustique, nous avons l'occasion de corriger certaines erreurs dues à l'étape de pré-segmentation acoustique.

Cette re-segmentation sera utilisée pour améliorer le résultat de segmentation de notre système et aussi dans un cas où nous fusionnerons plusieurs résultats de segmentation.

6. Résultats du système de référence final : Evaluations RT 2003 et ESTER

Les résultats présentés dans cette section sont les derniers résultats obtenus sans aucune information *a priori*. Ils ont été obtenus lors des dernières évaluations concernant la transcription enrichie de documents de type journaux télévisés et radiodiffusés : RT 2003 et ESTER.

Tableau II. 4 Pourcentage d'erreur de segmentation de notre système de référence pour les corpus RT 2003 et ESTER sans aucune information *a priori* et sans re-segmentation LIA

Corpus	Pré-Segmentation Acoustique	Langue	Parole Manquée	Fausse Alarme Parole	Erreur Locuteur	Erreur de Segmentation Totale
RT 2003 Dev	LIA	Anglais	2,70 %	3,80 %	13,20 %	19,65 %
RT 2003 Eva	LIA	Anglais	2,00 %	2,90 %	14,30 %	19,25 %
ESTER Dev	Aucune	Français	0,00 %	0,00 %	13,85 %	13,85 %
ESTER Eva	Aucune	Français	0,00 %	0,00 %	17,75 %	17,75 %

Les résultats obtenus sur les corpus ESTER sont cohérents avec ceux obtenus sur les corpus RT 2003 en ce qui concerne l'erreur locuteur seule. La différence entre les erreurs de segmentation totales obtenues sur les deux corpus correspond à l'erreur de détection parole/non-parole (parole manquée et fausse alarme parole). Dans les conditions d'évaluation

ESTER, les segments non-parole sont exclus du calcul de l'erreur performance de segmentation ce qui explique les valeurs nulles de *fausse alarme parole* et de *parole manquée*, cependant ils ne sont pas exclus de l'entrée du système de segmentation. Comme on pouvait s'y attendre, la langue des documents n'a pas d'influence sur l'erreur de segmentation.

Si nous regardons en détail la sortie du système, concernant les erreurs, une partie des erreurs de segmentation est produite par les locuteurs *minoritaires* qui parlent très peu par rapport à la totalité de l'enregistrement (2 secondes sur un document de 30 minutes). Ces locuteurs n'ont pas suffisamment de données pour produire un modèle fiable et ils sont presque impossibles à détecter. Ceci n'a pas une grande influence sur l'erreur de segmentation qui correspond à un pourcentage de la totalité du temps de test, mais ceci pourrait poser des problèmes dans une tâche consistant à rechercher ces locuteurs "rares" en particulier. En effet, les locuteurs "rares" peuvent être justement ceux qui apportent le plus d'information dans une tâche de recherche par exemple. En cela, le critère d'évaluation NIST n'est pas idéal.

Enfin, pour mieux interpréter les résultats présentés il faut savoir que selon les chercheurs dans le domaine [NIST 2003], un erreur d'environ 6 % peut être due uniquement à la synchronisation entre la transcription de référence et le résultat du système. A partir de cette conclusion, nous pouvons dire qu'un document qui obtiendrait 6 % d'erreur de segmentation serait correctement segmenté.

7. Conclusion

Cette partie a été dédiée à la segmentation en locuteurs de documents audio. Dans un premier temps nous avons présenté notre système de segmentation initial issu en totalité de l'état de l'art. Ce système nous a permis de participer aux évaluations NIST SpRec 2002 sur la segmentation en locuteurs.

Ce système a été ensuite amélioré au niveau de la paramétrisation et aussi en lui rajoutant un module d'estimation du nombre de locuteurs présents dans un enregistrement. Ce module est indispensable pour les enregistrements de type journaux télévisés et radiodiffusés proposés dans les évaluations NIST et ESTER qui contiennent un grand nombre de locuteurs (plusieurs dizaines pour un enregistrement d'une heure environ).

Malgré les progrès faits par rapport à notre système de segmentation initial, nous sommes assez loin d'une segmentation parfaite. Plusieurs directions de recherche sont envisageables pour réduire le taux d'erreur actuel qui se situe entre 15 % et 20 %. Une

meilleure détection des silences peut conduire à une réduction d'erreur de 4 % environ. Une détection plus précise des début et fin d'interventions de chaque locuteur peut conduire également à une réduction d'erreur.

Par hypothèse, spécifique au cadre des évaluations NIST et ESTER, nous ne disposons d'aucune information concernant les documents analysés pour la segmentation en locuteurs. Nous allons voir maintenant dans la troisième partie que cette contrainte n'est pas toujours aussi stricte dans la réalité, et nous évaluerons l'apport d'informations *a priori* pour la segmentation en locuteurs.

Troisième partie :

Intégration d'informations *a priori* dans un système de segmentation en locuteurs

Par hypothèse de travail la segmentation en locuteurs de documents audio est généralement faite sans aucune information *a priori* : sans données de référence pour le locuteurs, sans connaître le nombre de locuteurs, etc.. Cette partie est dédiée à l'intégration d'informations *a priori* dans un système de segmentation en locuteurs et à l'évaluation de ce que cela peut apporter en terme de performances. Deux grandes catégories d'informations seront discutées : les information issues d'une annotation manuelle (certaines) et les informations issues d'une annotation automatique (incertaines). Des résultats expérimentaux seront présentés et discutés pour chaque type d'information.

1. Introduction

Cette section concerne la partie la plus importante de notre travail exploratoire sur la segmentation en locuteurs. Nous allons envisager ici l'utilisation d'informations supplémentaires pour la segmentation en locuteurs. Ces informations correspondent à des connaissances *a priori* sur les enregistrements analysés comme la connaissance ou non du nombre de locuteurs par exemple.

Rappelons tout d'abord que l'hypothèse faite jusqu'ici en segmentation en locuteurs est qu'aucune information *a priori* n'est disponible. Ceci signifie que nous n'avons aucune information concernant le nombre de locuteurs présents dans l'enregistrement, concernant la nature de la parole (bande large, bande étroite, préparée, spontanée), concernant la présence ou l'absence d'un genre (uniquement locuteurs femme par exemple). Par ailleurs, aucune donnée audio de référence n'est censée être disponible pour aucun des locuteurs. Le système présenté dans la partie précédente respectait toutes ces contraintes.

Ces limitations ne sont pas aussi strictes dans la réalité, notamment dans certaines applications et conditions de test pour lesquelles on peut facilement avoir des informations *a priori* sur les documents analysés. Par exemple, le type de document (journal télévisé, conversation téléphonique, réunion) est généralement connu. Aussi, dans certains cas, nous pouvons facilement obtenir des modèles *a priori* pour certains locuteurs. Par exemple, dans le cas d'une réunion nous pouvons demander à tous les participants de se présenter au début pour obtenir quelques secondes de données de référence pour chaque locuteur. Dans le cas de journaux télévisés, nous pouvons facilement avoir des données de référence pour le présentateur du journal à partir de journaux précédents, ainsi que pour quelques personnalités qui reviennent fréquemment dans l'actualité.

Tableau III. 1 Différences entre documents audio (données SpRec 2002)

Information	Téléphonique	Réunion	Journaux TV
Tour de parole moyen	1,69 sec	1,12 sec	5 sec
Probabilité du locuteur majoritaire	68 %	49 %	55 %
Nombre de locuteurs	2	Peut être connu *a priori*	Toujours Inconnu
Données de référence	Possible pour les 2 locuteurs	Possible pour tous les locuteurs	Possible pour certains locuteurs seulement

En regardant le Tableau III. 1 on remarque qu'il existe des différences entre les types de documents. Par exemple, la durée moyenne du tour de parole est largement supérieure pour les documents de type journaux télévisés. Nous pouvons s'attendre alors à des taux d'erreurs meilleurs par rapport aux documents de type réunion.

Un système de segmentation classique n'utilise pas les différences entre les documents à son avantage. C'est exactement ce que nous allons essayer de faire dans cette partie. Avant d'entrer plus dans les détails, il faut se rendre compte que pour utiliser des informations *a priori*, en plus d'une connaissance des données, il faut aussi bien connaître les points forts et les points faibles du système de segmentation. Si par exemple le système de segmentation détecte déjà bien le nombre de locuteurs automatiquement une telle information *a priori* sera inutile. Dans la suite de cette partie, nous distinguerons les informations *a priori* qu'il est possible d'obtenir par annotation manuelle (informations *certaines*), des informations *a priori* qu'il n'est possible d'obtenir que par annotation automatique. Ces dernières informations seront donc considérées comme *incertaines* puisque l'annotation automatique peut avoir fait des erreurs.

2. Informations *a priori* obtenues par annotation manuelle (certaines)

Les informations que nous avons présenté dans le Tableau III. 1 sont dans la grande majorité des informations qui peuvent être obtenues à partir d'une intervention manuelle. Nous pouvons également les appeler informations certaines. Nous pouvons savoir par exemple avec certitude que dans une réunion il y a un nombre exact de locuteurs connu grâce à une liste de présence par exemple, ou nous disposons de données de référence fiables pour

apprendre des modèles de locuteurs à l'issue de l'annotation manuelle d'une partie d'un document. Dans les sous-sections suivantes, nous envisageons différentes informations manuelles et évaluons leur apport pour notre système de segmentation en locuteurs.

2.1. Connaissance *a priori* du nombre de locuteurs

Nous avons vu dans le Tableau III. 1 que dans certains cas il est possible de connaître le nombre réel de locuteurs présents dans un enregistrement. Dans le cas de données téléphoniques classiques nous savons qu'il y a seulement deux locuteurs, dans le cas de données de type réunion le nombre de locuteurs est supérieur à deux mais peut être connu (à partir d'une liste de participants ou d'un compte rendu).

Le cas le plus difficile semble être les données de type journaux télévisées pour lesquelles le nombre de locuteurs ne peut pas être connu à l'avance. La seule possibilité est alors de l'estimer de façon automatique.

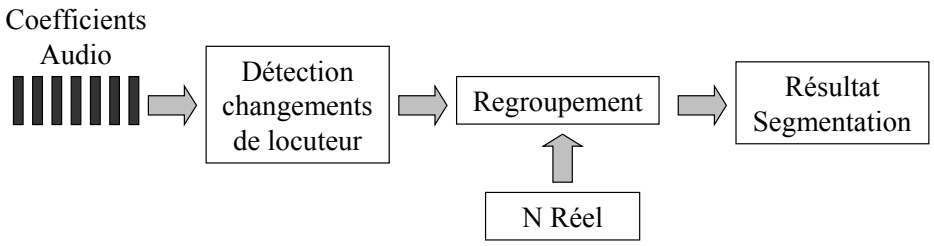


Figure III. 1 Système de segmentation quand le nombre de locuteurs est connu

La Figure III. 1présente ce que pourrait être un système de segmentation lorsque le nombre réel de locuteurs est connu à l'avance. Cependant, comme le processus de regroupement est un processus automatique qui peut générer des erreurs il n'est pas évident que le meilleur critère d'arrêt soit le nombre réel de locuteurs. Nous avons remarqué lors des expérimentations qu'il est difficile de détecter les locuteurs minoritaires, qui parlent très peu sur l'enregistrement (exemple : 2 secondes sur un document de 30 minutes). Nous pouvons dire alors que ce que nous souhaitons détecter est plutôt le nombre optimal de locuteurs, c'est à dire le nombre de locuteurs qui minimise l'erreur de segmentation. Nous rappelons ici que l'erreur de segmentation utilisée lors des évaluations NIST et ESTER est un pourcentage de temps mal classifié. Une erreur de quelques secondes qui correspond à un locuteur minoritaire non-détecté a peu d'influence directe sur l'erreur de segmentation. Vu la remarque précédente

concernant les locuteurs minoritaires, nous pouvons donc nous attendre à ce que le nombre optimal de locuteurs soit inférieur au nombre réel de locuteurs.

En ce qui concerne le nombre estimé de locuteurs, une valeur petite du paramètre de réglage λ de la formule II. 8 de la partie précédente peut générer un plus grand nombre de locuteurs. Le paramètre devra alors être réglé pour s'approcher du nombre optimal de locuteurs. En conséquence, on peut s'attendre à ce que :

$$N\,Optimal \leq N\,Estimé \leq N\,Réel \qquad \text{(III. 1)}$$

Pour justifier l'inégalité III. 1, nous avons fait des tests sur tous les corpus de journaux télévisés dont nous disposons. Nous avons considéré trois situations pour le nombre de locuteurs utilisé comme critère d'arrêt pour le regroupement : a) le nombre de locuteurs est fixé comme étant égal au nombre de locuteurs réel, b) le nombre de locuteurs est estimé de façon automatique selon la méthode proposée dans le paragraphe 4.2 de la partie précédente, c) le nombre de locuteurs est fixé comme étant égal au nombre de locuteurs optimal. Le nombre de locuteurs optimal est déduit *a posteriori* du résultat de segmentation.

Tableau III. 2 Pourcentage d'erreur de segmentation pour différents critères d'arrêt sur le nombre de locuteurs présents dans un enregistrement

Corpus	BN 2002	RT 2003 Dev	RT 2003 Eva	ESTER Dev	ESTER Eva
Optimal	16,50 %	14,54 %	14,03 %	12,85 %	16,52 %
Estimé	20,74 %	19,65 %	19,25 %	13,68 %	17,75 %
Réel	22,50 %	24,76 %	16,29 %	14,32 %	19,64 %

Tableau III 3 Nombre total de locuteurs utilisé comme critère d'arrêt dans les différents corpus (cumulé sur tous les fichiers d'un même corpus)

Corpus	BN 2002	RT 2003 Dev	RT 2003 Eva	ESTER Dev	ESTER Eva
Optimal	225	46	56	147	154
Estimé	242	47	50	147	153
Réel	300	69	57	147	154

Les résultats présentés dans le Tableau III. 2 montrent que pour notre système il est généralement mieux d'estimer le nombre de locuteurs plutôt que d'utiliser la connaissance *a priori* du nombre réel de locuteurs. Il existe un seul cas (RT 2003 Eva) où l'erreur de

segmentation obtenue avec la connaissance *a priori* du nombre de locuteurs est inférieure à l'erreur dans le cas de l'estimation du nombre de locuteurs. Nous pouvons remarquer également que la différence entre l'erreur de segmentation obtenue dans le cas optimal et celle obtenue dans le cas estimé est moins importante dans le cas du corpus ESTER. L'explication vient du modèle du monde utilisé qui est calculé à partir du document de test courant. La durée de documents dans le cas du corpus BN 2002 est de 2 minutes, dans le cas RT 2003 Dev de 10 minutes et dans le cas RT 2003 Eva de 30 minutes. Le corpus ESTER est quant à lui composé majoritairement de documents de 60 minutes. Le modèle du monde estimé dans ce dernier cas sur plus de données semble donc plus fiable pour l'estimation du nombre de locuteurs et sans doute aussi pour l'étape suivante de regroupement.

Concernant l'inégalité III. 1, nous pouvons affirmer qu'elle est respectée à une exception près (RT 2003 Eva). Cependant, les trois nombres considérés sont assez proches les uns des autres. Par exemple pour le corpus RT 2003 Eva seulement 7 locuteurs sur un total de 57 présents n'ont pas été détectés. La différence est encore plus faible dans le cas de fichiers de grande taille (ESTER Dev et ESTER Eva) où les trois nombres sont quasiment égaux.

Pour conclure, la connaissance *a priori* du nombre de locuteurs ne semble pas être une information extrêmement utile puisque nous disposons déjà d'un très bon estimateur du nombre de locuteurs, proche d'une estimation optimale en terme d'erreur de segmentation. Cette connaissance *a priori* pourrait cependant être au moins utilisée pour vérifier l'estimation du nombre de locuteurs surtout sur des documents de taille importante.

2.2. Données de référence disponibles pour tous les locuteurs

Dans cette section, nous allons supposer que nous disposons de données de référence pour tous les locuteurs présents dans un enregistrement.

Ceci est tout à fait réaliste dans le cas de données téléphoniques et de réunions. Dans le cas de réunions nous pouvons par exemple demander à tous les locuteurs de se présenter au début de l'enregistrement pour obtenir ainsi quelques secondes nécessaires pour obtenir par adaptation un modèle pour chaque locuteur. Dans le cas de conversations téléphoniques, à partir du moment où il y a seulement deux locuteurs, les données de référence peuvent être facilement obtenues par transcription manuelle d'une courte partie de la conversation.

En utilisant les données de référence, des modèles sont entraînés pour chaque locuteur. Les modèles de locuteur dans notre cas sont obtenus par adaptation MAP (moyennes

seulement) à partir d'un modèle du monde entraîné sur le fichier courant de la même façon que dans la section 3.3 de la partie précédente.

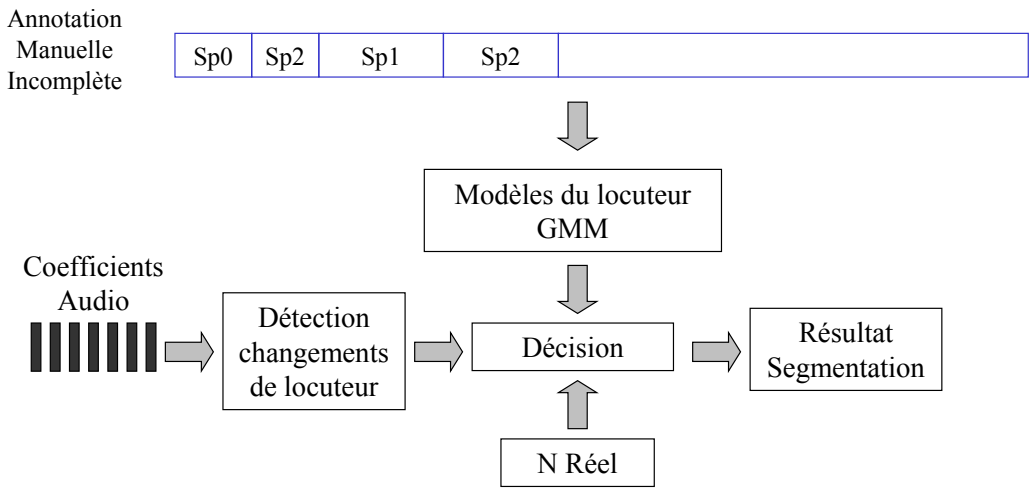


Figure III. 2 Système de segmentation quand des données de référence sont disponibles pour tous les locuteurs

Ensuite, pour chaque segment issu de la détection de changements de locuteur, une décision concernant le locuteur le plus probable est prise. La décision est faite segment par segment plutôt que trame par trame, car la décision prise sur un segment entier est plus fiable que la décision prise sur une seule trame. Pour chaque segment, la probabilité $P(X/M)$ est calculée pour chaque locuteur M et le locuteur le plus probable est sélectionné.

Nous avons étudié le cas ou différentes quantités de données (entre 0 et 60 secondes) sont disponibles pour chaque locuteur. Pour des raisons de comparaison, nous avons présenté aussi l'erreur de segmentation obtenue par le système de segmentation sans aucune information *a priori*. L'erreur de segmentation est dans tous les cas la somme entre l'erreur de classification locuteur et l'erreur de segmentation parole/non-parole. Cette dernière est, par contre, la même dans les deux cas (référence et *a priori*) et en conséquence, toute amélioration du résultat total correspondra à une amélioration de l'erreur *locuteur*.

Comme les données *a priori* pour les locuteurs sont issues des documents de test (comme cela serait le cas sur un document partiellement annoté manuellement), les zones utilisées pour calculer des modèles de locuteurs sont ensuite exclues lors du calcul de l'erreur de segmentation. C'est pour cette raison que la courbe d'erreur de segmentation de référence présentée dans la Figure III. 3 et dans la Figure III. 4 n'est présente pas des valeurs constantes et varie légèrement.

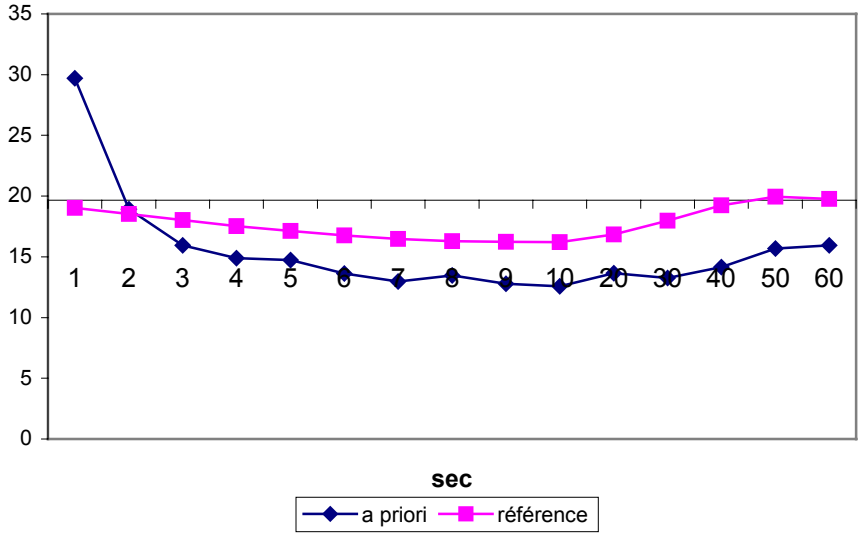


Figure III. 3 Pourcentage d'erreur de segmentation en fonction de la quantité de données *a priori* disponible pour chaque locuteur pour le corpus RT 2003 Dev

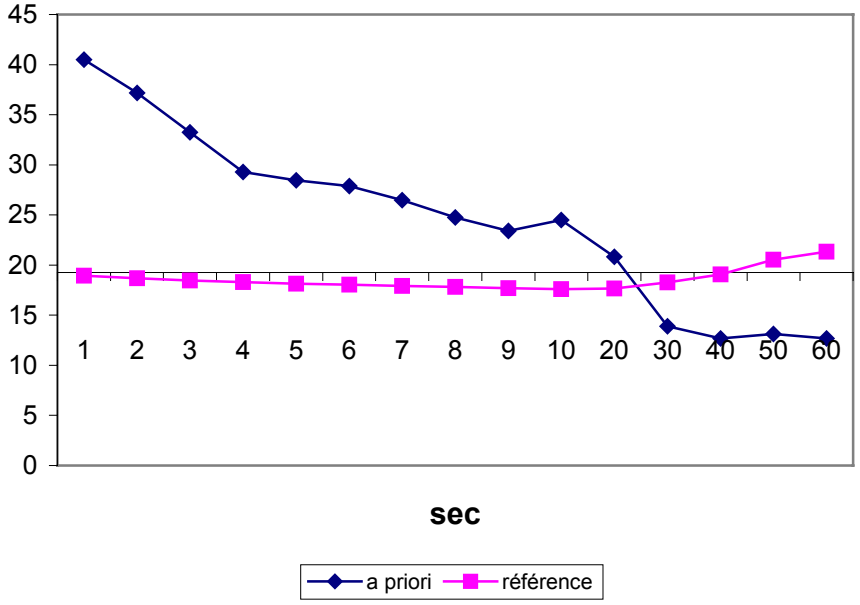


Figure III. 4 Pourcentage d'erreur de segmentation en fonction de la quantité de données *a priori* disponible pour chaque locuteur pour le corpus RT 2003 Eva

Tableau III. 4 Pourcentage d'erreur de segmentation quand des données *a priori* sont disponibles pour tous les locuteurs

Données *a priori*	0 sec	1 sec	2 sec	20 sec	60 sec
RT 2003 Dev *a priori*	19,65 %	29,70 %	18,92 %	13,65 %	15,95 %
RT 2003 Dev Référence	19,65 %	19,04 %	18,52 %	16,85 %	19,77 %
RT 2003 Eva *a priori*	19,25 %	40,51 %	37,19 %	20,82 %	12,67 %
RT 2003 Eva Référence	19,25 %	18,95 %	18,69 %	17,67 %	21,33 %

L'erreur de segmentation obtenue en utilisant des modèles de locuteurs est de plus en plus faible au fur et à mesure qu'augmente la taille des données disponibles pour chaque locuteur. Dans le cas du corpus RT 2003 Dev une amélioration est visible à partir de 2 secondes par rapport au corpus RT 2003 Eva ou une amélioration n'est visible qu'à partir de 20 secondes. Cette différence entre le moment de croisement des deux courbes peut s'expliquer par les tailles de documents qui influent sur la qualité du modèle du monde (10 minutes vs 30 minutes), par le nombre de locuteurs présents dans les documents qui augmente la difficulté de la décision et, le plus important, par le nombre de locuteurs qui n'ont pas suffisamment de données pour avoir un modèle fiable.

En conclusion, les données *a priori* pour tous les locuteurs peuvent être utilisées si nous avons *suffisamment* de données pour chaque locuteur. Cependant la quantité nécessaire variera en fonction du nombre de locuteurs présents dans l'enregistrement.

Concernant ces expérimentations, un autre aspect important est le temps d'exécution. Dans le cas de données *a priori* disponibles (simple décision par maximum de vraisemblance) on passe à 0,1 temps réel par rapport à 3 fois le temps réel au minimum pour le système de segmentation sans données *a priori*.

Nous avons parlé dans cette section du cas ou des données de référence sont disponibles pour tous les locuteurs. Cependant, dans le cas de journaux télévisés, il est généralement impossible d'avoir des données pour tous les locuteurs. Nous pouvons obtenir dans ce cas des données de référence pour quelques locuteurs particuliers seulement (le présentateur du journal, certaines personnalités célèbres). Cet aspect sera abordé dans la section 3.2.

3. Informations *a priori* obtenues pas annotation automatique (incertaines)

Les informations que nous avons évalué dans la section 2 sont ce que nous avons appelé des information certaines. Dans ce paragraphe, nous nous intéressons à des informations appelées incertaines qui sont issues du résultat d'un système de traitement automatique de la parole comme par exemple un système de pré-segmentation acoustique ou un système de suivi du locuteur.

3.1. Une pré-segmentation acoustique disponible *a priori*

Nous avons déjà vu que la pré-segmentation acoustique est une étape généralement indispensable pour un système de segmentation en locuteurs. Son but est d'empêcher que les segments de silence, bruit ou musique soient étiquetés comme des segments contenant de la parole. Le problème qui nous intéresse dans cette section est jusqu'à quel niveau de macro-segmentation en classes acoustiques il faut aller dans une pré-segmentation acoustique pour améliorer un système de segmentation en locuteurs.

Généralement, la pré-segmentation acoustique est utilisée en reconnaissance de la parole continue, notamment dans le cas de la transcription automatique de journaux télévisés pour pouvoir adapter le système de reconnaissance à certaines conditions acoustiques particulières (parole homme, parole femme, parole téléphonique etc.) [Gauvain 2002].

Pour la segmentation en locuteurs, en adoptant uniquement le point de vue de la métrique d'évaluation, une pré-segmentation qui sélectionne uniquement les segments contenant de la parole semble suffisante. Par contre, en regardant le processus de segmentation dans son ensemble, une pré-segmentation acoustique plus fine pourrait être utile. Une pré-segmentation en genre pourrait par exemple empêcher de regrouper dans la même classe un segment appartenant à une femme et un segment appartenant à un homme.

Dans cette section, nous allons expérimenter une pré-segmentation acoustique automatique à plusieurs niveaux :

- parole/non-parole ;
- parole/non-parole et homme/femme ;
- parole/non-parole, homme/femme et bande large/bande étroite (parole studio/parole téléphonique) ;

- parole/non-parole, homme/femme, bande large/bande étroite et parole sur musique/parole dégradée ;

A titre de comparaison, nous avons aussi fait des expérimentations avec une pré-segmentation acoustique manuelle en parole/non-parole, homme/femme et bande large/bande étroite. Il est important de noter que dans la pré-segmentation acoustique manuelle, nous avons également une information implicite sur les changements de locuteur puisque pour les journaux télévisés il y a toujours un segment de silence entre les interventions de deux locuteurs différents.

Nous allons voir dans nos expérimentations si notre système a vraiment besoin d'une pré-segmentation acoustique ou si son étape de regroupement est suffisamment performante pour ne jamais regrouper des segments appartenant à deux classes acoustiques différentes.

Le Tableau III. 5 indique les performances de l'étape de pré-segmentation acoustique présentée dans la section 5.1 de la partie précédente sur les deux parties du corpus RT 2003. Les performances de détection parole/non-parole sont très légèrement différentes par rapport à celles présentées dans le Tableau II. 4 de la partie précédente, parce que nous avons retiré manuellement toutes les publicités pour cette expérience. La raison de procéder ainsi s'explique par le fait que nous ne disposions pas de transcription de référence pour les séquences de publicité et que celles-ci sont difficiles à annoter.

Tableau III. 5 Erreurs de pré-segmentation acoustique sur les corpus RT 2003 Dev et Eva

Corpus	DM Parole	FA Parole	Erreur Genre	Erreur Téléphone/Non-téléphone
RT 2003 Dev	2,3 %	2,2 %	1,5 %	0,1 %
RT 2003 Eva	1,8 %	2,7 %	5,5 %	3,0 %

Tableau III. 6 Erreurs de segmentation en locuteurs sur les corpus RT 2003 Dev et Eva pour différents niveaux de pré-segmentation acoustique utilisés comme information *a priori*

	Système LIA-HMM	Système LIA-HMM	Système CLIPS	Système CLIPS
Pré-segmentation acoustique	RT 2003 Dev	RT 2003 Eva	RT 2003 Dev	RT 2003 Eva
Manuelle S/NS-Genre-T/NT	13,1 %	15,4 %	14,1 %	10,6 %
Automatique S/NS	15,4 %	26,9 %	19,7 %	17,1 %
Automatique S/NS-Genre	14,8 %	27,4 %	18,7 %	18,6 %
Automatique S/NS-Genre-T/NT	15,1 %	18,1 %	19,0 %	18,2 %
Automatique S/NS-Genre-T/S/MS/DS	25,9 %	30,5 %	19,1 %	27,6 %

Nous avons comparé deux systèmes LIA et CLIPS basés sur deux architectures différentes : architecture évolutive et architecture ascendante. Comme on pouvait s'y attendre les meilleurs résultats dans les deux cas sont obtenus avec l'utilisation de la pré-segmentation manuelle. Les résultats les moins performants sont obtenus dans le cas de la pré-segmentation la plus fine (S/NS-Gendre-T/S/MS/DS) qui est inadaptée car les classes acoustiques dans ce cas ne sont pas toujours disjointes : un locuteur (le présentateur par exemple) peut facilement passer de la classe MS (parole sur musique) à la classe S (parole propre). Une pré-segmentation acoustique trop fine conduira alors à deux étiquettes différentes pour un même locuteur et donc à une augmentation du taux d'erreur de segmentation en locuteurs. Pour les trois autres cas restant, on peut voir que l'architecture ascendante (CLIPS) semble moins sensible à la granularité de la pré-segmentation acoustique. La technique utilisant la distance BIC semble se comporter comme une pré-segmentation acoustique intrinsèque, et lors du regroupement, les segments de conditions acoustiques différentes ne sont pas regroupés dans la même classe. Le système CLIPS semble aussi plus robuste par rapport au changement de corpus, les performances sur Dev et Eva étant comparables. Le système LIA-HMM obtient par contre des résultats meilleurs sur le corpus Dev. En ce qui concerne le corpus Eva, le niveau de pré-segmentation acoustique a une influence importante sur le résultat de segmentation. L'explication réside dans le fait que dans le cas du corpus Eva nous avons 20 % de parole téléphonique par rapport au corpus Dev où il y en a seulement 7,7 %.

3.2. Utilisation *a priori* d'un modèle pour certains locuteurs seulement

Nous avons indiqué dans l'introduction que dans le cas de journaux télévisés, il est presque impossible de connaître *a priori* le nombre réel de locuteurs présents dans l'enregistrement et en conséquence, il est impossible d'avoir un modèle *a priori* pour tous les locuteurs présents.

En revanche, il est possible d'avoir des données de référence pour certains locuteurs comme le présentateur et les reporters [Moraru 2004a]. Dans cette section, nous allons nous intéresser uniquement au présentateur. Pour ce locuteur particulier, il est très facile d'obtenir une grande quantité de données à partir d'enregistrements de journaux télévisés qui ont été diffusés dans les jours précédents. La quantité de données est suffisamment importante pour pouvoir nous permettre d'apprendre un modèle acoustique de type GMM directement par l'algorithme EM sans avoir besoin de passer par l'adaptation d'un modèle du monde existant.

Nous pouvons, dans un premier temps, utiliser ce modèle pour enrichir le résultat de la segmentation. Nous pouvons identifier le présentateur parmi les locuteurs détectés par le système de segmentation. Dans ce cas, la segmentation en locuteurs n'est pas modifiée, simplement l'étiquette correspondant au présentateur n'est plus anonyme.

Pour aller plus loin, nous pouvons envisager d'utiliser le modèle du présentateur également pour améliorer les performances du système de segmentation.

La solution que nous proposons ici est d'utiliser le modèle du locuteur connu pour obtenir une pré-segmentation en *parole locuteur connu/parole locuteur inconnu*. Le système de segmentation sera ensuite appliqué uniquement sur la partie du document étiquetée comme locuteur inconnu. A la fin, les deux résultats, celui de la pré-segmentation en locuteur *connu/inconnu* et celui de la segmentation en locuteurs seront assemblés pour produire le résultat final de segmentation.

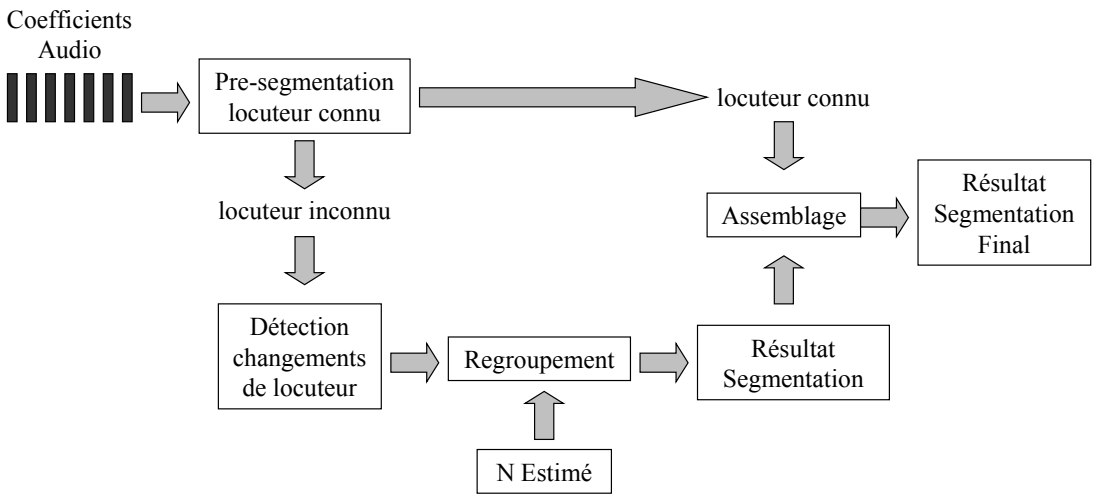


Figure III. 5 Utilisation de données appartenant à un locuteur comme information *a priori* pour la segmentation en locuteurs

Dans l'architecture présentée sur la Figure III. 5, il existe deux processus qui sont en compétition en ce qui concerne le locuteur connu : la pré-segmentation qui utilise le modèle du locuteur connu et le regroupement. Le fait d'avoir un bon modèle pour un locuteur particulier, dans notre cas pour le présentateur du journal télévisé, n'est pas une condition suffisante pour espérer améliorer le résultat de segmentation. Il faut aussi que le locuteur en question génère suffisamment d'erreurs à l'issue du regroupement pour que l'utilisation d'un modèle de référence soit utile.

Nous avons utilisé une partie (4 fichiers sur 6) des deux corpus ESTER (Dev et Eva) pour notre expérimentation. L'erreur de segmentation *initiale* est l'erreur obtenue par le système sans l'aide des données du locuteur connu et constitue notre référence. L'erreur de segmentation *présentateur* est la partie de l'erreur de segmentation initiale qui est due au présentateur du journal. Dans le cas d'une erreur de segmentation *présentateur* importante, une pré-segmentation en locuteur pourra aider le système de segmentation. C'est le cas de trois de fichiers de notre corpus (FrInt 1, FrInt 2, RFI1130 1).

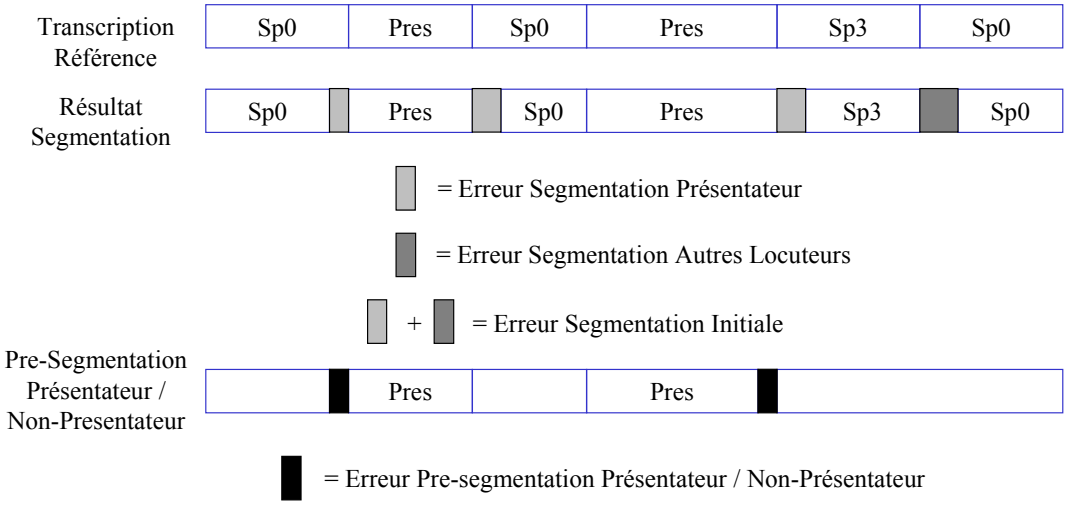


Figure III. 6 Taux d'erreur utilisés dans les résultats concernant les expérimentations de la section 3.2

Tableau III. 7 Erreur de segmentation initiale et erreur de segmentation due au locuteur connu pour le corpus ESTER Dev

Fichier	FrInt 1	FrInt 2	RFI0930 1	RFI1130 1
Erreur de segmentation *initiale*	7,31 %	8,49 %	14,55 %	23,47 %
Erreur de segmentation *présentateur*	7,05 %	7,56 %	1,31 %	4,27 %

A partir du corpus ESTER Train, nous avons extrait environ 20 minutes pour chaque présentateur et aussi plusieurs heures de données ne contenant pas le locuteur cible. A partir de ces données, nous avons entraîné pour chaque locuteur cible (présentateur) un modèle locuteur et un modèle du monde (spécifique à chaque radio). Dans les deux cas, les modèles sont des GMM 128 diagonales et la paramétrisation utilisée est identique à celle du système de segmentation (16 MFCC et énergie). Chaque segment de parole est attribué ou non au locuteur connu en utilisant un seuil sur le rapport de vraisemblance (selon la formule I. 19 de la partie Etat de l'art). Le système de pré-segmentation est réglé sur le corpus ESTER Dev et appliqué tel quel sur le corpus ESTER Eva. Le réglage est fait uniquement au niveau du seuil de décision mais il pourrait être fait aussi au niveau de la paramétrisation et de la taille du modèle par exemple.

Les résultats pour le corpus ESTER Eva sont présentés dans Tableau III. 8. Nous avons rajouté cette fois l'erreur de *pré-segmentation présentateur/non-présentateur* qui est l'erreur du système de suivi du présentateur ; nous avons aussi ajouté l'erreur de segmentation *finale* qui est l'erreur obtenue par le système de segmentation (les différents types d'erreur

sont expliqués dans la Figure III. 6) en utilisant successivement la pré-segmentation présentateur/non-présentateur et la segmentation en locuteurs.

Tableau III. 8 Utilisation de données de référence disponibles pour un locuteur (présentateur), résultats sur le corpus ESTER Eva

Fichier	FrInt 3	FrInt 4	RFI0930 2	RFI1130 2
Erreur de segmentation initiale	**13,29 %**	**6,89 %**	**13,67 %**	**25,29 %**
Erreur de segmentation présentateur	6,83 %	5,48 %	2,24 %	6,14 %
Erreur pré-segmentation présentateur/non-présentateur	3,84 %	3,63 %	5,16 %	1,67 %
Erreur de segmentation finale	**12,35 %**	**5,64 %**	**17,42 %**	**21,26 %**

Nous obtenons une amélioration en termes d'erreur de segmentation sur les fichiers FrInt 1, FrInt 2, RFI1130 2 où une partie importante de l'erreur de segmentation globale était due initialement au présentateur. Pour le seul fichier (RFI0930 2) sur lequel nous avons eu une dégradation au final, l'erreur de segmentation due au présentateur est déjà très faible ce qui veut dire que la majorité de l'erreur de segmentation initiale est due aux parties appartenant à d'autres locuteurs inconnus. Dans ce cas, la pré-segmentation en présentateur/non-présentateur cascadée avec le système de segmentation ne fait que rajouter des erreurs. L'erreur de pré-segmentation se retrouve alors sur l'erreur de segmentation finale (13,67 - 2,24 + 5,16 = 16,59 qui est assez proche de 17,42).

3.3. Utilisation d'un ou plusieurs résultats de segmentation disponibles *a priori*

L'information *a priori* la plus intéressante [Moraru 2004b], mais en même temps, la plus difficile à utiliser est la présence d'un résultat de segmentation issu d'un autre système de segmentation en locuteurs, car il est difficile d'établir une correspondance entre les étiquettes de locuteurs fournies par chacun des systèmes de segmentation.

3.3.1 Deux systèmes cascadés (ELISA hybride*)*

Le cas le plus facile à traiter est celui où nous disposons d'un seul autre résultat de segmentation. Dans ce cas, on peut par exemple utiliser ce résultat pour initialiser un autre système de segmentation, comme cela est présenté dans la Figure III. 7.

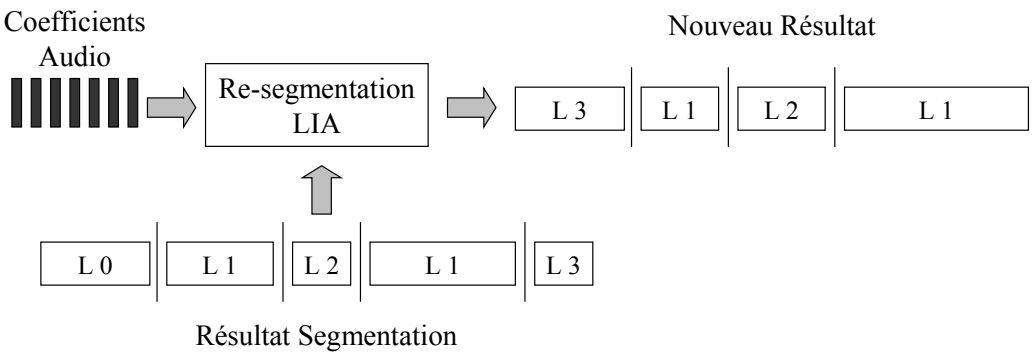


Figure III. 7 Utilisation d'un résultat de segmentation comme information *a priori* pour un autre système de segmentation

L'idée est de faire une re-segmentation à partir d'un résultat de segmentation initial. Dans notre cas, le résultat de segmentation utilisé pour initialiser la re-segmentation LIA est le résultat obtenu par le système CLIPS qui utilise aussi la pré-segmentation acoustique décrite dans la section 5.1 de la partie précédente. Si nous voulons obtenir un changement important par rapport au résultat initial il faut que les deux systèmes utilisent des stratégies différentes pour être capables de corriger les erreurs de chacun d'entre eux. Nous allons appeler par la suite ce système "*ELISA Hybride*".

3.3.2 Fusion d'étiquettes (ELISA fusion*)*

Si nous sommes dans le cas où nous avons à notre disposition plusieurs résultats de segmentation, un premier problème à résoudre est d'établir une correspondance entre les différents locuteurs qui ont été détectés par chacun des systèmes de segmentation. En effet, à l'issue de la segmentation en locuteurs, l'identité des locuteurs n'est pas connue. Les personnes présentes dans un enregistrement sont identifiées uniquement par leur nom symbolique de type L0, L1, …, Ln, etc. En plus, chaque système de segmentation en locuteurs utilise son propre ensemble de noms symboliques, c'est à dire que pour le même document, le même locuteur peut s'appeler différemment selon le système de segmentation utilisé.

Cependant, ce qui reste commun à tous les systèmes de segmentation est le signal audio. Chaque système va étiqueter à sa façon un segment de signal mais le segment reste commun à tous les systèmes. A partir de cette remarque nous proposons ici une méthode de fusion de résultats de segmentations par concaténation d'étiquettes.

La méthode proposée consiste tout simplement à concaténer pour chaque trame de signal les étiquettes données par chaque système. Soit par exemple quatre systèmes de

segmentation en locuteurs, l'étiquette issue de la fusion sera obtenue comme indiqué dans le Tableau III. 9.

Tableau III. 9 Exemple de fusion d'étiquettes pour quatre systèmes de segmentation

Trame	Système 1	Système 2	Système 3	Système 4	Fusion
i	S1	T4	S1	F1	S1T4S1F1
i + 1	S2	T4	S1	F3	S2T4S1F3

Bien sûr, avec cette seule fusion d'étiquettes, le nombre d'étiquettes différentes de locuteurs généré est très grand. Sur les documents issus du corpus RT 2003 Eva on obtient environ 150 étiquettes de "locuteurs virtuels" pour 30 minutes de signal, parmi lesquels on trouve :

- des étiquettes associées à une grande quantité de données, elles correspondent à des locuteurs qui peuvent être considérés comme correctement détectés par tous les systèmes ;
- des étiquettes associées à une petite quantité de données concentrée sur un seul segment court, elles correspondent à des zones qui peuvent être considérées comme erronées, leur influence sur l'erreur finale de segmentation étant par ailleurs marginale ;
- des étiquettes associées à peu de données mais cette fois ci dispersées parmi plusieurs segments courts, ces zones peuvent être considérées comme des zones d'indécision.

La re-segmentation présentée dans la section 5.2 de la partie précédente n'est pas seulement utilisée pour mieux redistribuer les données entre les locuteurs, elle peut aussi éliminer certains locuteurs dont les interventions sont trop courtes. C'est précisément cette qualité de la re-segmentation que nous allons utiliser pour diminuer le très grand nombre initial d'étiquettes de locuteurs issu de la fusion de labels pour arriver à un nombre raisonnable. La re-segmentation est appliquée de façon itérative et après chaque itération, les locuteurs ayant moins de 3 sec de signal attribué sont éliminés. Apres une première itération, le nombre de locuteurs est déjà réduit à environ 50.

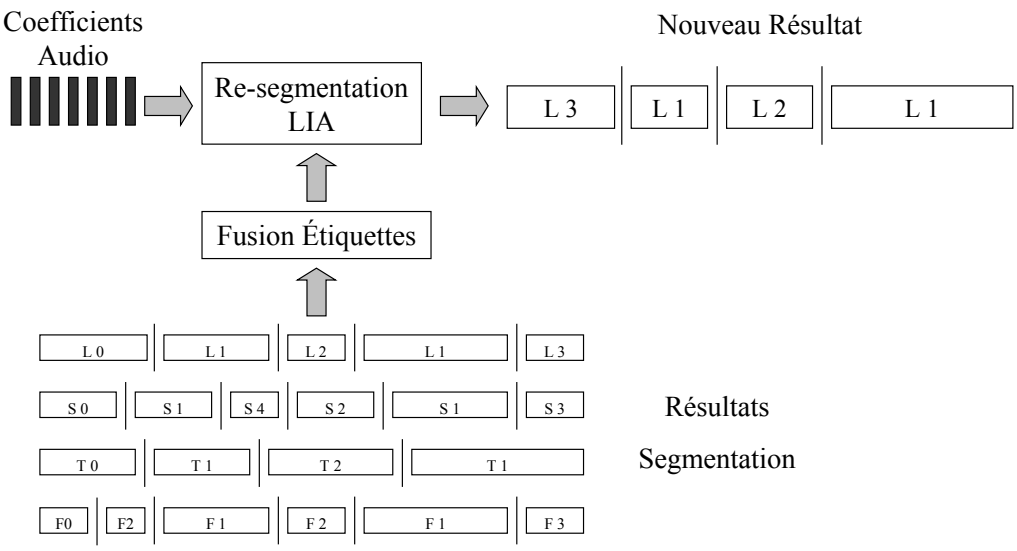


Figure III. 8 Utilisation de plusieurs résultats de segmentation comme information *a priori* pour un autre système de segmentation

La méthode est efficace si les systèmes de segmentation utilisent des stratégies de segmentation différentes. Nous pouvons alors espérer que par fusion, les systèmes de segmentation vont arriver à corriger leurs erreurs entre eux. Un exemple de fusion pour deux systèmes (CLIPS et LIA) est présenté dans la Figure III. 9.

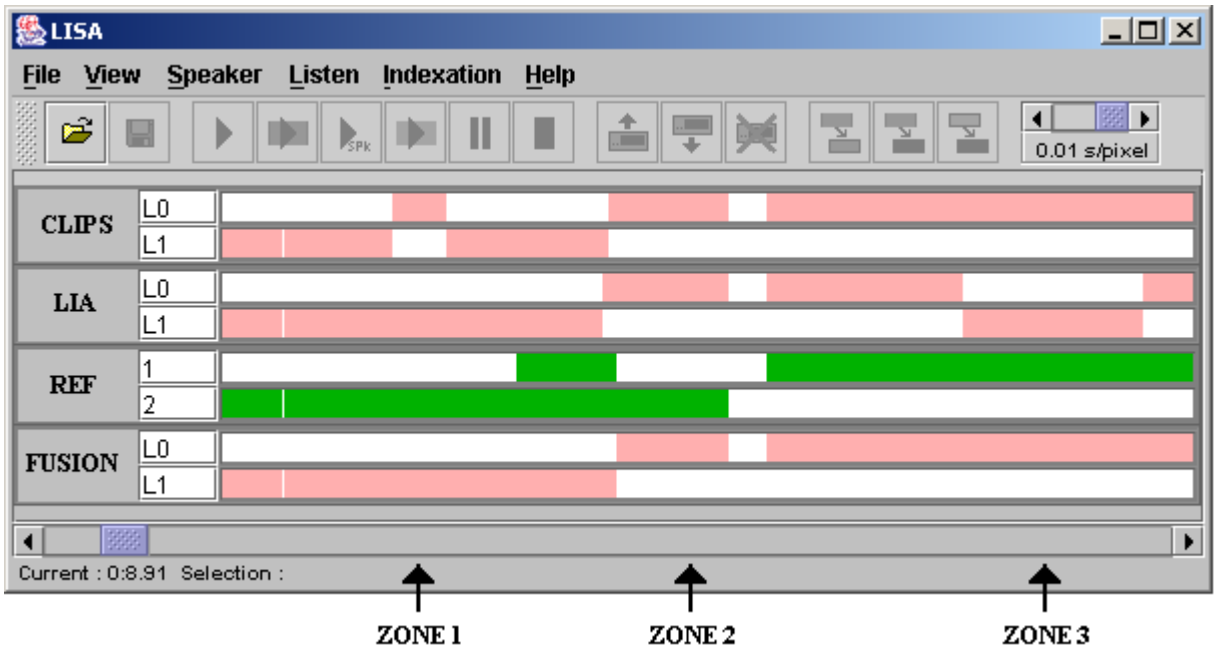


Figure III. 9 Exemple de résultat final de segmentation après l'utilisation de la fusion d'étiquettes

Dans la zone 1, le système LIA donne le résultat correct par rapport à la transcription de référence (REF) ; le résultat final de segmentation (FUSION) est correct. Le système LIA a corrigé l'erreur faite par le système CLIPS. Dans la zone 3 le système CLIPS donne le résultat correct par rapport à la transcription de référence (REF) ; le résultat final de segmentation (FUSION) est correct. Le système CLIPS a corrigé l'erreur faite par le système LIA. Dans la zone 2, malheureusement aucun système n'a pu donner le résultat correct et le résultat final est aussi incorrect.

Tableau III. 10 Un ou plusieurs résultats de segmentation disponibles *a priori* : Expérimentations sur le corpus RT 2003 Eva

Système	DM Parole	FA Parole	Erreur Locuteur	Erreur Segmentation
CLIPS	2,0 %	2,9 %	14,3 %	19,25%
LIA 1	1,1 %	3,8 %	12,0 %	16,90%
LIA 2	1,1 %	3,8 %	19,8 %	24,71%
ELISA Fusion	1,1 %	3,8 %	9,3 %	14,24%
ELISA Hybride	1,1 %	3,8 %	8,0 %	12,88%

Le Tableau III. 10 est une synthèse des résultats présentés lors des évaluations RT 2003. Tous le systèmes utilisent la pré-segmentation acoustique présentée dans la section 5.1 de la partie précédente. Les différences concernant les deux erreurs de pré-segmentation acoustiques (DM et FA Parole) sont un effet de bord du au fait que le système LIA traite le signal par blocs de 0,2 secondes et pas par trame de 10 ms comme le système CLIPS. Les deux systèmes LIA sont fondés sur la stratégie décrite dans la section 5.2 de la partie Etat de l'art, la différence entre les deux étant au niveau de l'adaptation. Pour le système ELISA Fusion nous avons utilisé quatre sorties de segmentation : CLIPS, LIA 1, LIA 2 et ELISA Hybride.

Le meilleur résultat est obtenu par le système ELISA Hybride qui utilise les segments purs en terme de locuteur issus de la stratégie hiérarchique du système CLIPS et la puissance de décodage des HMMs utilisées par la re-segmentation LIA. Les systèmes CLIPS et LIA 1 obtiennent des résultats comparables avec un léger avantage pour le système LIA.

En ce qui concerne l'approche ELISA Fusion, une analyse détaillée montre qu'elle améliore les résultats sur 2 fichiers sur un total de trois dans le corpus. En revanche, sur le troisième document, une dégradation importante est observée. Elle correspond au cas où un

des systèmes a divisé un vrai locuteur en deux, ainsi lors de la fusion d'étiquettes, nous obtenons deux étiquettes différentes qui sont associées à une grande quantité de données et ne pourront plus être regroupées lors de la phase de re-segmentation.

3.4. Utilisation de multiples capteurs pour la segmentation en locuteurs (enregistrements de réunions, évaluations RT 04 Meeting)

Parmi les types d'enregistrements existants, les plus difficiles sont pour l'instant sans doute les enregistrements de réunions. La difficulté ne réside pas dans le nombre de locuteurs présents qui est souvent plus petit que dans le cas de journaux télévisés (entre 4 et 8 sur les données RT04-Meeting par rapport à 20-40 sur les journaux télévisés) mais dans le caractère spontané de la parole. Les locuteurs présents ont des interventions de durée variable, de moins d'une seconde jusqu'à plusieurs minutes. Il arrive aussi très souvent qu'ils se coupent la parole. Si on regarde les données des évaluations RT04-Meeting, pour environ un tiers de la durée de chaque enregistrement, nous avons plusieurs (jusqu'à 5) locuteurs qui parlent en même temps !

Si on regarde les différents systèmes de segmentation proposés dans la partie Etat de l'art, on peut se rendre compte qu'aucun système n'est capable de traiter des enregistrements où plusieurs locuteurs parlent en même temps. Par exemple, pour la détection de changements de locuteur, les deux hypothèses testées sur deux fenêtres consécutives sont : est-ce que les fenêtres appartiennent au même locuteur ou à deux locuteurs différents. De même, pour le regroupement, chaque segment de signal est inclus dans une seule classe, et nous n'avons pas la possibilité de l'inclure dans plusieurs classes, c'est à dire de designer des segments multi-locuteurs.

Par rapport à d'autres types de données, il existe plusieurs façons d'enregistrer une réunion :

- un seul microphone placé au milieu des participants ;
- un microphone pour chaque participant ;
- plusieurs microphones placés dans la salle de réunion (par exemple un dans chaque coin de la salle).

Le premier cas est celui qui donnera *a priori*, les plus mauvais résultat étant donné l'éloignement résultant pour certains locuteurs de la réunion. Ce cas correspond aux évaluations NIST SpRec 2002. Pour le deuxième cas, nous pouvons nous attendre à des meilleurs résultats étant donné que la segmentation semble réduite à seulement une détection

de silence sur chaque microphone. Cependant, l'expérience montre que le microphone de chaque locuteur capte également les interventions des personnes proches de lui, les choses ne sont donc pas aussi simples. Une somme des signaux de tous les microphones individuels était aussi proposée à la segmentation pour les évaluations NIST SpRec 2002 à titre de comparaison. De façon assez surprenante, notre système a montré peu de différences (environ 1 % en absolu) sur ces deux types de données.

Des réunions enregistrées avec plusieurs microphones placés uniformément dans la salle de réunion ont été proposées pour les évaluations NIST RT-04 Meeting. La difficulté principale dans ce genre de situation est la grande variabilité du signal selon le microphone. Selon la position du microphone, on peut entendre ou non certains locuteurs et aussi la qualité du signal peut varier beaucoup selon la proximité du locuteur qui parle. Selon la position du locuteur qui parle, on peut se retrouver dans une situation favorable dans certains cas : par exemple quand deux locuteurs parlent en même temps, ils peuvent être captés par des microphones différents, pour être ensuite séparés [Jin 2004] ce qui n'est pas possible si la réunion est enregistrée avec un seul microphone.

Nous pouvons faire l'hypothèse, si on admet que les locuteurs ne changent pas de place pendant la réunion, que pour chaque microphone, il y aura un ou deux locuteurs qui seront suffisamment proches du micro pour être majoritaires sur le signal associé. Avec cette hypothèse, et sachant que le système de segmentation déjà présenté détecte généralement bien le locuteur majoritaire sur un signal audio, nous pouvons par exemple appliquer le système de segmentation et ensuite sélectionner le locuteur majoritaire sur le signal issu de chaque microphone. Le résultat final sera ensuite la réunion des résultats obtenus individuellement sur chaque microphone.

A partir du résultat de segmentation obtenu sur chaque microphone, l'algorithme suivant appelé "Stratégie de Fusion sur Microphones Individuels" (SFMI) publié dans [Fredouille 2004] peut être proposé :

- Etape 1 : Le temps total de parole le plus long attribué à un seul locuteur, tous microphones confondus, est sélectionné. Toutes les interventions de ce locuteur seront considérées comme lui appartenant dans le résultat final de segmentation ;
- Etape 2 : Les segments sélectionnés à l'étape 1 sont éliminés sur tous les autres microphones ;

- Etape 3 : Si le temps total de parole restant est inférieur à 30 secondes, le processus est arrêté et tout le signal restant est attribué à un nouveau locuteur. Dans le cas contraire une nouvelle itération est lancée en retournant à l'étape 1.

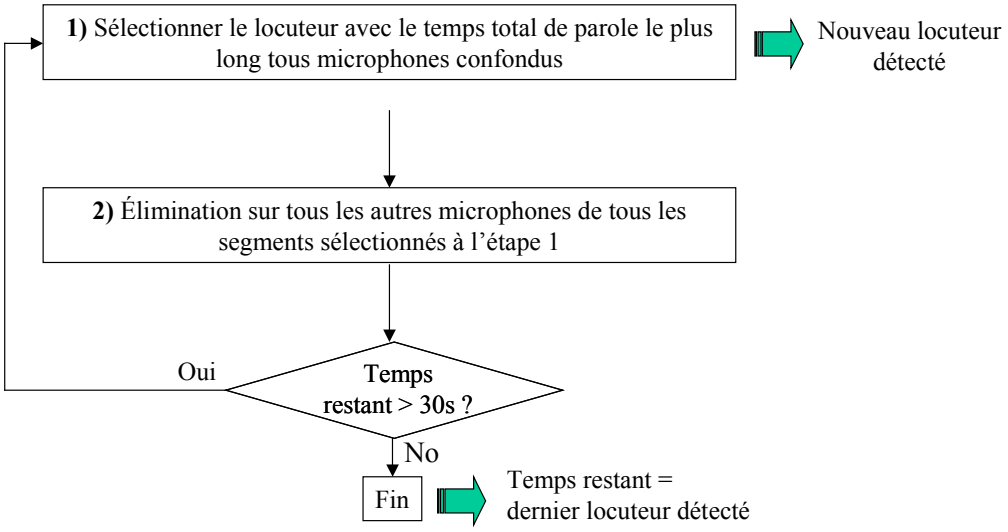


Figure III. 10 L'algorithme de "Stratégie de Fusion sur Microphones Individuels" (SFMI)

Une seule règle supplémentaire à l'algorithme décrit dans la Figure III. 10 est considérée pour l'étape 1 : si le temps de parole sélectionné par rapport au temps total du signal dépasse 60 % du signal total, le locuteur sélectionné est ignoré. Cette règle est utilisée pour éviter les erreurs de segmentation dues à la mauvaise qualité du signal audio sur certains microphones disposés dans la salle de réunion. Pour illustration, un exemple d'enregistrement de réunion avec le résultat de la pré-segmentation acoustique est présenté dans la Figure III. 11. Le signal sur un des capteurs ($3^{\text{ième}}$ signal) est composé uniquement de silence et sur un autre ($4^{\text{ième}}$ signal) uniquement de bruit de fond. Sur le quatrième canal (bruit de fond), le système de segmentation ne détecte qu'une longue intervention d'un locuteur fictif, d'où la nécessité de cette dernière règle pour éliminer l'effet de ce problème sur notre algorithme SFMI.

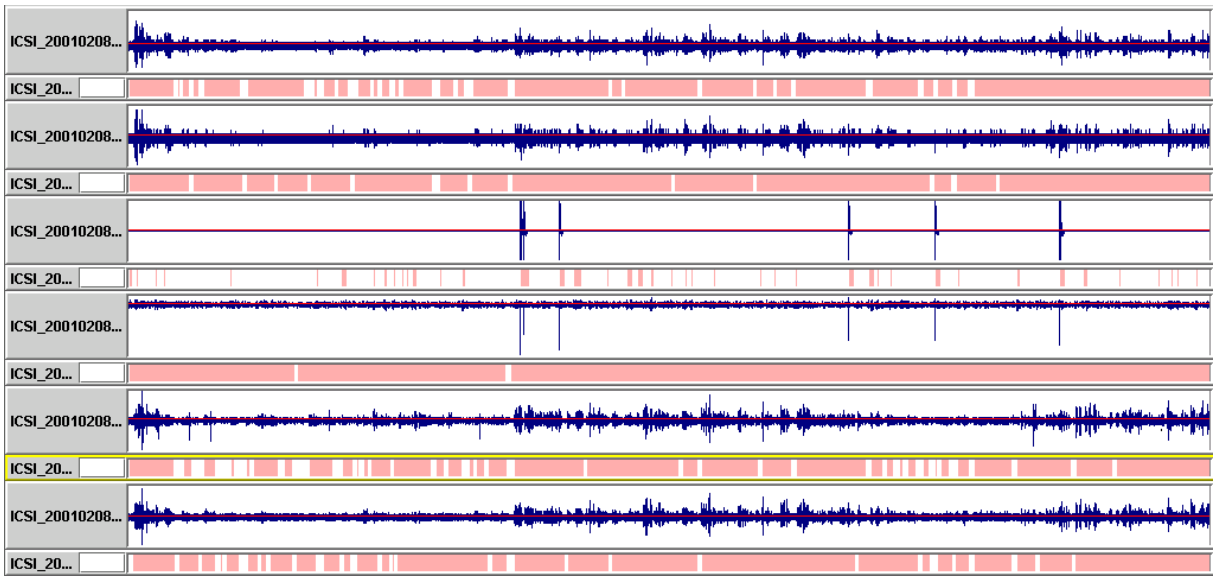


Figure III. 11 Pré-segmentation acoustique parole/non-parole pour un enregistrement d'une réunion avec plusieurs capteurs.

Tableau III. 11 Pourcentage d'erreur de segmentation pour le corpus Meet 2004 (RT 04 Meeting)

Fichier	Locuteurs	Microphones	Erreur segmentation multi-microphone SFMI	Erreur segmentation 1 microphone
CMU_20030109-1530	4	1	39,8 %	39,8 %
CMU_20030109-1600	4	1	17,8 %	17,8 %
ICSI_20000807-1000	6	6	19,1 %	19,9 %
ICSI_20011030-1030	10	6	44,2 %	68,4 %
LDC_20011121-1700	3	7	7,7 %	41,4 %
LDC_20011207-1800	3	8	26,7 %	32,3 %
NIST_20030623-1409	6	7	13,9 %	67,6 %
NIST_20030925-1517	4	6	22,7 %	42,9 %
Total	-	-	22,6 %	39,0 %

Le Tableau III. 11 présente le pourcentage d'erreur de segmentation obtenu avec notre algorithme SFMI (3ème colonne) sur le corpus RT Meeting 2004. Le nombre de microphones est le nombre prévu *a priori* pour l'enregistrement et pas forcément le nombre effectif, car il est possible sur certains d'entre eux d'avoir seulement une faible quantité de signal ou pas de

signal du tout. Nous avons également fait une expérience à titre de comparaison en simulant un enregistrement de chaque réunion avec seulement un microphone (résultats reportés dans la dernière colonne nommée *erreur segmentation 1 microphone*). Le signal, pour cet unique microphone *virtuel,* a été obtenu en prenant pour chaque segment de taille fixe (1 seconde) sur chacun des microphones, la partie de signal de plus forte énergie.

Le meilleur résultat est obtenu pour un nombre de locuteurs réduit et un grand nombre de microphones (LDC_20011121-1700). Le résultat le moins bon est obtenu pour la situation inverse (ICSI_20011030-1030) : 10 locuteurs pour seulement 6 microphones. Concernant les résultats obtenus sur un seul microphone, le taux d'erreur obtenu est beaucoup plus important par rapport au cas où l'on essaie d'exploiter tous les capteurs. Ceci montre encore une fois la nécessité d'utiliser des capteurs multiples pour pouvoir segmenter correctement des signaux où les locuteurs parlent souvent en même temps. Il faut bien préciser ici que les zones où plusieurs locuteurs parlent en même temps étaient bien considérées dans le calcul du taux d'erreur.

Malgré la difficulté de la tache, l'erreur de segmentation moyenne (22,6%) est comparable avec le résultat obtenu sur des corpus de journaux télévisés (RT 2003) réputés plus faciles. On peut remarquer par contre une grande variabilité des résultats individuels sur chaque document. On passe de 7,7 % à 44,2 % entre deux enregistrements différents. Même pour des enregistrements concernant les mêmes locuteurs et les mêmes conditions d'enregistrement (exemple : CMU) nous avons une grande variabilité dans d'erreur de segmentation. Ceci est spécifique aux enregistrements de type réunion qui peuvent présenter des degrés de spontanéité différents (pourcentage de signal multi-locuteur, durée des tours de parole très variable) par rapport aux enregistrements de journaux télévisés qui ont une structure généralement plus stable.

4. Informations *a priori* qui se sont avérées non-utilisables

Les informations présentés dans cette section sont chronologiquement les premières informations *a priori* qui ont attiré notre attention mais malheureusement, elles se sont avérées être difficilement utilisables pour notre système de segmentation. Nous les reportons tous de même à titre informatif dans ce manuscrit.

4.1. Utilisation de la probabilité *a priori* du locuteur majoritaire

La première information que nous avons essayé d'utiliser est la probabilité d'apparition du locuteur majoritaire. Si nous regardons toujours le Tableau III. 1, nous pouvons voir par exemple que la probabilité d'apparition du locuteur majoritaire peut être relativement élevée : il est notamment surprenant de voir que dans les conversations téléphoniques, elle est de 68 % ; ces conversations sont donc relativement asymétriques. On retrouve ce genre de configuration dans des dialogues issus de centres d'appels de type client/agent par exemple, où chaque locuteur joue un rôle particulier dans une conversation.

Une possibilité pour prendre en compte cette information *a priori* est de remplacer le critère simple de *maximum de vraisemblance* utilisé lors d'une phase de re-segmentation, par un critère de *maximum a posteriori* qui prend alors en compte la probabilité *a priori* de chacun des locuteurs, en utilisant la formule de Bayes (équation III. 2).

$$P(M/X) = P(X/M)\frac{P(M)}{P(X)} \qquad \text{(III. 2)}$$

Lors d'une phase de re-segmentation où les modèles de locuteurs ont été appris sur chaque classe, à l'issue de la phase de regroupement, tous les locuteurs sont considérés équiprobables. Dans ce cas, le calcul de $P(M/X)$ est équivalent au calcul de $P(X/M)$. Pour utiliser l'information concernant la probabilité d'apparition du locuteur majoritaire, il faut d'abord sélectionner le locuteur majoritaire à partir du résultat de segmentation et ensuite lui assigner une probabilité différente (obtenue grâce à une connaissance *a priori*) lors de la re-segmentation.

Une condition indispensable est la vérification préalable du fait que la probabilité du locuteur majoritaire obtenue par le système (que nous appellerons *PLM Système*) est différente de la probabilité *a priori* du locuteur majoritaire (que nous appellerons *PLM Référence*), sinon, cela signifie que notre système est déjà capable d'estimer correctement la probabilité d'un locuteur majoritaire, et dans ce cas, l'utilisation d'une telle information *a priori* est inutile.

Tableau III. 12 Comparaison entre la probabilité *a priori* et la probabilité *a posteriori* (obtenue automatiquement par notre système) du locuteur majoritaire

Corpus	PLM Référence	PLM Système
RT 2003 Dev	31 %	30 %
RT 2003 Eva	18 %	16 %
ESTER Dev	16 %	15 %
ESTER Eva	18 %	15 %

A l'exception du corpus RT 2003 Dev, la probabilité *a priori* du locuteur majoritaire est quasiment constante (en moyenne 17 %). Le corpus RT 2003 Dev contient des parties de début de journaux et pas des journaux entiers. C'est sans doute pour cette raison que la probabilité *a priori* du locuteur majoritaire est différente. Le locuteur majoritaire est généralement le présentateur du journal qui parle beaucoup plus au début de journal pour présenter les titres des informations que dans la suite qui contient aussi des reportages, des rubriques financières, de la météo, et du sport où le présentateur n'intervient pas.

Concernant cette information, nous pouvons voir que le système que nous avons proposé obtient déjà une probabilité du locuteur majoritaire proche de la probabilité *a priori*. Donc, il semble difficile d'utiliser cette information pour améliorer le résultat de segmentation. Des expériences conduites sur des conversations asymétriques, créées artificiellement, ont été conduites au CLIPS [Moraru 2003a], mais les améliorations obtenues sont assez décevantes et nous ne les décrivons pas en détail dans ce manuscrit.

4.2. Utilisation de la durée moyenne d'un tour de parole

Si nous regardons encore une fois le Tableau III. 1, nous pouvons voir par exemple que la durée moyenne d'un tour de parole dans le cas de journaux télévisés est largement supérieure à celle dans le cas de documents téléphoniques.

Pour exploiter cette information sur la durée moyenne des tours de parole, deux possibilités sont envisageables. La première consiste comme dans la cas précédent à modifier la formule de Bayes en fonction de la durée moyenne du tour de parole désirée. On doit calculer la probabilité d'avoir le locuteur M étant donné la trame courante X et le fait que les n trames précédentes appartiennent au locuteur N. Si nous faisons l'hypothèse que la trame courante ne dépend pas du résultat sur les n trames précédentes, alors :

$$P(M/X,nN) = P(X/M,nN)\frac{P(M/nN)P(nN)}{P(X/nN)P(nN)} = P(X/M)\frac{P(M/nN)}{P(X)} \quad \text{(III. 3)}$$

La probabilité *a priori* $P(M/nN)$ peut être estimée par exemple à partir d'un histogramme de taille de segments construit sur des données de développement.

La deuxième possibilité est d'utiliser un HMM ergodique avec un état par locuteur pour re-segmenter la sortie du système de segmentation. Dans ce cas, la probabilité de rester dans le même état peut être ajustée en fonction de la taille moyenne des tours de parole estimée sur le type de données à traiter.

Comme dans le cas de la prise en compte de la probabilité *a priori* du locuteur majoritaire, il faut vérifier au préalable si le système automatique n'estime pas déjà bien la longueur moyenne des tours de parole (TPM). Si c'est le cas, toute prise en compte de la durée moyenne du tour de parole sera inutile.

Tableau III. 13 Comparaison entre la durée moyenne *a priori* et *a posteriori* (obtenue par le système) d'un tour de parole

Corpus	TPM Référence	TPM Système
RT 2003 Dev	5,01 sec	4,91 sec
RT 2003 Eva	5,70 sec	4,39 sec
ESTER Dev	3,95 sec	4,27 sec
ESTER Eva	3,68 sec	4,39 sec
Phone 2002	1,69 sec	1,76 sec

Comme dans le cas du paragraphe précédent, nous pouvons voir que le système que nous avons proposé obtient déjà un tour de parole moyen proche de sa valeur réelle même dans le cas de la parole téléphonique spontanée (Phone 2002) supposée plus difficile à traiter. En conséquence, cette information ne semble pas être très utile pour améliorer un résultat de segmentation.

5. Conclusion

Dans cette partie, nous nous sommes intéressés notamment à l'utilisation d'informations *a priori* pour la segmentation en locuteurs.

On rappelle ici que par hypothèse, spécifique au cadre des évaluations NIST et ESTER, un système de segmentation en locuteurs ne dispose généralement pas d'informations *a priori* concernant les documents à analyser. Dans certaines applications pratiques et sur certaines données, il est cependant possible de disposer d'informations utiles *avant* de traiter l'enregistrement. Une résumé des différents résultats obtenus, dans cette partie, se trouve dans le tableau suivant :

Tableau III. 14 Synthèse des résultats concernant les différentes informations *a priori* utilisées et selon différents types de données (journaux télévisés et réunions). Amélioration (valeur négative) ou dégradation (valeur positive) en terme de pourcentage d'erreur de segmentation

Information	Nombre de locuteurs	Données pour tous les locuteurs	Un modèle pour un seul locuteur	Pré-segmentation Acoustique	Segmentation d'un autre système	Segmentation sur une autre voie
Journaux	**+ 5 %**	-8 %	- 4 %	- 8 %	- 6 %	-
Réunions	-	-	-	-	-	- 18 %

Nous avons appelé *information certaine*, toute information qui peut être obtenue par une annotation manuelle comme par exemple le nombre de locuteurs ou une courte annotation de référence incomplète. Parmi les informations présentées dans Tableau III. 14 seulement la connaissance *a priori* du nombre de locuteurs par rapport à l'estimation en aveugle ne semble pas apporter une amélioration à notre système, notamment à cause du fait que la méthode d'estimation du nombre de locuteurs que nous avons proposée est déjà assez précise. Par contre, l'utilisation de données de référence disponibles pour chaque locuteur peut conduire à une réduction importante de l'erreur de segmentation mais aussi du temps d'exécution. Le temps d'exécution passe de plusieurs fois le temps réel à environ 0,1 fois le temps réel.

Les informations dites *incertaines* sont des informations qui peuvent être issues d'un autre système automatique de traitement de la parole. Concernant les journaux télévisés, toutes les informations *incertaines* utilisées *a priori* (pré-segmentation acoustique, pré-segmentation en présentateur/non présentateur, pré-segmentation issue d'un autre système) nous ont mené généralement vers une réduction de l'erreur de segmentation. Cependant, certaines doivent être utilisées avec précaution. Si nous avons par exemple plusieurs résultats de segmentation disponibles, une erreur importante présente dans un des systèmes peut nous mener à une augmentation de l'erreur finale de segmentation.

Concernant le dernier aspect traité, la présence de plusieurs capteurs, ceci est une situation spécifique aux enregistrements de réunions. La difficulté élevée de ce type de documents est due à la spontanéité de la conversation (les locuteurs parlent en même temps pour une durée importante par rapport à la taille du document, grande variabilité des durées de tour de parole à travers le document). Nous avons proposé une stratégie originale qui nous permet d'extraire la parole appartenant à chaque locuteur à partir de résultats de segmentation obtenus individuellement sur chaque microphone. Cette approche a donné des résultats encourageants malgré la difficulté de la tâche. Le taux d'erreur de segmentation y est comparable avec celui obtenu sur des documents de type journaux télévisés.

Quatrième partie :

Application à la recherche d'information multimédia

Cette partie concerne l'application des outils de segmentation en locuteurs développés dans cette thèse, à la recherche d'information multimédia. Deux directions principales de recherche sont explorées : l'utilisation de l'information *locuteur* pour l'indexation de documents vidéo et, inversement, l'utilisation de l'information vidéo pour la segmentation en locuteurs qui est une tâche purement audio. Ces travaux de recherche sont faits dans le contexte des évaluations TREC VIDEO 2002 et 2003 sur des grandes collections de données audio-visuelles (plus de 100 heures de signal dans le cas de TREC 2003).

1. Introduction

Dans cette partie nous allons aborder le problème de l'indexation d'une grande collection de documents vidéo. Il s'agit en fait ici de répondre à deux questions principales : est-ce que la segmentation en locuteurs peut apporter des informations utiles pour l'indexation de documents vidéo ? Et, inversement, est-ce que les informations extraites à partir du canal vidéo peuvent être utiles pour améliorer le résultat d'un système de segmentation en locuteurs ?

La possibilité d'avoir de l'information au format numérique a sans doute beaucoup d'avantages : possibilité de transfert rapide, reproduction illimitée, facilité de stockage. Comme pratiquement toute l'information disponible aujourd'hui est sous format numérique, la recherche de documents vidéo basée sur le contenu est devenue une tâche importante.

La difficulté d'indexation de documents audio et vidéo [Berrut 1997] par rapport aux documents textuels réside dans la distance entre le support de représentation (le signal numérique) et les informations contenues dans le document. Les documents textuels sont représentés par une chaîne de caractères. Les informations sémantiques sont donc plus facilement extraites à partir des mots et des phrases. Les documents audio et vidéo sont représentés quant à eux par une chaîne d'octets issus de la numérisation de signaux électriques, leur contenu sémantique est dans ce cas plus difficile à extraire. Malgré les possibilités offertes par le format numérique pour y ajouter des informations sémantiques, la représentation de documents audio et vidéo a été développée, dans un premier temps, sans réellement envisager la possibilité de manipuler l'information selon le contenu.

Smeaton passe en revue dans [Smeaton 2001] les principales techniques d'extraction d'informations à partir de documents audio et vidéo. Pour les documents audio, les techniques décrites sont : le repérage de mots clés (word spotting), la reconnaissance du locuteur et l'extraction de mots clés à partir de la reconnaissance automatique de la parole continue. Pour les documents vidéo, l'extraction d'information la plus simple est la détection de plans vidéo [Cooper 2003]. Chaque plan peut être ensuite décrit par une image ou par une suite d'images.

Dans la classification de Smeaton, on remarque que le signal audio et le signal vidéo sont traités de façon séparée. C'est le cas de la plupart des systèmes de traitement d'information multimédia [Beal 2002]. Pour indexer des documents vidéo, on utilise par exemple des systèmes de transcription automatique de journaux télévisés ou des systèmes de

suivi vidéo mais appliqués de façon indépendante, avec faible interaction entre le traitement du signal vidéo et le traitement du signal audio. Une possibilité d'utilisation conjointe d'informations audio et vidéo est présentée dans [Benois-Pineau 2002]. Les auteurs retrouvent à partir d'un "leitmotiv audio-visuel" des plans vidéo associés à un certain instrument musical. La détection de frontières de plans vidéo est utilisé dans ce cas conjointement avec le traitement de l'audio. L'instrument musical considéré comme étant harmonique et percurssif est retrouvé à partir de corrélation avec des enveloppées spectrales appartenant aux instruments connus. Cependant, comme les deux signaux peuvent provenir d'une même source, il peut exister des informations représentées dans les deux flux de données. Un système optimal de traitement de l'information devra en conséquence exploiter les informations redondantes entre les deux signaux.

Les travaux présentés dans cette partie s'inscrivent dans ce contexte : nous nous proposons en effet d'utiliser les informations audio et vidéo pour améliorer les performances de chaque source d'information, le but final étant l'extraction d'informations sémantiques à partir de documents multimédia.

Les applications les plus intéressantes concernent les collections de journaux télévisés notamment en raison de leur intérêt pour le grand public (retrouver un reportage à la demande, retrouver les déclarations d'une certaine personnalité politique, etc.) mais aussi en raison de leur problématique spécifique (traitement de flux audio et vidéo continus et hétérogènes, grand nombre de locuteurs, conditions audio très variables, présence de plusieurs sources d'informations différentes : audio, vidéo, sous-titrage, etc.). Nous sommes conscients que les expériences reportées dans ce cadre ne seront pas immédiatement applicables sur d'autre types de données (problème de généricité), mais nous avons également fait le choix de ces données parce qu'elle sont largement utilisées dans la communauté *recherche d'information multimédia*, ce qui nous permettra de nous comparer à d'autres travaux, le cas échéant.

2. Recherche d'information multimédia : généralités et métriques d'évaluation

Une des spécificités des travaux sur les données multimédia est la nécessité de traiter le flux audio qui est continu et hétérogène. Les signaux de tests sont des enregistrements de longue durée et peuvent contenir plusieurs dizaines de locuteurs, différentes classes

acoustiques (parole, silence, musique, etc.) et différentes qualités de parole (parole propre, parole bruitée, etc.). Il faut donc sélectionner les extraits ou segments d'intérêt sur les documents analysés. Par exemple, pour la segmentation en locuteurs, il faut d'abord sélectionner les zones contenant de la parole. Par ailleurs, les documents vidéo et audio se visualisent séquentiellement et leur durée peut être longue. Les demandes des utilisateurs portent alors sur des parties seulement, et non sur tout le document : il faut dans ce cas identifier un découpage temporel du document.

Les travaux de [Hjelsvold 1994] montrent un système de modélisation et d'interrogation de données vidéo et modélisent de façon très détaillée les informations vidéo. Afin de renseigner au mieux les données, la représentation des vidéos dans ce modèle est alors manuelle. Le système permet à l'utilisateur d'interroger les données vidéo selon cinq types de requêtes : la navigation structurelle (par une table des matières), la navigation par clip (les vidéos sont résumées par quelques images à partir desquelles l'utilisateur peut les visionner), la navigation par résumé (les utilisateurs peuvent lire le résumé des vidéos et ainsi faire leur choix), les requêtes de contenu simple (par les descriptifs faits dans les vidéos ou leurs sous-séquences), et les requêtes thématiques (par des annotations générales sur chaque vidéo). Le système a été développé pour un corpus de journaux télévisés, et est donc adapté notamment aux besoins spécifiques de ce type de document.

Les performances d'un système de recherche d'information sont évaluées par les mesures de *rappel* et *précision*. La segmentation en locuteurs est en revanche, évaluée par l'erreur de segmentation comme nous l'avons vu dans la deuxième partie de la thèse. La question que l'on peut se poser est alors : que vaut un système de segmentation en locuteur évalué en termes de taux d'erreur dans une tâche plus générale de recherche d'information évaluée en termes de rappel et précision ? Dans la recherche d'information, sur des documents vidéo, l'unité utilisée est généralement le plan vidéo (durée variable de quelques secondes à plusieurs dizaines de secondes). Dans la segmentation en locuteurs, l'unité utilisée est plutôt la trame audio (de durée fixe, typiquement 20 ms).

Soit $Pert$ le nombre de documents pertinents dans une collection, qui contiennent un trait recherché. Soit Ret le nombre de documents que le système a trouvé, contenant le trait recherché. Le rappel R et la précision P sont définis par les formules IV. 1. Si nous voulons avoir une mesure de performance unique nous pouvons utiliser le F Ratio donné par la formule IV. 2. Bien sûr, comme le choix d'un document pertinent et le plus souvent soumis au choix d'un seuil de décision, on peut aussi tracer l'évolution du rappel en fonction de la

précision, pour différentes valeurs de seuil, de façon similaire à la construction d'une courbe ROC en théorie de la décision binaire, par exemple.

$$R = \frac{Ret \ \wedge \ Pert}{Pert} \qquad P = \frac{Ret \ \wedge \ Pert}{Ret} \tag{IV. 1}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{IV. 2}$$

Si en plus, le système de recherche d'information donne en sortie une liste ordonnée de documents selon un score de confiance préalablement défini, nous pouvons alors définir la précision moyenne. Cette mesure est la moyenne des précisions calculées après chaque document pertinent retrouvé par le système :

$$\overline{P} = \frac{\sum_{i=1}^{Pert} P_i}{Pert} \tag{IV. 3}$$

Dans les mêmes conditions, nous pouvons évaluer le système en tenant compte seulement des N premiers documents retrouvés ($Ret = N$). Nous pouvons alors calculer la précision, le rappel et le F Ratio uniquement sur les N documents considérés. Par exemple, si N est égal à 1000, nous pouvons calculer R 1000, P 1000 et F 1000.

La relation entre la mesure de performance d'un système de segmentation en locuteurs (proportionnelle au temps d'un document) et les mesures utilisées dans la recherche d'information n'est pas évidente. Il est impossible d'établir une relation mathématique précise entre le rappel et la précision et l'erreur de segmentation. Selon la tache visée, un mauvais système de segmentation en terme d'erreur de segmentation pourra donner des résultats acceptables en termes de rappel et précision sur certaines tâches, comme la tâche de détection de monologues qui sera présentée plus tard dans cette partie.

3. Contexte de travail : la campagne d'évaluation TRECVID

Les travaux présentés dans cette partie s'inscrivent dans le contexte des évaluations TRECVID (Text REtrieval Conference VIDeo) 2002, 2003 et 2004 organisés par NIST. Nous

allons commencer par une courte présentation de la campagne, suivie des expérimentations concernant les différentes tâches proposées auxquelles nous nous sommes intéressé. Nous nous sommes inspiré de cette campagne pour proposer aussi nos propres tâches. Ces dernières ont été proposées pour montrer plus spécifiquement l'apport de la segmentation en locuteurs sur l'indexation de documents vidéo et réciproquement, pour montrer l'apport du canal vidéo pour la segmentation en locuteurs.

Le Tableau IV. 1 présente une synthèse des différentes tâches auxquelles nous nous sommes intéressés. Le détail de toutes ces tâches sera décrit plus précisément dans la suite de cette partie.

Tableau IV. 1 Les tâches proposées dans la campagne TREC ou inspirées de la même campagne sur lesquelles nous avons travaillé

Trait à détecter ou tâche	Traitement audio	Traitement vidéo	Contexte d'évaluation	Données
Parole	Détection activité vocale	Détection plans vidéo	TREC officiel	Vieux films
Monologue	Segmentation en locuteurs	Détection plans vidéo ; Détection visage	TREC officiel	Vieux films
Personne X	Identification du locuteur	Détection plans vidéo	TREC officiel	Journaux télévisés
Segmentation en locuteurs audio/vidéo	Segmentation en locuteurs	Détection plans vidéo	Inspirée de TREC	Journaux télévisés
Présentateur	Segmentation en locuteurs	Détection plans vidéo	Inspirée de TREC	Journaux télévisés
Segmentation en histoires	Segmentation en locuteurs ; Détection silences ; Détection de jingles ; Reconnaissance de parole	Détection plans vidéo	TREC officiel	Journaux télévisés

3.1. Introduction

La campagne d'évaluation TREC est organisée chaque année depuis 1992 par NIST (National Institute of Standards and Technology). Le but de la campagne est d'encourager les travaux scientifiques en recherche d'information en fournissant de grandes collections de données, des procédures d'évaluation uniformes, et un cadre expérimental pour des laboratoires intéressés à comparer leurs résultats entre eux.

A partir de l'année 2001, les données vidéo ont commencé à faire partie des données d'évaluation. Les données sur lesquelles les évaluations ont été faites pendant les années 2001 et 2002 étaient des enregistrements de vieux films. Trois tâches principales ont été définies alors :

- détection de plans vidéo ;
- extraction de traits à partir de la vidéo ;
- recherche interactive.

La première tâche consiste évidemment à détecter les frontières de plans vidéo. Elle a été nécessaire pour réaliser les deux autres tâches. Un plan vidéo est défini comme une séquence continue de trames vidéo similaires d'un point de vue visuel. Les évaluations TREC utilisent le terme "*continuous camera shots*", un plan (shot) étant le résultat des trois actions suivantes : démarrer la caméra vidéo, filmer une séquence vidéo, arrêter la caméra vidéo. Les moments du démarrage et l'arrêt de la caméra vidéo constituent les frontières d'un plan vidéo. Dans le cas de transmissions télévisées un changement de caméra constitue aussi une frontière.

La *détection de frontières de plans vidéo* est une première structuration simple d'un document vidéo ; même si quelques problèmes demeurent lors de fondus enchaînés par exemple, la détection de frontières de plans vidéo est un problème quasiment résolu.

L'extraction de traits est toujours évaluée en termes de rappel et précision calculés à partir du nombre de plans vidéo pertinents. Le plan vidéo est donc l'unité élémentaire pour notre travail sur les documents vidéo. Le score officiel de la campagne TRECVID est la précision moyenne, définie par la relation IV. 3.

Nous avons utilisé pour notre partie expérimentale, le système de détection de frontières de plans vidéo du laboratoire [Quénot 2001]. Le système est capable de détecter deux types de frontières : *abruptes* (la frontière est située entre deux images complètement différentes) et *progressives* (une transition linéaire sur plusieurs images consécutives).

Pour la deuxième catégorie de tâches (*extraction de traits*), les participants doivent rechercher des plans vidéo contenant un trait particulier. Selon le support d'information utilisé pour extraire le trait recherché, nous avons trois catégories distinctes :

- traits vidéo : paysage extérieur, paysage intérieur, visage, texte, etc. ;
- traits audio : parole, musique instrumentale ;
- traits audio-vidéo : monologue (présence d'un visage et de la voix d'une seule personne simultanément).

La *recherche interactive* est une tâche spécifique au domaine de la recherche d'informations. Elle est considérée comme la tâche la plus importante pour un utilisateur de moteur de recherche d'informations. Soit la requête suivante : "*Je voudrais tous les plans vidéo qui contiennent des images de la ville de New York*.". Un opérateur humain transforme ce besoin en une requête exprimée dans le langage du moteur de recherche. Le moteur de recherche sélectionne ensuite des documents pertinents dans une grande collection de données selon des thèmes définis auparavant. Il existe deux possibilités de recherche : manuelle et interactive. La recherche interactive permet à l'opérateur humain de reformuler sa requête en fonction des résultats de la requête initiale si ceux-ci ne le satisfont pas.

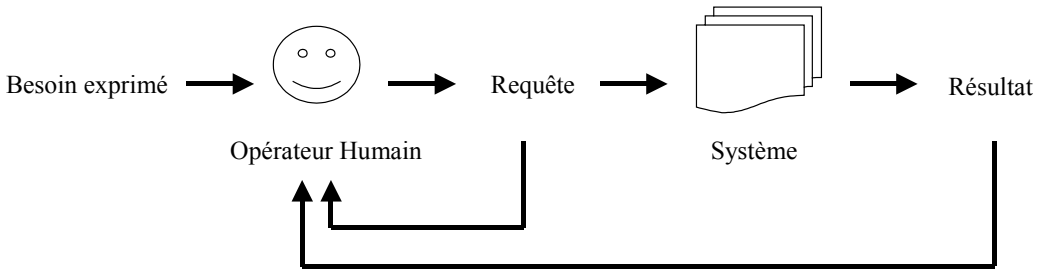


Figure IV. 1 La recherche d'information interactive

A partir de l'année 2003, les données de vieux films ont été remplacées par des données de type journaux télévisés. Une nouvelle tâche est apparue aussi : la segmentation en histoires (*story segmentation*) d'un document audio-vidéo.

Nous allons décrire d'abord le corpus expérimental et ensuite les tâches qui ont suscité notre intérêt parmi les tâches proposées dans la campagne d'évaluation. Les tâches qui ont attiré notre attention sont l'extraction de segments contenant de la parole et l'extraction de segments contenant un monologue. La tâche monologue est particulièrement intéressante pour nous, notamment parce qu'elle nécessite la fusion d'informations audio et vidéo, et permet

également d'utiliser la sortie de notre système de segmentation en locuteurs. Pour chaque tâche nous allons donner la définition, et la méthode de résolution employée.

3.2. Corpus vidéo expérimental

Les expérimentations ont été faites sur les bases de données proposées dans les évaluations TREC Vidéo 2002 et TRECVID 2003. Pour des raisons de temps d'exécution certaines expérimentations, comme celles concernant la segmentation en locuteurs, ont été réalisées sur un sous-ensemble du corpus TRECVID 2003, appelé TREC 4F 2003. Le Tableau IV. 2 présente de façon synthétique les principales informations concernant le corpus expérimental.

Tableau IV. 2 Corpus expérimental TREC Vidéo

Corpus	Fichiers	Taille fichiers (minutes)	Plans vidéo	Type	Trait à détecter ou tâche
TREC 2002	23	env. 13	1850	Vieux films	Parole ; Monologue
TREC 2003	121	30	35220	Journaux télévisés	Personne X ; Segmentation en histoires
TREC 4F 2003	4	30	685	Journaux télévisés	Segmentation en locuteurs audio/vidéo ; Détection présentateur

Le corpus TRECVID 2002 est composé exclusivement de vieux films issus du corpus Open Video (http://www.open-video.org/) et Internet Archive (http://www.archive.org/). La qualité des données audio est 16 bits, 16 kHz et la qualité des données vidéo est MPEG-1, résolution 352x240 à une fréquence vidéo de 29,97 Hz. La source vidéo est noir et blanc et parfois les films peuvent être muets. La qualité audio est aussi assez variable, parfois la bande son peut s'étendre seulement jusqu'à 4 kHz. Nous avons utilisé le corpus TRECVID 2002 pour les tâches d'extraction de plans vidéo contenant de la parole et du monologue.

Le corpus TRECVID 2003 est complètement différent du corpus de l'évaluation précédente. Cette fois, le corpus est beaucoup plus important par rapport à l'année précédente, et il est composé majoritairement d'enregistrements de journaux télévisés en anglais CNN, ABC. La même quantité de données est également disponible pour le développement. La qualité des données audio est toujours 16 bits, 16 kHz et la qualité des données vidéo est

MPEG-1, résolution 352x240 à une fréquence vidéo de 29,97 Hz. Mais cette fois l'image est couleur et la qualité du son est constante, la bande passante du signal sonore est entièrement utilisée. Nous avons utilisé le corpus TRECVID 2003 pour la tâche d'extraction de plans personne X et aussi pour la segmentation en histoires.

Pour toutes les données vidéo du corpus TRECVID 2003, il est important de préciser qu'il n'existait aucune annotation de référence détaillée. Seules une segmentation automatique en plans vidéo et une transcription automatique étaient fournies. Pour évaluer les résultats concernant la segmentation en locuteurs sur ces données, nous avons donc dû nous même obtenir une segmentation de référence pour chaque fichier.

Nous avons donc choisi 4 fichiers du corpus TRECVID 2003, deux fichiers CNN et deux fichiers ABC : 19980202_CNN, 19980202_ABC, 19980204_CNN, 19980204_ABC pour lesquels nous avons fait une segmentation manuelle de référence en locuteurs. Chaque fichier contient entre 18 et 27 locuteurs. Le format de la transcription de référence est similaire à celui utilisé pour les évaluations NIST Rich Transcription 2003. La quantité de données de test est aussi comparable à la quantité de données des évaluations NIST Rich Transcription 2003, 4 fichiers de 30 minutes *versus* 3 fichiers de 30 minutes. De même, les séquences de publicité sont exclues de la segmentation. Nous allons appeler ce corpus pour la suite TREC 4F 2003.

3.3. Campagne d'évaluation TREC : description des tâches d'extraction

La tâche d'extraction de traits à partir de la vidéo (*video feature extraction)* consiste à former une liste ordonnée de plans vidéo, du plus probable au moins probable de contenir un trait recherché. L'existence d'une pré-segmentation préalable en plans vidéo est une condition indispensable. Le résultat de la segmentation en plans vidéo du CLIPS[1] est distribué à tous les participants comme pré-segmentation de référence.

L'extraction de traits est toujours évaluée en termes de rappel et précision calculés à partir du nombre de plans vidéo pertinents. Le plan vidéo est donc l'unité élémentaire pour notre travail sur les documents vidéo. Le score officiel de la campagne TRECVID est la précision moyenne, définie par la relation IV. 3.

La collection de données pour les évaluations est divisée en une partie test et une partie développement. Les données de développement sont fournies sans aucune ou avec peu

[1] http://www.itl.nist.gov/iaui/894.02/projects/t2002v/CSBR_mpeg7/README

de vérité terrain et sont utilisées pour le réglage des paramètres du système. Les données de test sont destinées uniquement à l'évaluation.

Tous les traits sont considérés comme étant binaires, c'est-à-dire qu'ils sont soit présents soit absents dans un plan vidéo. Les traits auxquels nous nous sommes intéressé sont détaillés dans les sous-parties suivantes.

3.3.1 Extraction de plans contenant de la parole : critère audio pur pour la recherche d'un document vidéo

La définition de ce trait est la suivante : "*A human voice uttering words is recognizable as such in this segment*", elle correspond à la présence dans un plan vidéo d'une voix humaine prononçant des mots et reconnaissable comme telle.

Pour traiter cette tâche, nous avons, à partir d'environ quatre heures de données de développement, entraîné deux modèles de mélange de gaussiennes de taille 128 : un modèle parole et un modèle non-parole. Le modèle non-parole est un modèle multi-conditions entraîné sur plusieurs classes non-parole : silence, musique, bruit. La paramétrisation est identique à la paramétrisation du système de segmentation. Aucune paramétrisation spécifique à la tâche n'a été faite au moment de l'évaluation. Pour chaque plan vidéo X le rapport de vraisemblance sous forme logarithmique est calculé :

$$llr(X) = \log P(X / M_{Parole}) - \log P(X / M_{NonParole}) \qquad \text{(IV. 4)}$$

Les plans sont ensuite ordonnés selon cette mesure pour produire une liste ordonnée des plans vidéo les plus probables.

3.3.2 Extraction de plans Monologue : critère conjoint audio-vidéo pour la recherche d'un document vidéo

La définition d'un plan monologue est la suivante : "*A single person is at least partially visible and speaks for a long time without interruption by another speaker. Pauses are ok if short*.", une seule personne est au moins partiellement visible et parle pendant un temps long sans être interrompue par une autre personne ; les courtes pauses sont admises.

C'est une tâche qui nécessite la fusion des sources audio et vidéo d'un document vidéo. La condition de visibilité d'une personne implique l'utilisation d'un détecteur de visage et la condition *parole* implique l'utilisation d'un système de segmentation en locuteurs.

Notre recherche de plans *monologue* est illustrée dans la Figure IV. 2. Le système de détection de visage a trouvé plusieurs plans vidéos où un visage est présent. Le premier plan contient un visage mais aussi la parole de plusieurs locuteurs (L0 et L1) donc, il n'est pas considéré comme monologue. Le troisième et le huitième plans contiennent un visage et également la parole d'un seul locuteur. Ces deux plans sont alors considérés comme des plans *monologue* corrects (en gris sur le résultat).

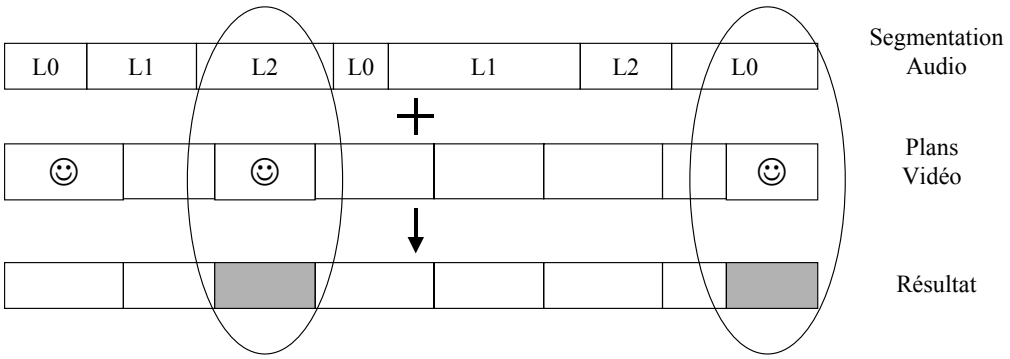


Figure IV. 2 La détection de plans vidéo contenant du monologue

Pour ces expériences, nous avons collaboré pour la détection de visage avec le Laboratoire de Technologie de l'Information (LTI) de Singapour. La détection de visage est faite à partir d'un système de l'Université Carnegie Mellon (CMU) [Rowley 1997], basé sur l'utilisation d'un réseau de neurones. Le système est capable de détecter plusieurs visages sur une image. Les performances de ce système ne sont pas affectées par l'orientation du visage face à la caméra, ni par la distance par rapport à la caméra.

Le système de détection de visages est appliqué uniquement sur les images clés de chaque plan vidéo. Les images dites clés sont sélectionnées selon trois critères : les images doivent se trouver vers le milieu du plan vidéo, elles doivent être fortement contrastées et avoir peu de variation par rapport aux images qui les précèdent et qui les suivent.

La détection de visage est une condition imposée par la définition d'un plan monologue lors des évaluations TREC. Les monologues dits narratifs, caractérisés par la continuité de la parole du même locuteur et l'absence de son visage dans des plans voisins à ceux où ce visage à été détecté, ne sont pas inclus dans la définition d'un plan monologue utilisée lors de ces évaluations. La détection de ce type de monologue pourra faire l'objet

d'une suite de ces travaux. Un tel monologue pourrait être détecté par la stratégie suivante : la détection de visages nous indique la présence d'un visage sur un plan vidéo et la segmentation en locuteurs nous indique ensuite la présence du même locuteur sur ce plan vidéo et sur un ou plusieurs plans vidéo qui le suivent.

Le système de segmentation audio sélectionne d'abord les parties parole obtenues selon la méthode présentée dans 3.3.1. Ensuite, le système de segmentation en locuteurs présenté dans la deuxième partie de la thèse est utilisé uniquement sur les zones de parole. Au moment de cette évaluation (2002), l'estimation du nombre de locuteurs n'étant pas encore implémentée, le système utilisé fixe donc le nombre de locuteurs à deux lors du critère d'arrêt de la phase de regroupement.

Les plans vidéo sont ensuite ordonnés selon une combinaison entre la longueur du segment de parole appartenant à un seul locuteur et le résultat de la détection de visage. Le résultat de la détection de visage est dans ce cas un rapport entre la surface du visage détecté et la surface totale de l'image.

3.3.3 Extraction de plans Personne X : critère multimodal (audio-video-texte) pour la recherche d'un document vidéo

La définition d'un plan personne X est la suivante : *"segment contains video of person X",* le segment (plan vidéo) contient la présence de la personne X. La tâche a été proposée dans la campagne d'évaluation TRECVID 2003. La présence de la personne est considérée seulement du point de vue visuel, mais toute information disponible (audio, vidéo, transcription, etc.) peut être utilisée. Dans notre cas, la personne X est l'ancien secrétaire d'état américaine Madleine Albright. Un exemple de réponse à la requête "je voudrais avoir tous les plans vidéo contenant la présence de Madeleine Albright" est donné dans la Figure IV. 3.

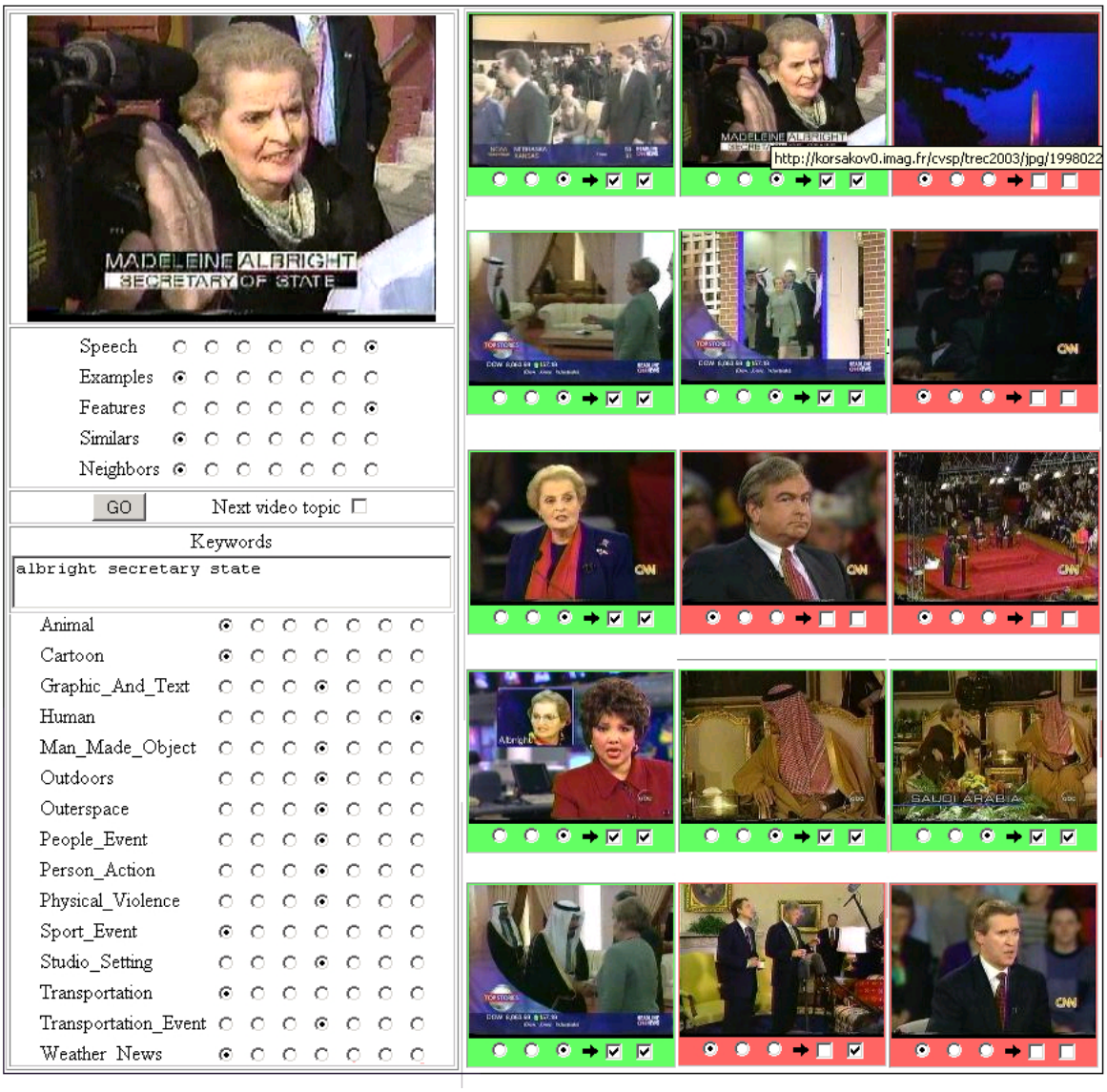


Figure IV. 3 Recherche de plans vidéo contenant Madleine Albright dans une collection de documents vidéo.

La requeté présentée en exemple dans la Figure IV. 3 était faite à partir des mots clés issus d'une transcription automatique et à partir d'une annotation manuelle du corpus vidéo selon certains traits prédéfinis (extérieur, studio, météo, etc.). L'utilisateur peut sélectionner les plans vidéo qui lui semblent pertinents et raffiner ensuite sa recherche.

Dans le système que nous avons utilisé, les sources d'informations considérées sont : vidéo, transcription automatique, audio. L'information vidéo est représentée par les frontières de plans vidéo. La transcription automatique est la transcription fournie par le laboratoire LIMSI [Gauvain 2002] à tous les participants à l'évaluation. Nous avons utilisé cette transcription uniquement pour détecter les plans vidéo contenant de la parole. Ensuite, nous

avons utilisé une approche issue de la reconnaissance du locuteur : pour chaque plan vidéo considéré comme étant uni-locuteur, un rapport de vraisemblance selon deux modèles M_{PersX} et $M_{NonPersX}$ est calculé :

$$llr(X) = \log P(X / M_{PersX}) - \log P(X / M_{NonPersX}) \qquad \text{(IV. 5)}$$

Les deux modèles M_{PersX} (le modèle du locuteur cible Madleine Albright) et $M_{NonPersX}$ (le modèle du monde) sont des modèles de mélanges de gaussiennes d'ordre 256 avec des matrices de covariances diagonales. Le modèle $M_{NonPersX}$ est entraîné sur tout le corpus de développement sauf les segments contenant la personne cible.

Comme la quantité de données disponibles pour le locuteur cible est d'environ 90 secondes, le modèle M_{PersX} a été obtenu par adaptation MAP d'un modèle du monde. Le modèle du monde est entraîné à partir données femme issues de la base de données Hub4 [LDC 1998].

Les données d'entraînement correspondant à la personne cible (Madleine Albright) sont issues de conditions d'enregistrements très variées. Le locuteur cible peut être interviewé dans le studio, il peut faire une déclaration devant plusieurs centaines de personnes, donc dans un milieu très bruité, il peut aussi être appelé au téléphone. Les conditions d'enregistrement dans les données d'entraînement ne sont pas forcement les mêmes que celles des données de test. Elles peuvent se retrouver partiellement ou pas du tout dans les données de test. Cette différence entre les données d'entraînement et les données de test peut avoir des conséquences très négatives sur le résultat du système de reconnaissance du locuteur utilisé.

3.4. Premiers résultats sur les évaluations TREC

Dans cette section, nous présentons les résultats que nous avons obtenu pour les tâches d'extraction automatique de traits à partir de la vidéo. Ces résultats correspondent aux systèmes proposés lors des évaluations TREC 2002 (Parole et Monologue) et TREC 2003 (Personne X). Les résultats sont synthétisés dans le Tableau IV. 3. A titre informatif nous présentons dans le Tableau IV. 4 les résultats pour le système de détection de visage utilisé pour la tâche Monologue.

Tableau IV. 3 Résultats officiels TREC : extraction de plans vidéo contenant un trait (TREC 2002 : parole, monologue et TREC 2003 : personne X). Les mesures de performance utilisées sont : précision (P 1000), rappel (R 1000), F ratio (F 1000) sur 1000 plans et la précision moyenne (PM)

Tâche	P 1000	R 1000	F 1000	PM
Parole	0,997	0,721	0,837	0,649
Monologue	0,014	0,368	0,027	0,117
Personne X	0,004	0,095	0,008	0,010

Tableau IV. 4 Performance du système de détection de visages : nombre de plans pertinents sur premiers 100 (D 100), nombre de plans pertinents sur premiers 1000 (D 1000) et précision moyenne

D 100	D 1000	PM
70	118	0,176

La précision moyenne obtenue pour la tâche parole, qui est le score officiel de l'évaluation nous situe 7ème sur 13 participants dans le classement final sachant que l'écart entre les différents participants est très faible.

La précision est élevée car il existe un très grand nombre de plans vidéo contenant de la parole dans le corpus : 1382 plans vidéo contenant de la parole parmi les 1850 plans vidéo du corpus de test. La tâche est également facilitée par le fait que, par rapport aux journaux télévisés, le flux audio est assez homogène : il y a très peu de parole bruitée et très peu de parole superposée sur la musique.

En ce qui concerne la tâche monologue nous allons regarder d'abord les résultats du système de détection de visage seul présentés dans Tableau IV. 4. Les résultats du système de détection de visage que nous avons utilisé sont très moyens à cause de la simplicité de l'approche. Les résultats du système de détection de visage sont ensuite combinés avec le résultat du système de segmentation en locuteurs, comme expliqué dans la section 3.3.2.

Le système soumis est classé 3ième dans le bilan final sur 9 participants. Malgré le faible taux de précision de la détection de visage, et malgré le fait que le système de segmentation utilisé est celui de 2002 seulement, on obtient finalement des résultats qui se comparent favorablement aux autres participants. Il faut noter que sur cette tâche, le nombre de plans vidéo pertinents est beaucoup plus faible par rapport à la tâche précédente. Il existe 38 plans vidéo contenant du monologue parmi les 1850 plans vidéo du corpus.

Pour la tâche Personne X, notre approche purement audio basée sur des méthodes spécifiques à la reconnaissance du locuteur ne donne pas de résultats acceptables. L'explication de ce taux de précision moyen extrêmement faible vient du fait que la personne cible parle seulement sur 5 plans vidéo parmi les 42 plans pertinents (et sur une collection de 35000 plans au total). On voit donc bien que pour cette tâche, il aurait fallu utiliser un système de reconnaissance de visage d'une part, et un système utilisant les informations de transcription du flux audio d'autre part [Amir 2003]. Ceci dit, au vu des résultats officiels TRECVID 2003, même avec cela, les performances sur cette tâche très difficile, restent assez faibles.

Nous pouvons tout de même noter que les 5 plans vidéo où la personne cible parle sont classés selon le rapport de vraisemblance dans les 6 % premiers documents parmi les 30000 plans contenant de la parole. Une des explications possibles pour ces résultats décevants, même pour les plans où la *personne X* parle, est la grande variété des conditions audio sur les données de test. Une autre explication possible est que généralement, dans les conditions "classiques" (type NIST) de vérification du locuteur, le rapport entre le nombre de locuteurs cible et le nombre d'imposteurs est d'environ 1/10. Dans notre cas le rapport est beaucoup plus faible, environ 1/6000 (5/30000) ce qui change complètement la difficulté de la tâche.

Le fait de sélectionner les plans vidéo où la personne cible parle parmi les 6 % de la totalité des plans vidéo laisse tout de même envisager une utilisation de modèles de locuteurs pour éliminer une partie des documents à traiter dans une tâche de type *personne X*.

3.5. Enseignements tirés des campagnes TREC pour notre travail de thèse

Dans le contexte de ces évaluations, nous nous sommes intéressé aux tâches qui peuvent montrer l'intérêt du traitement de l'audio en général et de la segmentation en locuteurs en particulier pour l'indexation de documents vidéo.

Cette participation nous a permis de nous familiariser avec les particularités du domaine de la recherche d'information multimédia sur de grandes collections de données. Les métriques d'évaluation sont notamment complètement différents de celles utilisées dans la segmentation en locuteurs. La difficulté des tâches est très variable et pour certaines d'entre elles, comme la tâche personne X, l'utilisation d'informations issues de modalités différentes est indispensable pour arriver à des résultats acceptables.

A partir des résultats obtenus, nous pouvons conclure que la difficulté de chacune des tâches semble corrélée avec le nombre de modalités qu'il faut utiliser pour leur mise en œuvre,

mais aussi avec la probabilité *a priori* d'apparition du trait dans le corpus d'évaluation. La tâche la plus simple est la détection de plans vidéo contenant de la parole, où seulement la voie audio est utilisée et la probabilité d'apparition *a priori* du trait dans le corpus est très élevée (1382 plans vidéo contenant de la parole parmi les 1850 plans vidéo du corpus 2002). La tâche la plus difficile est la tâche personne X, une approche multimodale est nécessaire pour des résultat acceptables (l'identification vocale n'apporte que très peu d'aide sur cette tâche) et la probabilité d'apparition *a priori* du trait dans le corpus est très faible (42 plans vidéo parmi environ 30000 plans vidéo du corpus 2003).

Par ailleurs, nous avons vu lors de ces évaluations que le traitement de l'audio peut être utilisé pour l'indexation d'une collection de documents vidéo. Mais, nous n'avons pas vu encore si inversement, le traitement de la vidéo peut être utilisé pour le traitement de l'audio, notamment pour la segmentation en locuteurs qui est notre principal intérêt.

Au vu de ces enseignements, tirés de notre participation officielle à l'évaluation TREC, nous avons donc bâti un plan d'expériences constitué de tâches inspirées de la campagne TREC, afin de voir plus précisément encore l'apport d'un système de segmentation en locuteurs pour la recherche d'information multimodale aussi que, en contrepartie, l'apport de la modalité vidéo pour la segmentation en locuteurs.

4. Tâches inspirées de la campagne d'évaluation TRECVID

Nous avons vu dans l'introduction de ce chapitre que la plupart des systèmes de traitement d'informations multimédia traitent séparément les signaux audio et vidéo. Comme les deux signaux proviennent de la même source, il semble qu'un système optimal de traitement de l'information devrait exploiter les informations conjointes entre les deux signaux.

Dans cette partie, nous essayons de proposer une utilisation conjointe de l'information audio et vidéo pour améliorer les performances de chaque source d'information prise séparément. Nous nous sommes intéressés à l'utilisation de frontières de plans vidéo pour améliorer notre système de segmentation en locuteurs et pour rechercher un locuteur particulier dans une vidéo : le présentateur du journal. Concernant la détection du présentateur, nous tenterons également d'exploiter la structure particulière d'un journal télévisé pour effectuer cette tâche sans aucun modèle *a priori* du locuteur cible.

Enfin, nous tenterons d'évaluer l'apport de notre système de segmentation en locuteurs sur une tâche de segmentation de vidéo à un plus haut niveau sémantique : la segmentation en histoires présentée dans la section 5.

4.1. Utilisation de la détection de frontières de plans vidéo dans un système de segmentation en locuteurs (comment l'information vidéo peut aider au traitement de l'audio ?)

Dans cette section, nous avons essayé d'améliorer les performances de notre système de segmentation en locuteurs présenté dans la deuxième partie de la thèse, en utilisant l'information de la détection de frontières de plans vidéo.

En regardant les enregistrements vidéo disponibles, nous avons remarqué que la plupart des plans vidéo contiennent la parole d'un seul locuteur. Il existe toutefois environ 20 % de plans vidéo contenant la parole de plusieurs locuteurs.

La détection de changements de locuteur présentée dans la deuxième partie de la thèse, est basée uniquement sur le traitement de l'audio. L'utilisation de l'information vidéo, non disponible sur les données des évaluations NIST SpRec 2002 et RT 2003, n'a jamais été envisagée dans la littérature.

Nous rappelons ici qu'une bonne détection de changements de locuteur doit donner des segments purs et de taille suffisamment grande pour pouvoir ensuite calculer des mesures fiables sur ces segments, son but étant de détecter tous les changements de locuteur tout en gardant un faible taux de fausses alarmes.

Nous proposons alors de détecter les changements de locuteur en utilisant aussi la détection de frontières de plans vidéo. En ce qui concerne la détection de changements de locuteur qui se trouvent *à l'intérieur* d'un plan vidéo, celle-ci sera laissée entièrement à la charge de la partie audio du système de segmentation.

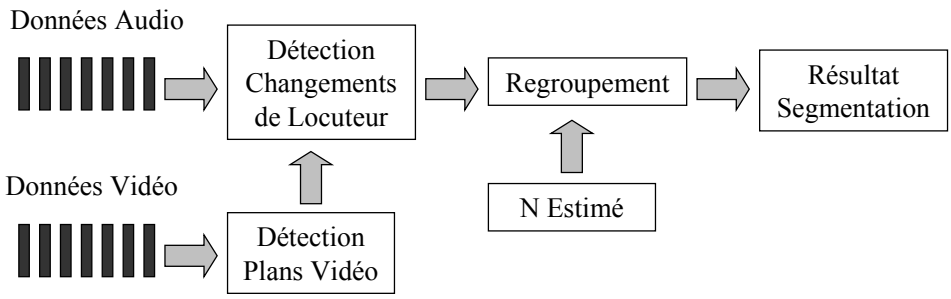


Figure IV. 4 Le système audio/vidéo de segmentation en locuteurs

Le système *audio/vidéo* de segmentation en locuteurs est présenté dans Figure IV. 4. Tous les changements de locuteur détectés, soit par la voie audio, soit par la voie vidéo, seront ensuite envoyés vers le module de regroupement et seront traités de la même façon que pour le système de segmentation audio pur.

Dans nos expérimentations, nous allons étudier et comparer trois différentes détections de changements de locuteur (Figure IV. 5) : uniquement audio (BIC) comme dans la deuxième partie de la thèse, uniquement vidéo en utilisant la détection de frontières de plans vidéo (Plans) et en assemblant les deux voies audio et vidéo (BIC + Plans).

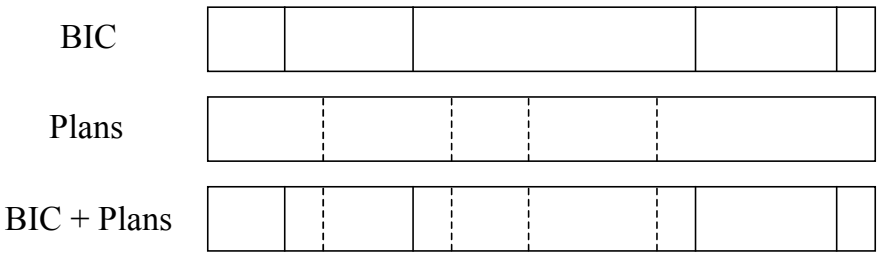


Figure IV. 5 Differents detecteurs de changements de locuteur

Les résultats de la segmentation en locuteurs pour les trois cas présentés dans la Figure IV. 5 seront présentés dans la section 4.3.1.

4.2. Identification sans données a priori du présentateur d'un journal télévisé sur un document vidéo

Parmi les locuteurs présents dans un document vidéo d'un journal télévisé, il en existe un qui peut être très important pour d'autres tâches comme la segmentation en histoires ou la création d'un résumé automatique ou tout simplement pour mieux structurer le document. Ce locuteur est le présentateur du journal télévisé ou radiodiffusé. Une première possibilité pour extraire les interventions du présentateur est l'utilisation d'un modèle du locuteur pour le présentateur. Un modèle GMM classique peut être appris sur des données audio issues de journaux précédents. On peut facilement obtenir une quantité de données suffisamment grande pour apprendre un modèle fiable. C'est ce que nous avons présenté dans la section 3.2 de la troisième partie de cette thèse.

Le principal désavantage d'une telle approche est que le présentateur peut changer pendant une série d'enregistrements du même journal. Par exemple, le présentateur du journal

le week-end n'est souvent pas le même que le présentateur du journal des autres jours de la semaine. Dans ce cas, la seule solution est de réapprendre un nouveau modèle du locuteur pour le nouveau présentateur. La portabilité d'une telle approche est donc assez limitée.

Une autre approche consiste à utiliser nos connaissances concernant le tournage d'une émission d'information télévisée ou radiodiffusé. La structure d'une émission d'information peut être décrite et se retrouve parfois, à quelques différences près, d'une chaîne à l'autre, et même d'un pays à l'autre.

Un système de segmentation, par contre, ne peut pas donner des identités aux locuteurs détectés. Les locuteurs sont identifiés par des noms génériques tells quels : L0, L1, L2, etc. La question que nous intéresse est alors : est il possible de reconnaître parmi les noms génériques le présentateur ?

Regardons maintenant deux exemples de résultats de segmentation automatique en locuteurs obtenus sur deux types de journaux : a) radiodiffusé et b) télévisé, et sur deux langue différentes : a) français et b) anglais. Chaque ligne représente un locuteur et chaque zone noire indique une intervention détectée par notre système de segmentation automatique pour le locuteur considéré.

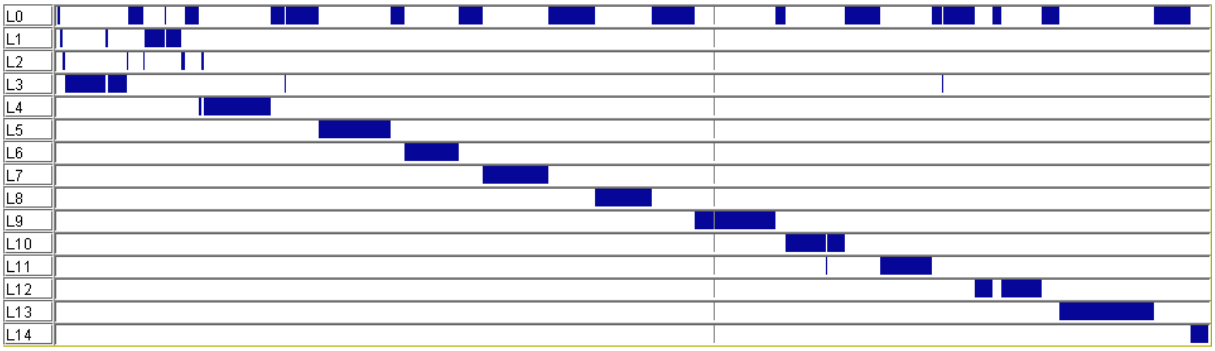


Figure IV. 6 Sortie d'un système de segmentation en locuteurs sur un journal français radiodiffusé de 20 minutes. Le présentateur est le locuteur L0 (premiere ligne)

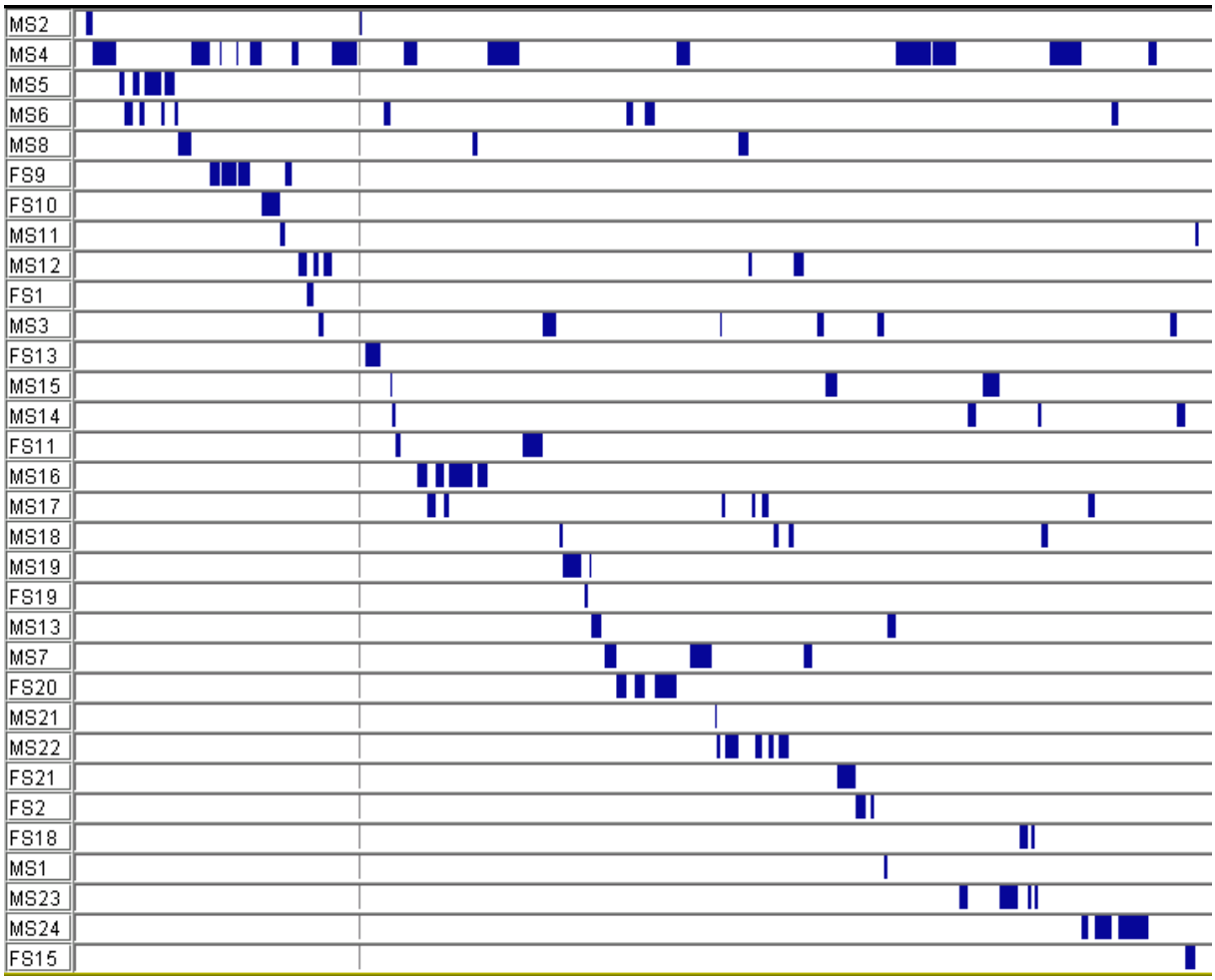


Figure IV. 7 Sortie d'un système de segmentation en locuteurs sur un journal anglais télévisé de 30 minutes. Le présentateur est le locuteur MS4 (deuxieme ligne)

En regardant les résultats de segmentation présentés dans la Figure IV. 6 et la Figure IV. 7 nous voyons que la ligne du présentateur semble identifiable car, contrairement aux autres locuteurs, qui parlent de façon occasionnelle et pour une période relativement courte, le présentateur du journal a des interventions qui sont réparties de façon relativement uniforme sur tout le journal.

En tenant compte de l'observation précédente, la ligne du présentateur peut être identifiée sans modèle *a priori*, uniquement à partir des règles empiriques suivantes :

- le présentateur du journal parle toujours pendant les cinq premières minutes d'un journal télévisé ;
- les interventions du présentateur d'un journal sont réparties de façon uniforme sur tout le journal contrairement aux interventions d'un reporter qui parle uniquement pendant son reportage ;

- le présentateur du journal est généralement le locuteur majoritaire du document, le locuteur qui parle le plus, même si ceci n'est pas systématique.

La première règle est appliquée telle quelle. Toute personne qui parle pendant les cinq premières minutes du journal est considérée comme étant un présentateur potentiel. Pour la deuxième règle, l'enregistrement est divisé uniformément en intervalles de taille fixe (3 minutes chacun). Pour chaque locuteur, on compte le nombre d'intervalles sur lesquels il ne parle pas. Le locuteur qui a le plus petit nombre d'intervalles sans intervention est sélectionné. Enfin, si après l'application de cette seconde règle nous avons toujours plus d'un présentateur potentiel, le critère final est la quantité de parole attribuée à chaque locuteur (troisième règle).

4.3. Expérimentations et résultats

Les résultats de cette section ont été obtenus sur le corpus TREC 4F 2003 composé de 4 enregistrements de journaux télévisés anglais (ABC, CNN) issus du corpus d'évaluation TREC 2003. Nous présentons les expériences faites avec notre système de segmentation audio/vidéo en locuteurs décrit dans la section 4.1 et celle correspondant à la recherche des interventions du présentateur décrite dans la section 4.2.

4.3.1 Segmentation audio/vidéo en locuteurs

Pour l'étape de détection de changements de locuteur, nous avons utilisé les stratégies présentées dans la Figure IV. 5. Les trois types de segments mis à l'entrée du module de regroupement sont : uniquement audio (par BIC), uniquement vidéo (par détection de frontières de plans vidéo) et combiné audio/vidéo (en fusionnant les changements détectés sur les deux modalités).

Tableau IV. 5 L'erreur de segmentation sur le corpus TREC 4F 2003 en utilisant différentes stratégies de détection de changements de locuteur

Détection de changements de locuteur	Audio (BIC)	Vidéo (Plans)	Audio/Vidéo (BIC + Plans)
Erreur de Segmentation	29,72 %	30,09 %	26,78 %

Les résultats du Tableau IV. 5 montrent que la meilleure détection de changement de locuteur est la détection audio/vidéo. En fait, l'utilisation des frontières de plans vidéo permet de fournir des segments homogènes issus de plans uni-locuteur qui sont majoritaires (presque

80% après analyse) et la détection de changements de locuteur basée classiquement sur le canal audio permet de découper en segments homogènes les plans longs contenant de la parole de plusieurs personnes. La détections de changements de locuteur conjointe audio/vidéo mène ainsi à une réduction de l'erreur de segmentation initiale de 3 % en absolu.

4.3.2 Extraction de plans où le présentateur parle

Cette partie décrit les résultats expérimentaux obtenus concernant la recherche du présentateur d'un journal, sans données *a priori*, selon les règles qui exploitent la structure d'un journal télévisé, énoncées dans la section 4.2.

Comme dans l'expérience précédente, le système de segmentation en locuteurs utilise toujours les trois stratégies de détection de changements de locuteur présentées auparavant : audio, vidéo et audio/vidéo. Les résultats de recherche du présentateur sont présentés d'abord en terme d'erreur de détection (comme cela a été fait dans la section 3.2 de la partie précédente : erreur proportionnelle au temps de signal mal étiqueté), et d'autre part en terme de rappel et précision concernant les plans vidéo où le présentateur parle.

Tableau IV. 6 L'erreur de détection du présentateur sur le corpus TREC 4F 2003

Détection de changements de locuteur	Audio (BIC)	Vidéo (Plans)	Audio/Vidéo (BIC + Plans)
Erreur de détection	13,90	15,61	10,80

Tableau IV. 7 Taux de rappel, précision et F ratio pour l'extraction de plans où le présentateur parle, sur le corpus TREC 4F 2003

Détection de changements de locuteur	Audio (BIC)	Vidéo (Plans)	Audio/Vidéo (BIC + Plans)
Rappel	0,69	0,72	0,71
Précision	0,93	0,84	0,93
F ratio	0,79	0,78	0,81

L'erreur de détection présentée dans le Tableau IV. 6 est obtenue comme dans la Figure IV. 8, à partir d'un étiquetage de référence et d'une sortie automatique indiquant "présentateur" ou "inconnu".

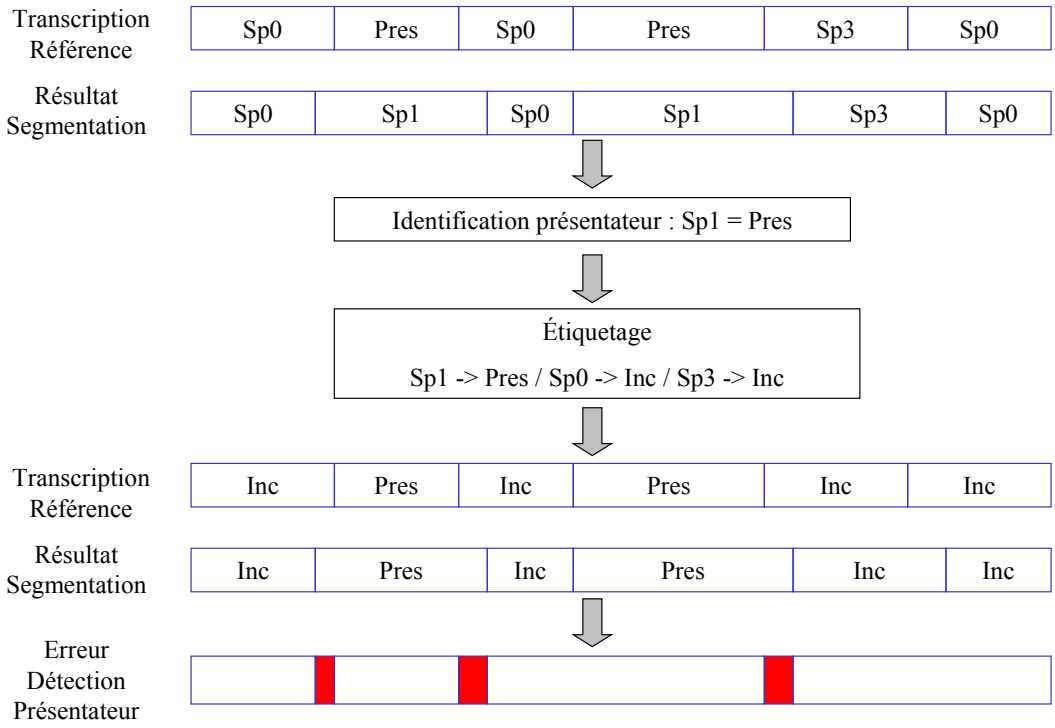


Figure IV. 8 Calcul de l'erreur de détection du présentateur

Les résultats obtenus dans le Tableau IV. 6 confirment les résultats obtenus par le système de segmentation audio/vidéo présentés dans le Tableau IV. 5. L'utilisation de la détection de frontières de plans vidéo combinée avec la détection audio de changements de locuteur donne au final une réduction d'erreur de détection de 3% en absolu, comme pour la segmentation en locuteurs. Ce taux d'erreur n'est pas vraiment comparable avec celui qu'obtiendrait un système de suivi du locuteur classique. Par exemple, les taux d'erreur en suivi de locuteur obtenus par les participants aux évaluations ESTER sont beaucoup plus bas, jusqu'à 2% d'erreur de détection. Mais il ne faut pas oublier que notre résultat est obtenu sans modèle *a priori* du présentateur. L'avantage est que notre système peut fonctionner sans nécessiter un ré-apprentissage lorsqu'un nouveau locuteur apparaît. Son défaut est, qu'en revanche, son utilisation est limitée à des documents de type journaux télévisés présentant une structure particulière.

Les améliorations sont aussi visibles dans le Tableau IV. 7 en termes de meilleur taux de précision et meilleur F ratio pour un taux de rappel comparable avec la détection vidéo.

D'un point de vue recherche d'information, la détection de plans où le présentateur parle semble donc assez robuste et pouvoir apporter un aide importante à d'autres taches d'un plus haut niveau d'abstraction comme le résumé automatique de vidéos ou la segmentation en histoires qui fera le sujet de la prochaine section.

5. Vers une segmentation vidéo multimodale : la segmentation en histoires

A partir de l'année 2003, une nouvelle tache est apparue pour les évaluations TRECVID. C'est la *segmentation en histoires*, qui consiste à détecter des frontières de sujets de journaux télévisés.

Une histoire est définie comme un segment de journal télévisé avec un sujet central (*news focus*) cohérent qui contient au moins deux énoncés déclaratifs. Tout autre segment cohérent est considéré comme étant divers (*miscelaneous*) et correspond à des séquences de publicité, des jingles, ou des conversations spontanées (*chit-chat*) entre les reporters.

5.1. Un premier système de segmentation en histoires proposé au CLIPS

La détection de frontières d'histoires sur des documents audio-vidéo illustre parfaitement le problème de *gap sémantique* bien connu en recherche d'information multimodale. L'information que nous désirons retrouver (les frontières d'histoires) est très éloignée de toute information qu'on peut extraire directement à partir du signal audio ou vidéo, comme la détection de plans vidéo ou la détection de changements de locuteur.

Une histoire peut contenir plusieurs plans vidéo. Par exemple, le présentateur du journal annonce le sujet et donne la parole à un reporter, plusieurs personnes sont interviewés par le reporter, et le sujet est ensuite clôturé à nouveau dans le studio par le présentateur. Dans d'autres cas, un plan vidéo peut contenir plusieurs histoires. C'est le cas typique des plans de début de journal ou le présentateur annonce successivement les principaux titres qui constituent autant d'*histoires* différentes.

Une histoire peut également être précédée d'un court générique (jingle) comme la rubrique météo par exemple. Une approche qui tiendrait compte de multiples informations issues de multiples détecteurs est alors indispensable.

Pour la détection de frontières d'histoires nous avons fait plusieurs observations empiriques des données de développement TREC 2003. A partir de ces observations nous pouvons dire que dans certains cas :

a. un long silence peut correspondre à un changement d'histoire ;
b. un changement de plan vidéo peut correspondre à un changement d'histoire ;
c. un changement de locuteur peut correspondre à un changement d'histoire ;

d. le début ou la fin d'une intervention du présentateur du journal peut correspondre à un changement d'histoire ;
e. un jingle (ou générique) peut correspondre à un changement d'histoire ;
f. certains mots clés détectés à partir de la transcription automatique peuvent correspondrent à un changement d'histoire.

Au vu de ces hypothèses, un premier système de segmentation en histoires a été développé au CLIPS et présenté aux évaluations TREC 2004. Nous ne présenterons pas tous les détails techniques des détecteurs utilisés dans ce système, mais insisterons seulement sur l'intérêt des détecteurs utilisant l'information *locuteur* pour cette tâche, qui correspondent aux observations c) et d) faites précédemment (détection de changements de locuteur et détection du présentateur). Tous les détails techniques sur le système du CLIPS de segmentation en histoires sont présentés dans [Besacier 2004]. Pour résumer, la difficulté vient du fait que chaque détecteur donne lui même des frontières qu'il va falloir ensuite fusionner pour rendre une décision finale de frontières d'histoires.

5.2. Premiers résultats expérimentaux

La segmentation en histoires est évaluée en termes de taux de rappel et précision. Dans la *vérité terrain* disponible, les histoires se succèdent l'une après l'autre sans intervalle de pause entre elles et sans se superposer. Une frontière d'histoire est considérée comme correctement détectée si le système a placé une frontière d'histoire dans un intervalle de 5 secondes autour de la frontière de référence.

Les performances pour la tâche de segmentation en histoires, obtenues séparément pour les différentes informations utilisées dans le système du CLIPS, sont présentées dans Tableau IV. 8. Les performances du système total résultant de la fusion de toutes ces informations multimodales sont également présentées sur la dernière ligne de ce tableau.

Tableau IV. 8 Segmentation en histoires, taux de rappel, taux de précision et F Ratio (évaluation réalisée sur 50 heures de signal vidéo du corpus TREC 2003)

Information	Rappel	Précision	F Ratio
a) Pauses	0.613	0.344	0.44
b) Frontières Plans	0.934	0.142	0.25
c) Changements de locuteur, BIC	**0.782**	**0.176**	**0.29**
d) Détection Présentateur	**0.205**	**0.702**	**0.32**
e) Jingles	0.028	0.735	0.05
f) Mots clés issus de la transcription automatique	0.252	0.657	0.36
Système total	**0.616**	**0.418**	**0.50**

Ces résultats montrent que l'information *locuteur* peut être intéressante pour aider un système de segmentation de plus haut niveau sémantique comme la segmentation en histoires. La détection de changement de locuteurs permet une bonne présélection des frontières d'histoire (rappel important), même si, bien sûr, ça n'est pas une information très précise car beaucoup d'histoires contiennent des interventions de plusieurs locuteurs. La détection de présentateurs est en revanche une information relativement précise (P=0.7) puisque le début de certaines *histoires* correspond exactement au début d'une nouvelle intervention du présentateur. Finalement, les performances du système global sont très encourageantes comparées à celles obtenues par les autres participants à la tâche (CLIPS vient de se classer 3e/6 participants sur cette tâche lors des évaluations TREC 2004), même avec une stratégie de fusion encore très naïve.

6. Conclusions

Dans cette partie, nous nous sommes intéressés à des tâches d'indexation d'une grande collection de documents vidéo. Depuis 1992, NIST (National Institute of Standards and Technology) organise la campagne d'évaluation TREC avec comme but d'encourager les travaux dans le domaine de la recherche d'information en fournissant des grandes collections

de données. Depuis 2001, les données vidéo ont commencé à faire partie des données d'évaluation.

Dans le contexte de ces évaluations, nous nous sommes intéressé particulièrement à l'apport de l'information *locuteur* extraite à partir du canal audio sur l'indexation de documents vidéo. Notre participation aux campagnes 2002 et 2003 nous a permis d'avoir une première expérience de la tâche de recherche d'information multimodale sur de grandes quantité de vidéos : métriques d'évaluation différentes, grande écart de difficulté entre les tâches, nécessité d'utiliser une information multimodale, etc.

De cette expérience, nous avons défini nos propres tâches de recherche, intéressantes d'un point de vue *locuteur*, dans le but de répondre à deux questions principales : est-ce que la segmentation en locuteurs peut apporter des informations utiles pour l'indexation de documents vidéo ? Et, inversement, est-ce que les informations extraites à partir du canal vidéo peuvent être utiles pour améliorer le résultat d'un système de segmentation en locuteurs ?

Nous nous sommes donc intéressé à l'apport de l'information vidéo sur une tâche purement audio qui est la segmentation en locuteurs. Les résultats obtenus montrent une réduction d'environ 3 % en absolu de l'erreur de segmentation en utilisant un système de segmentation audio/vidéo.

Nous avons également proposé une méthode originale pour la détection du présentateur d'un journal télévisé. L'originalité de l'approche vient de la non-utilisation d'un modèle probabiliste du locuteur cible, comme cela est le cas dans un système de suivi de locuteur classique. Le présentateur est identifié uniquement à partir du résultat du système de segmentation en locuteurs et des quelques règles empiriques issues de connaissances *a priori* sur la structure des documents de journaux télévisés.

Dans la dernière partie, nous abordons la problématique de la segmentation d'un document vidéo en utilisant des informations issues de détecteurs différents (texte, audio, vidéo) : la segmentation en histoires. Nous avons vu que les informations *locuteur* sont exploitables pour améliorer un système général de segmentation de vidéo, notamment la sortie d'un détecteur de présentateur, qui joue un rôle important.

Conclusions et perspectives

Cette partie résume le travail réalisé dans cette thèse et met en évidence nos contributions. Nous y présentons également les perspectives à court terme et à long terme de nos travaux.

1. Conclusions

Dans cette thèse, nous avons abordé le problème de la segmentation en locuteurs de documents audio et son application dans un cadre plus vaste, au delà du domaine audio, qui est l'indexation d'une grande collection de documents vidéo.

La quantité de données numériques qui est disponible pour le grand public sur le web, sur les chaînes média numériques (télévisions et radio) ou dans les archives de différents instituts et bibliothèques est de plus en plus importante. L'annotation selon le contenu d'une telle quantité de données est indispensable pour un accès facile à ce type de corpus. La segmentation en locuteurs est une des premières étapes vers une annotation structurée d'un document multimédia. Nous exposons dans un premier temps les principales contributions de cette thèse avant d'évoquer les perspectives de ce travail.

1.1. Evaluation d'un système sur le long terme et sur plusieurs types de données.

Notre travail a suivi la méthodologie définie dans le chapitre d'introduction. Dans un premier temps, nous avons réalisé un système de segmentation en locuteurs de référence. Ce système, pour lequel nous retrouvons les principales étapes de l'architecture générale d'un système de segmentation (la paramétrisation, la détection de changements de locuteur et le regroupement) a fait ses preuves pendant les évaluations NIST SpRec 2002. Lors de ces évaluations, malgré sa simplicité, notre système a notamment obtenu le meilleur résultat sur la tâche de segmentation en locuteurs de documents de réunion.

Cependant, ce système initial ne répondait pas directement aux trois problèmes énoncés dans l'introduction de cette thèse. Ces trois problèmes étaient : la possibilité d'estimer de façon automatique le nombre de locuteurs présents dans un document (p1), la possibilité de traiter un flux audio hétérogène contenant de la parole de qualité variable, de la musique, du silence, etc. (p2) et la nécessité de traiter de la parole spontanée avec nombreux recouvrements entre les interventions des locuteurs (p3). Les deux premiers problèmes sont spécifiques aux données de type journaux télévisés et radiodiffusés. Comme la plupart de nos travaux ont été faits sur ce type de données, ces deux problèmes ont été abordés en priorité dans cette thèse.

Notre système initial a été alors amélioré au niveau de la paramétrisation et en proposant une méthode d'estimation du nombre de locuteurs présents dans un enregistrement. Ce module, indispensable pour les enregistrements de type journaux télévisés et radiodiffusés proposés dans les évaluations NIST et ESTER, répond au premier problème énoncé (p1).

Ces améliorations sont le résultat d'une expérience accumulée au fur et à mesure des différentes campagnes d'évaluations auxquelles nous avons participé pendant plusieurs années (voir récapitulatif Tableau 1). Ces évaluations nous ont permis de tester notre système de référence dans des conditions expérimentales solides, sur une grande quantité des données de différents types (téléphoniques, réunions, journaux) et dans deux langues différentes : anglais et français.

Tableau 1 Evolution des taux d'erreur de notre système sur 3 ans et sur différentes données. Synthèse de résultats sur les corpus NIST SpRec 2002, NIST RT 2003 et NIST RT 2004 Meeting

Données	2002	2003	2004
Données Téléphoniques	16,58 %	-	-
Données Journaux	30,33 %	19,25 %	-
Données Réunions	50,20 %	-	22,6 %
Solutions proposées	-	p1 + p2	p3

1.2. Utilisation d'informations *a priori* pour la segmentation en locuteurs

L'expérience accumulée lors de ces campagnes d'évaluation nous a permis d'aborder ensuite une partie plus originale de la thèse. Par hypothèse, spécifique au cadre de ces évaluations, nous ne disposions d'aucune information concernant les documents analysés pour la segmentation en locuteurs.

Cependant, dans certaines applications pratiques et sur certaines données, il est possible de disposer d'informations *a priori* utiles avant de traiter l'enregistrement. Il existe deux grands types d'informations auxquelles nous pouvons avoir accès : des informations issues d'une annotation automatique et des informations issues d'une annotation manuelle.

L'utilisation des informations *a priori* issues d'une annotation automatique faite par un autre système, nous mènent généralement vers une réduction de l'erreur de segmentation. Nous avons vu notamment qu'une première annotation en classes acoustiques permet de traiter efficacement des documents hétérogènes (problème p2) en conservant un système de

segmentation en locuteurs performant. Dans le même esprit, une pré-segmentation en locuteurs issue d'un autre système, ou la sortie d'un système de suivi du locuteur cible, peuvent être utilisés avantageusement comme étape préliminaire pour affiner un système de segmentation. Par exemple, l'utilisation en cascade de notre système et de celui développé au LIA, nous a permis d'obtenir le deuxième meilleur résultat officiel lors des évaluations NIST RT 2003 parmi huit autres participants.

Concernant la deuxième catégorie d'informations, les informations issues d'une annotation manuelle, la connaissance *a priori* du nombre de locuteurs ne semble pas apporter d'amélioration à notre système, notamment à cause du fait que la méthode d'estimation automatique du nombre de locuteurs que nous avons proposée est déjà assez précise. Par contre, l'utilisation d'une annotation de référence incomplète qui génère des données de référence pour chaque locuteur conduit à une réduction importante de l'erreur de segmentation mais aussi du temps d'exécution. L'utilisation de ce type d'informations est tout à fait réaliste dans certaines conditions pratiques et permet d'optimiser le résultat d'un système de segmentation classique.

Le dernier aspect encourageant a été l'utilisation du signal audio issu de plusieurs capteurs différents, situation spécifique aux enregistrements de réunions. Sans utiliser de techniques de séparation de sources, nous avons extrait la parole appartenant à chaque locuteur à partir de résultats de segmentation obtenus individuellement sur chaque capteur. Cette approche a donné de bons résultats malgré la difficulté de la tâche due à la spontanéité de la conversation et nous a permis de proposer une première piste pour le troisième problème (p3). Le taux d'erreur de segmentation obtenu sur des données de réunion en 2004 est même comparable avec celui obtenu sur des documents de type journaux télévises.

Il est tout de même important de noter que nos conclusions en ce qui concerne l'influence des informations *a priori* sur la segmentation en locuteurs, sont en partie influencées par la mesure de performance utilisée. Le taux d'erreur actuel est calculé trame par trame et correspond à une proportion de signal mal étiqueté, ce qui donne la possibilité d'obtenir des performances acceptables sans détecter par exemple les locuteurs qui ne parlent pas beaucoup dans un document. Ainsi, nous sommes arrivés à la conclusion qu'il est mieux d'estimer le nombre de locuteurs plutôt que d'utiliser la connaissance *a priori* du nombre réel de locuteurs présents dans un enregistrement. Cependant, nous sommes bien conscients que certains de ces locuteurs *minoritaires*, mal détectés par notre système, pourraient justement s'avérer être les plus intéressants dans un tâche de recherche d'informations.

1.3. Application à la recherche d'informations multimédia

Dans la dernière partie de la thèse, nous nous sommes intéressé à l'application de notre système à une tâche d'indexation d'une grande collection de documents vidéo. Nos travaux de recherche ont été faits dans le cadre des évaluations TREC (Text REtrieval Conference) organisées également par NIST. Cette fois ci, l'évaluation était faite au niveau du plan vidéo qui est considéré comme une unité d'indexation élémentaire, et les mesures de performance utilisées étaient spécifiques au domaine de la recherche d'informations (précision, rappel).

Notre insertion dans ce domaine nouveau s'est faite via la participation officielle du laboratoire à certaines tâches proposés dans les évaluations TRECVID 2002 et 2003. De cette expérience, nous avons défini nos propres tâches de recherche, intéressantes d'un point de vue *locuteur*, dans le but de répondre à deux questions principales : est-ce que la segmentation en locuteurs peut apporter des informations utiles pour l'indexation de documents vidéo ? Et, inversement, est-ce que les informations extraites à partir du canal vidéo peuvent être utiles pour améliorer le résultat d'un système de segmentation en locuteurs ?

Concernant l'apport de l'information vidéo sur une tâche purement audio de segmentation en locuteurs, les résultats obtenus ont montré une réduction d'environ 3 % en absolu de l'erreur grâce à un système de segmentation en locuteurs "audio-visuel".

A partir d'un résultat de segmentation en locuteurs, nous avons proposé une méthode originale pour la recherche de plans vidéos contenant le présentateur d'un journal télévisé. L'originalité de l'approche est due à l'absence d'un modèle probabiliste du locuteur cible, présent dans tout système de suivi de locuteur classique. Le présentateur est identifié uniquement à partir du résultat du système de segmentation en locuteurs et avec quelques règles empiriques issues de connaissances sur la structure d'un journal télévisé.

Les dernières expérimentations présentent une approche multimodale pour la segmentation en histoires de documents vidéo. Dans cette approche, nous avons montré que les informations *locuteur* obtenues avec les méthodes présentées dans cette thèse, permettent d'aider à améliorer un système de découpage plus sémantique d'un document vidéo.

2. Perspectives

Malgré les progrès récents dans la segmentation en locuteurs concrétisés par une réduction d'erreur de segmentation lors des évaluations NIST (voir Tableau 1) nous sommes loin d'une méthode de segmentation en locuteurs parfaite.

Plusieurs questions restes ouvertes, notamment la difficulté de détecter les locuteurs qui ne parlent pas beaucoup appelés *minoritaires* et la difficulté de traiter la parole spontanée contenant du recouvrement entre différents locuteurs. Concernant le premier problème, une solution qui pourrait être envisagée est de donner plus d'importance à la phase de détection de changements de locuteur. Les interventions de locuteurs minoritaires sont généralement correctement détectées lors de la phase de détection de changements de locuteur, mais leur faible quantité de données n'arrive pas à former une classe à part lors de la phase de regroupement de segments. Le critère d'arrêt de la phase de regroupement, qui est actuellement utilisé par notre système, est une coupure du dendrogramme à une hauteur fixe. Par des observation empiriques, nous avons remarqué que cette méthode privilégie la création de classes finales qui ont une grande taille dès le début du regroupement. L'utilisation d'une méthode d'élagage de l'arbre du dendrogramme, pour la sélection du résultat final du regroupement, pourrait être une solution menant à de meilleurs résultats, privilégiant ainsi les classes qui contiennent moins de données au départ (locuteurs minoritaires).

Concernant le traitement de la parole spontanée (les locuteurs qui parlent en même temps, qui se coupent la parole, longueur d'interventions de tailles très variables, etc.), notamment dans le cas d'enregistrements de type réunion, les techniques utilisées dans la séparation de sources peuvent s'avérer particulièrement utiles. Ces méthodes permettent, à partir de signaux captés par plusieurs microphones, de séparer les différents locuteurs qui parlent en même temps à condition que le nombre de microphones soit suffisamment important par rapport au nombre de locuteurs superposés.

La segmentation en locuteurs est une meta-donnée importante pour la transcription enrichie de documents multimédia. Jusqu'à présent, la reconnaissance de la parole utilise la segmentation en locuteurs comme donnée *a priori* mais, par contre la segmentation en locuteurs n'utilise pas le résultat de la reconnaissance de la parole. La segmentation en locuteurs est faite uniquement à partir des indices purement acoustiques, mais l'utilisation d'indices linguistiques et prosodiques est parfaitement envisageable. Les informations linguistiques comme les mots clés pourraient être utilisées pour trouver les fins et les débuts des interventions de locuteurs. Les informations prosodiques comme la fréquence

fondamentale ou le débit verbal qui peuvent varier d'un locuteur à l'autre pourraient être utilisées pour la détection de changements de locuteur notamment dans des conditions de parole spontanée mais aussi pour nous mener vers la détection d'autres types de meta-données comme les *hésitations (disfluencies)* très fréquentes dans les conversations spontanées. Cette interaction entre plusieurs types d'informations différentes nous mènerait vers une intégration de notre système de segmentation en locuteurs dans un système de transcription enrichie qui reste un des objectifs principaux dans le domaine de la reconnaissance de la parole.

Un moteur de transcription enrichie peut utiliser plusieurs systèmes de segmentation en locuteurs en même temps. Nous avons proposé une méthode de fusion de résultats de segmentation en locuteurs mais cette méthode est appliquée uniquement sur les résultats finaux (ou sorties) des systèmes. Une suite logique serait d'envisager une interaction entre les systèmes également au niveau des résultats intermédiaires. Nous pourrions par exemple surestimer le nombre de locuteurs lors de la phase de regroupement hiérarchique, et utiliser ce résultat intermédiaire pour initialiser un autre système (comme nous l'avons fait pour le système LIA) et revenir ensuite dans le premier système.

Dans la partie de la thèse consacrée à l'application de notre système à une tâche de recherche d'information, nous avons proposé une première utilisation conjointe des informations audio issues de la segmentation en locuteurs et des informations vidéo. L'interaction entre les deux modalités reste cependant assez faible, dans le cas de la segmentation en locuteurs elle est réduite à seulement une concaténation du résultat de la détection de frontières de plans vidéo avec celui de la détection de changements de locuteur.

Plus d'interaction entre les différentes sources d'informations nous mènera vers des techniques de fusion de données plus complexes et vers de traitement de signaux plus complexes, notamment sur le signal vidéo. D'une façon générale nous pouvons parler de deux type d'interactions : *interactions tardives* et *interactions précoces*. Si les paramètres extraits à partir de chaque source d'information audio et vidéo sont traités de façon indépendante et si les résultats selon chaque source d'information sont fusionnés après le traitement, nous pouvons parler alors d'interaction tardive. Ce type d'interaction a été abordé dans la dernière partie de la thèse. Si les paramètres extraits à partir de chaque source d'information audio et vidéo sont fusionnés avant le traitement, nous pouvons parle alors d'interaction précoce. Le traitement appliqué pourrait être par exemple une détection de changements de locuteur suivi d'un regroupement hiérarchique à partir de vecteurs de paramètres audio-vidéo.

Enfin la segmentation en locuteurs peut être très bien utilisée directement dans un moteur de recherche de documents audio et/ou vidéo. Dans le système de recherche vidéo interactive de laboratoire CLIPS (http://korsakov0.imag.fr/cvsp/ctrec12.html), l'utilisateur peut faire sa requête à partir des mots clés, à partir d'une annotation selon différentes caractéristiques (paysage extérieur, intérieur studio, événement sportif, etc.), similarité d'image, etc. A partir des réponses proposées qui sont des plans vidéo, il peut sélectionner certains d'entre eux et raffiner sa recherche en demandant une recherche dans le voisinage temporel des plans vidéo sélectionnées. Avec un système de segmentation en locuteurs intégré, il pourrait alors raffiner sa recherche en demandant à partir des plans vidéo sélectionnés d'autre plans vidéo où les mêmes locuteurs parlent (voisinage acoustique).

Annexes

ANNEXE A

Démonstration de la formule de rapport de vraisemblance généralisé

Soit une séquence $x=\{x_n\},\ n=1\ldots N$ de N vecteurs acoustiques de dimension d modélisé par une gaussienne multidimensionnelle $G(\mu,\Sigma)$. Sa distribution de probabilité est donnée par :

$$P(x,G(\mu,\Sigma))=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \qquad \text{(A. 1)}$$

La vraisemblance de la séquence x par rapport a $G(\mu,\Sigma)$ est donnée par :

$$L(x;\mu,\Sigma)=\prod_{i=1}^{N}\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \qquad \text{(A. 2)}$$

Le logarithme de cette expression donne :

$$\log L(x;\mu,\Sigma)=-\frac{Nd}{2}\log(2\pi)-\frac{N}{2}\log|\Sigma|-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) \qquad \text{(A. 3)}$$

Soit deux séquences de vecteurs $x=\{x_n\},\ n=1\ldots N_1$ et $y=\{y_n\},\ n=1\ldots N_2$. Soit $z=\{z_n\},\ n=1\ldots N$ ($N=N_1+N_2$) l'union des deux séquences x et y. Soit les deux hypothèses suivantes pour les séquences de vecteurs x et y :

- H_0 : les séquences de vecteurs ont été générées par le même gaussienne multidimensionnelle $G(\mu,\Sigma)$;
- H_1 : les séquences de vecteurs ont été générées par deux gaussiennes différentes $x\approx G(\mu_1,\Sigma_1)$ et $y\approx G(\mu_2,\Sigma_2)$.

Le rapport de vraisemblance associé à ce test d'hypothèse est :

$$\lambda = \frac{L(z;\mu,\Sigma)}{L(x;\mu_1,\Sigma_1)\cdot L(y;\mu_2,\Sigma_2)} \tag{A. 4}$$

Le logarithme de ce rapport de vraisemblance est :

$$\begin{aligned} R &= -\log\lambda = -\log L(z;\mu,\Sigma) + \log L(x;\mu_1,\Sigma_1) + \log L(y;\mu_2,\Sigma_2) \\ &= \frac{N}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| + \frac{1}{2}\sum_{i=1}^{N}(z_i-\mu)^T\Sigma^{-1}(z_i-\mu) - \\ &- \frac{1}{2}\sum_{i=1}^{N_1}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) - \frac{1}{2}\sum_{i=1}^{N_2}(y_i-\mu_2)^T\Sigma_2^{-1}(y_i-\mu_2) \end{aligned} \tag{A. 5}$$

Si $\sigma[k,l]$ désigne l'élément de la k-ième ligne et de la i-ième colonne de Σ^{-1} et $z_i[j]$ et $\mu[j]$ les j-ièmes vecteurs éléments des vecteurs z_i et μ respectivement. Nous pouvons alors écrire :

$$\sum_{i=1}^{N}(z_i-\mu)^T\Sigma^{-1}(z_i-\mu) = \sum_{i=1}^{N}\left(\sum_{l=1}^{d}\sum_{k=1}^{d}(z_i[l]-\mu[l])\cdot\sigma[k,l]\cdot(z_i[k]-\mu[k])\right) \tag{A. 6}$$

Soit S la matrice de covariance estimée à partir des vecteurs et $s[k,l]$ ses éléments. Cette matrice est par définition de la forme :

$$S = \frac{1}{N}\sum_{i=1}^{N}(z_i-\mu)\cdot(z_i-\mu)^T \tag{A. 7}$$

$$s[k,l] = \frac{1}{N}\sum_{i=1}^{N}(x_i[k]-\mu[k])\cdot(x_i[l]-\mu[l]) \tag{A. 8}$$

Alors l'équation (A. 6) s'écrit :

$$\begin{aligned} \sum_{i=1}^{N}(z_i-\mu)^T\Sigma^{-1}(z_i-\mu) &= \sum_{l=1}^{d}\sum_{k=1}^{d}\sigma[k,l]\left(\sum_{i=1}^{N}(x_i[l]-\mu[l])\cdot(x_i[k]-\mu[k])\right) \\ &= N\sum_{l=1}^{d}\sum_{k=1}^{d}\sigma[k,l]\cdot s[k,l] = Ntr\left(\Sigma^{-1}S\right) \end{aligned} \tag{A. 9}$$

où $tr(\cdot)$ est la trace de la matrice.

Si nous faisons de plus l'hypothèse que $S = \Sigma$, ça veut dire que l'estimée de la matrice de covariance sur les échantillons est égale à la matrice de covariance théorique alors $\Sigma^{-1}S = I_d$ où I_d est la matrice identité d'ordre d.

Dans ce cas on peut écrire :

$$\sum_{i=1}^{N}(z_i - \mu)^T \Sigma^{-1}(z_i - \mu) = Nd \tag{A. 10}$$

Avec ce dernier résultat, l'équation (A. 5) :

$$\begin{aligned} R = -\log\lambda &= \frac{N}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| + \frac{1}{2}\sum_{i=1}^{N}(z_i - \mu)^T\Sigma^{-1}(z_i - \mu) - \\ &- \frac{1}{2}\sum_{i=1}^{N_1}(x_i - \mu_1)^T\Sigma_1^{-1}(x_i - \mu_1) - \frac{1}{2}\sum_{i=1}^{N_2}(y_i - \mu_2)^T\Sigma_2^{-1}(y_i - \mu_2) \\ &= \frac{N}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| + \frac{Nd}{2} - \frac{N_1 d}{2} - \frac{N_2 d}{2} \end{aligned} \tag{A. 11}$$

En tenant compte que $N = N_1 + N_2$ on retrouve la formule (II. 5) :

$$R = \frac{N_1 + N_2}{2}\log|\Sigma| - \frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| \tag{A. 12}$$

ANNEXE B

Bibliographie personnelle

Revues internationales en cours de soumission :

- "Approaches in broadcast news speaker segmentation", S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, L. Besacier, Computer Speech and Language, soumission prévue 1er Novembre 2004

Conférences internationales :

- "Speaker segmentation using audio-video information", D. Moraru, L. Besacier, G. Quénot, soumis à ICASSP 2005, Philadelphia, PA, USA, Mars 2005
- "Video Story Segmentation With Multi-Modal Features: Experiments on TRECVID 2003" L. Besacier, G. Quénot, S. Ayache, D. Moraru, Proc of ACM Multimedia Information Retrieval, New York, Etats-Unis, Octobre 2004
- "Using a priori information for speaker diarization", D. Moraru, L. Besacier, E. Castelli, A Speaker Odyssey 2004, Toledo, Espagne, Mai 2004
- "ELISA NIST RT 03 Broadcast News Speaker Diarization Experiments", D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, A Speaker Odyssey 2004, Toledo, Espagne, Mai 2004
- "The ELISA Consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation", D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, ICASSP 2004, Montreal, Canada, Mai 2004
- "Benefits of prior acoustic segmetation for automatic speaker segmentation", S. Meignier, D. Moraru, C. Fredouille, L. Besacier, J.-F. Bonastre, ICASSP 2004, Montreal, Canada, Mai 2004
- "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation", D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, I. Magrin-Chagnolleau. ICASSP 2003, Honk-Kong, China, Mai 2003
- "Towards Conversational Model for Speaker Segmentation", D., Moraru, L. Besacier, Proc of Sped 2003, Bucarest, Roumanie, Mai 2003

- "CLIPS-IMAG at TREC-11 : Experiments in Video Retrieval", G. M. Quénot, D. Moraru, L. Besacier, and P. Mulhem , 11th Text Retrieval Conference, Gaithersburg, MD, USA, Novembre 2002
- "Speech Translation for French in the NESPOLE! European Project", L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, D. Vaufreydaz, Eurospeech 2001, Aalborg, Danemark, Septembre 2001

Conférences nationales :

- "Premiers pas du CLIPS sur les données d'évaluation ESTER", R. Lamy, D. Moraru, B. Bigi, L. Besacier, JEP 2004, Fès, Maroc, Avril 2004
- "Segmentation selon le locuteur: les activités du Consortium ELISA dans le cadre de Nist RT03", D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J-F Bonastre, JEP 2004, Fès, Maroc, Avril 2004
- "Separation en locuteurs de conversations via IP", D. Moraru, L. Besacier, JEP 2002, Nancy, France, Juin 2002

ANNEXE C

Article Informations *a priori* conférence A Speaker Odyssey 2004

Using A Priori Information For Speaker Diarization

Daniel Moraru, Laurent Besacier, Eric Castelli

CLIPS-IMAG (UJF & CNRS)
BP 53 - 38041 Grenoble Cedex 9 - France

(daniel.moraru,laurent.besacier,eric.castelli)@imag.fr

Abstract

This paper presents an attempt to use supplementary information for audio data diarization. The approach is based on the use of *a priori* information about the speakers involved in dialogue. Those specific information are the number of speakers involved in conversation, and training data available for one speaker or for all the speakers involved in conversation. The experiments were mainly conducted on the 2003 Rich Transcription Diarization corpus both Dry Run Corpus and Evaluation corpus. The results show that knowing *a priori* the exact number of speakers seems not to be a very useful information. On the other hand, using *a priori* speaker models for one or all speakers involved in the conversation, may improve diarization performance when enough data is available to train reliable speaker models.

1. Introduction

Speaker diarization (or segmentation) is a new speech processing task resulting from the increase in the number of multimedia documents that need to be properly archived and accessed. One key of indexing can be speaker identity.

The goal of speaker diarization is to segment a N-speaker audio document in homogeneous parts containing the voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to a same speaker (clustering process). In most papers recently published in this new domain [1], [2], [3], an assumption is generally that no *a priori* information is available on the number of speakers involved in the conversation as well as on the identity of the speakers and on the nature of the conversation. A consequence of this is that no speaker reference is supposed to be available before segmenting an audio signal for instance.

This limitation may however not be so rough for some applications and conditions for which we can reasonnably hope to have *a priori* information. For instance, we may be informed of the type of conversation to be segmented (broadcast news, telephone or meeting data) for which speech signal quality and speaker turn length differ. One may also have some speaker reference models known *a priori*. For example, in a telephone meeting room system, one can ask to all the participants to orally present themselves before the meeting starts, in order to get few seconds of speech reference per speaker. In broadcast news documents, one may also have speaker references for a limited number of persons (news presenters for instance). Finally, information about the verbosity of speakers, i.e. wether they are known to speak a lot or not in a conversation, may be also an interesting information.

This paper is an attempt to investigate how helpful could be such kind of *a priori* information for speaker segmentation and how it compares with a reference state of the art segmentation system which uses no *a priori* information. Our state-of-the-art system was developped at CLIPS laboratory for both NIST 2002 and RT (See http://www.nist.gov/speech/tests/rt/rt2003/index.htm for more details) 2003 speaker segmentation evaluations.

Section 2 gives a brief description of the reference CLIPS diarization system used for the experiments. *Section 3* proposes some solutions for including *a priori* information in speaker diarization. The performance of the various propositions are shown and discussed in *Section 4*. Finally, *Section 5* concludes this work and gives some perspectives.

2. The Clips diarization system

This section presents the CLIPS diarization system [4] used during the NIST Rich Transcription Evaluation .

The CLIPS system is a state of the art diarization system based on a BIC (Bayesian Information Criterion) speaker change detector [5], [6], [7] followed by an hierarchical clustering. It uses the ELISA frame-work [8], [9] for the parametrization of the acoustic signal and for the speaker models used for clustering. The clustering stop condition is the estimation of the number of speakers using a penalized BIC criterion.

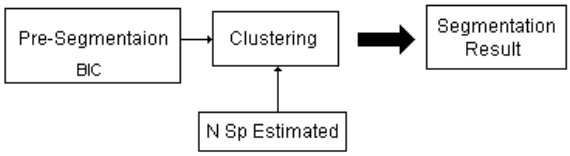


Figure 1: The CLIPS segmentation system

As a preliminary phase we use the LIA hierarchical acoustic pre-segmentation [10] in Speech / Non-Speech, gender and bandwidth. Finally we obtain 4 acoustic classes: male wideband, female wideband, male narrowband, female narrowband. The goal of the acoustic pre-segmentation is to eliminate the false alarm speech errors penalized by the error metric and to improve the performance of the clustering phase (e.g: by not clustering together segments labeled as male and female). The acoustic segmentation is applied individually on each class and the results are merged in the end.

A BIC approach is then used to define first potential speaker changes. A BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-Gaussian models with diagonal covariance matrices are used to model the two windows. A threshold is then applied on the BIC curve to find the most likely speaker change points which correspond to the local maximums of the curve.

Clustering starts by first training a 32 components GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion using a classical EM algorithm. Segments models are then trained using MAP adaptation [11] of the background model (means only). Next, BIC distances are computed between segment models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to the N speakers in the conversation).

The number of speakers is estimated (see section 3) using a penalized BIC criterion.

The signal is characterized by 16 mel Cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then the Cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied.

3. Including a priori information

Among the different types of *a priori* information, we are mainly interested in :

- number of speakers involved in the conversation;
- training data available for all speakers involved in conversation;
- training data available for only one or for few of the speakers involved in the conversation.

3.1 Knowing or not the number of speakers

3.1.1 Differences between the types of conversation

The number of speakers involved in dialogue is conversation type dependent: for telephone speech the number is fixed and is 2, for meeting speech the number can be usually known (the list of participants may be known before the meeting) but is greater than two, for broadcast news data the number of speakers is usually unknown and must be estimated.

If we know *a priori* the number of speakers involved in the conversation, since no estimation of the number of speakers is needed, our segmentation system changes as in the next figure:

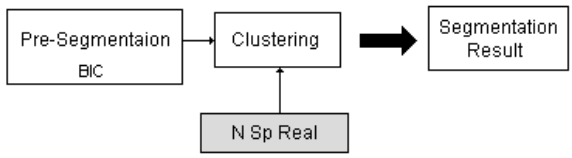


Figure 2: The segmentation system when the number of speakers is known

3.1.2 Estimating the number of speakers

When the number of speakers is unknown, it must be estimated. The estimation of the number of speakers involved in conversation is used as a stop criterion for the clustering phase of the segmentation system. In this case, the number of speakers is estimated using an algorithm based on a penalized BIC criterion.

At first the number of speakers is limited between 1 and 25. The upper limit depends usually on the recording size (e.g.: for 30 minutes audio files the limit is set to 25). Then, we select the number of speakers (NSp) that maximizes:

$$BIC(M)=\log L(X/M)-\lambda\frac{m}{2}NSp\log N_X \qquad (1)$$

where M is the model composed of NSp models, Nx is the total number of speech frames involved, m is a parameter that depends on the complexity of the speaker models and λ is a tuning parameter empirically set to 0.6. The first term is the overall log-likelihood of the data while the second term is used to penalize the complexity of the model. We need the second term because the log-likelihood of the data increases with the number of models (speakers) involved in the calculation of $L(X/M)$.

Let X_i and M_i be the speech data and respectively the model of speaker i. Then we have:

$$L(X/M)=L((X_1,X_2,\ldots,X_{NSp})/(M_1,M_2,\ldots,M_{NSp})) \qquad (2)$$

If we make the hypothesis that the data X_i depends only on the speaker model M_i (a speaker data does not depend on other speaker data and models) it can be shown that:

$$L(X/M)=\prod_i L(X_i/M_i) \qquad (3)$$

The number of speakers estimated this way is usually between the real number of speakers and what we call the *optimal* number of speakers namely the number of speakers that minimizes the segmentation error. We will see in an experiment of section 4 that this optimal number is not always the real number of speakers.

3.2 Training data available for all speakers

In this section we assume that we have training data available for all speakers involved in conversation.

This could be reasonably achieved for meeting and telephone data. Data for speakers involved in a meeting could easily be obtained by asking each participant to present himself at the beginning of the meeting. For telephone speech since there are only two speakers involved, data could be

obtained by manual segmentation of a short part of the conversation.

Using the available data, speaker models are trained for every speaker. The models are derived by MAP (means only) from a background model trained the same way it was trained for clustering (see section 2). Then a decision (see Fig 3) is made for every segment obtained from the BIC-pre-segmentation module, namely the segmentation system becomes in this case:

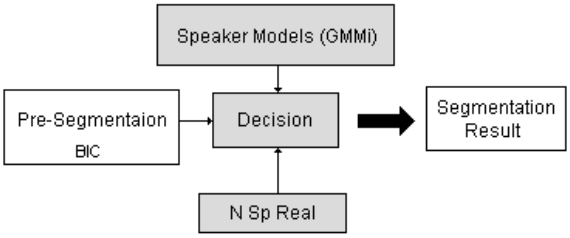


Figure 3: The segmentation system when training data is available for all speakers

The decision is made segment by segment and not frame by frame since experiments have proven that the decision made on an entire segment is more reliable than the decision made on a single frame. A maximum likelihood decision is made for each segment. Given a speech segment X and a speaker model M_i the probability that the segment belongs to that speaker is given by:

$$P(M_i/X)=P(X/M_i)\frac{P(M_i)}{P(X)} \quad (4)$$

Considering that the speakers and the segments are all equal as probability we are actually computing $P(X/M_i)$ for each segment.

However for broadcast news data it is usually impossible to obtain data for all speakers involved in conversation. We can easily obtain data for the news hosts using recordings from the same source (same TV or Radio station) but we could never obtain data for all speakers (e.g: people interviewed that speak for only 30 seconds during a 1 hour news journal).

This is a situation that corresponds to the next case. For the moment we treat only the particular case where training data is available for one of the speakers only.

3.3 Training data available for only one speaker

3.3.1 Pre-Segmenting the audio file

When we have data available for one speaker only, one way to use it is to identify the segments of the known speaker by pre-segmenting the entire recording and segment then the speech not labeled as the known speaker using the conventional segmentation system.

Assuming we have two models available: one for the known speaker M and a universal model for the unknown speakers U and let X be the speech segment data. Then the segment X is labeled to the known speaker if:

$$\log L(X/M)-\log L(X/U)>\alpha \quad (5)$$

where α is a threshold. This pre-segmentation is actually a tracking of the known speakers.

However, this method requires tuning data for the threshold and also the final diarization error may vary with respect to the quantity of training data available for the known speaker.

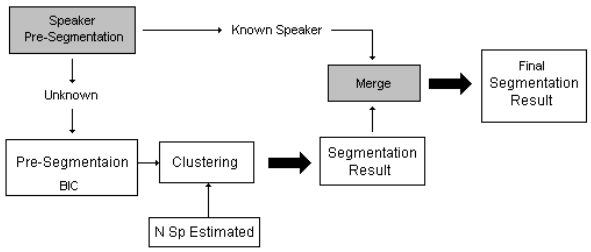


Figure 4 : The segmentation system when data is available for one speaker only (Speaker Pre-Segmentation)

This kind of system (Fig. 4) is actually a cascaded system made of a pre-segmentation and a diarization system. This means that the final system error is the cumulated error of both systems.

3.3.2 Post-Segmentation speaker labeling

There is another way to use the *a priori* known speaker model. We noticed, as stated above, that the estimated number of speakers is usually greater than the optimal number of speakers and smaller than the real number of speakers. The goal in this case would be to decrease the estimated number of speakers in order to get closer to the optimal number of speakers.

Let N_{Sp} be the estimated number of speakers, let X_i be the data labeled to speaker i and let M and U be the models corresponding to the known speaker and the unknown speakers, then we compute the likelihood ratio for all N_{Sp} speakers detected:

$$llr(i)=\log L(X_i/M)-\log L(X_i/U) \quad (6)$$

Using a Bayesian decision, the speaker i that maximizes the $llr(i)$ is matched to the known speaker. This simple decision seems to perform well since the right speaker was well matched to the a priori known model in all the test files used in the experiments of section 4.3.2. Now, suppose the known speaker was split in two or more speaker clusters by the segmentation system. The bad speaker clusters should also obtain a likelihood ratio close to the maximum.

So all speakers clusters that have a positive likelihood ratio and that are close enough to the maximum likelihood ratio should be re-labeled as the known speaker. As a measure of closeness we empirically decided that a speaker cluster is labeled as the known speaker only if its likelihood ratio is superior to a decision threshold fixed to 0.5 of the maximum ratio.

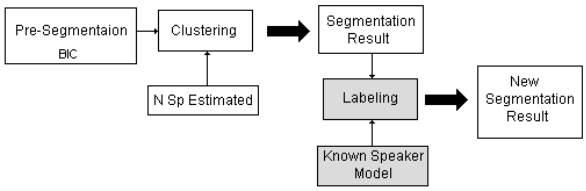


Figure 5: The segmentation system when data is available for one speaker only (Post-Segmentation Labeling)

As shown in Figure 4, some segments of the output segmentation can be re-labeled or not as the known speaker. It is equivalent to a constrained clustering.

4. Experiments And Results

Experiments were conducted on three databases, namely the broadcast news Dry Run and Evaluation database used during the broadcast news NIST RT 2003 Diarization Task and also on the french broadcast news ESTER [12] database. The segmentation error is the official RT 2003 metric used during the spring 2003 campaign.

The RT Dry Run database consisted of 6 news audio files recorded in 1998 from 6 different sources: MNB_NBW, PRI_TWD, NBC_NNW, CNN_HDL, VOA_ENG, ABC_WNT. All files were 10 minutes long, recorded at 16 KHz, 16bit. The number of speakers involved was between 6 and 18.

The RT Evaluation database consisted of 3 news files recorded in 2001 from three different sources: MNB_NBW, PRI_TWD, VOA_ENG. All files were 30 minutes long, recorded at 16 KHz, 16bit. The number of speakers involved was between 10 and 27.

As we can see there are 3 sources that are common to both RT Dry Run and Evaluation databases. Those sources will probably contain speakers that are common to both databases. In experiments of section 4.3, data available from one database will be used as training data for the experiments on the other database.

Finally the ESTER[1] database (the French version of the RT evaluation) see [12] consisted of 40h of audio data recorded from French radio broadcast news shows. It was divided in train data, dev data and test data. A subset was used for the experiment of section 4.3.1. The subset consisted of 21 train files (13h40) from two different sources France Inter, and RFI, 4 dev files (2h40) from the same sources and the 4 test files from the same sources. All files were 1h (RFI) or 20 minutes (France Inter) long, recorded at 16KHz, 16 bit. The number of speakers involved was between 13 and 39.

The performance measure used for the RT 03 segmentation evaluation is an error function based on an optimum one to one mapping of reference speaker IDs to system output speaker IDs. The measure of optimality will be the aggregation, over all reference speakers, of the time that is attributed to both the reference speaker and the corresponding system output speaker to which that reference speaker is mapped. This will always be computed over all speech, including regions of no speech (silence, music, noise, etc).

The measure excludes from scoring the overlapping speech regions. Also some segments, for which transcriptions were not provided (e.g. commercials) using an Unpartitioned Evaluation Map (uem) file, are excluded.

There are two kinds of errors: the speech detection errors and the speaker mapping errors:

- miss speech: a segment was incorrectly labeled as no-speech
- false alarm speech: a segment was incorrectly labeled as speech
- speaker error: a segment was assigned to the wrong speaker

Each audio segment was labeled as a correct diarized segment or as one of the error type segment.

The final score is the fraction of scored time (given by the uem file) that is not correctly labeled. It is the sum of speech error time and of the speaker error time over the scored speech time.

The score tool can be downloaded from the RT web site[2]. The same score is used for the French evaluations ESTER.

4.1 Knowing or not the number of speakers

Our first experiments concerned the estimation of the number of speakers involved in conversation. The following table presents the speaker segmentation error using the optimal, the estimated and respectively the real number of speakers obtained on both speech databases:

Table 1: Estimation of the number of speakers (% diarization error)

Corpus	N Optimal	N Estimated	N Real
Dry Run	14.54	19.65	24.76
Evaluation	14.03	19.25	16.29

The second column is the reference error that we will attempt to decrease using the data available for the speakers.

As stated before, what we call the *optimal* number of speakers is the number of speakers that minimizes the segmentation error and not the real number of speakers. It is usually smaller than the real number of speakers due to the fact that there are a lot of speakers especially in broadcast news data that do not speak enough to train a reliable statistical model (e.g: 4 seconds during a 30 minutes file).

Tests proved that the estimation of the number of speakers generates approximately 5% more absolute segmentation error compared to the optimal number of speakers.

It is however difficult to conclude if it is useful or not to know *a priori* the real number of speakers since the results on Dry Run database show that it is better to estimate the number of speakers and the results on the Evaluation database show that is better to know it *a priori*.

4.2 Training data available for all speakers

For our second experiment we assumed data was available for all speakers.

Different amount of data were available for training starting with 1 second per speaker till 60 seconds per speaker.

The experiment is more suitable for meeting data and certain broadcast news data like a round table discussion but we still used the RT03 Broadcast news data. The speaker data

[1] http://www.afcp-parole.org/ester/

[2] http://www.nist.gov/speech/tests/rt/rt2003/spring/

used to build the speaker models was directly extracted from the test data. This is typically what would be done if a human annotator had labeled a part of the conversation in order to have a few seconds of data for each speaker.

For this experiment, a synthesis of results is presented in the Table 2:

Table 2: Diarization error when data is available for all speakers

A priori data	0 sec	1 sec	2 sec	20 sec	60 sec
Dry Run a priori	19.65	29.70	18.92	13.65	15.95
Dry Run Reference	19.65	19.04	18.52	16.85	19.77
Evaluation a priori	19.25	40.51	37.19	20.82	12.67
Evaluation Reference	19.25	18.95	18.69	17.67	21.33

For comparison purposes we present the segmentation error obtained by the stand alone segmentation system (with no *a priori* models). The diarization error presented is the sum of speech detection error and of speaker error but, since we are using the same pre-segmentation in both cases (with and without training data) the gain or the loss in terms of performance is entirely due to the speaker error change. This reference performance for Dry Run and evaluation data changes slightly for each column of the table since the diarization error is always computed on the parts of speech that were not used to train speaker models. In other words the segments used for training are always excluded from scoring the diarization error. It is also important to note that those segments are excluded from scoring but not from system input.

Finally figures 5 and 6 present all results for different amount of data available from 1 second to 60 seconds.

Figure 6: Diarization error for the Dry Run Database according to the amount of training data available a priori

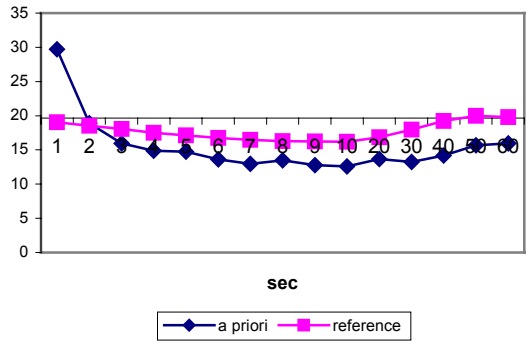
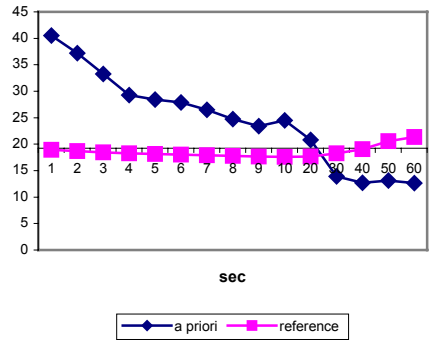


Figure 7: Diarization error for the Evaluation Database according to the amount of training data available a priori



The stand alone reference system segmentation error logically stays constant since it corresponds always to the same reference segmentation output (the slight differences are due to the fact that more and more segments are removed for scoring). On the other hand, the segmentation error decreases as more data is available to learn *a priori* speaker models[3].

There are 3 main differences between the two databases: the size of the recordings (10 minutes versus 30 minutes), the number of speakers involved in conversation and most important the number of speakers that do not have enough data. That is why the improvement is obtained very fast for the Dry Run (starting with 2 seconds data available) and much later for the Evaluation (starting with 20 seconds data available).

For 1 second per speaker of data available we obtain a high error rate since one second is not really enough to train a speaker model. Thus, the improvement actually depends on the amount of data available and on the number of speakers involved in conversation. For 60 seconds available, the error is mainly due to the false alarm speech error and to the speakers that do not have 60 seconds available (we used the maximum available speech in this case but in Eval03 there are speakers that speaks even less than 7 seconds).

In conclusion, as expected if we do not have *sufficient data* for the speakers involved in conversation it is better to use the stand alone segmentation system. However the *sufficient amount of data* varies with the number of speakers involved in conversation.

For all experiments presented in this section, there is also an important aspect to mention, namely the computation time of the segmentation when *a priori* data is available which was about 0.1 x Real Time on a P III (800 MHz) with 512MB physical memory running Red Hat Linux 7.3, compared to 10 x Real Time for the stand alone segmentation system (with no *a priori*).

4.3 Training data available for only one speaker

4.3.1 Pre-Segmenting the audio file

For this experiment we assumed that data was available only for the news speaker. As described in section 3.3.1 the idea was to pre-segment the audio document in known (news host) speaker speech and unknown speaker speech. For

[3] When we are saying we have 30 seconds of available data for example it means at most 30 seconds because there are speakers that do not have such an amount of data.

experimental reasons we preferred to use the ESTER database since it provides enough data splitted in training data, dev (tuning) data and test data.

The speaker models were 128 diagonal GMM directly trained on data extracted from the train corpus using the EM algorithm. The data available for every known speaker was between 20 and 70 minutes.

The unknown speaker models were also 128 diagonal GMM and were directly trained on data not labeled as the known speaker extracted from the train corpus. The unknown models were source dependent (1 unknown model per radio source). The data available for the unknown models was between 150 and 250 minutes.

At first the known / unknown speaker pre-segmentation was tuned on the dev part of ESTER. Then it was applied on the test files. The tuning was done only on the decision threshold from equation 5 but a finer tuning could also be done on speech parametrization and on speaker modeling.

Table 3 presents the experimental results for the dev files while Table 4 presents the results on the test files. For both dev and test corpora the "Reference Diarization Error" (first line) is the diarization error obtained by the CLIPS segmentation system from section 2 without any *a priori* information used. The "Known speaker error" (second line) is the diarization error obtained from the previous segmentation output (without any *a priori* information) by labeling the hypothesis speakers in either 'known' or 'unknown'. It is always inferior to the "Reference Diarization Error" and corresponds to the diarization errors related to the known speaker. When this error value is high, that indicates the potential of pre-segmenting the signal in known / unknown speaker. For instance, for the first dev signal (FrInter1), the "Reference Diarization Error" is 7.31% while the "Known speaker error" is 7.05%. That means that if we had a perfect known speaker tracking for pre-segmentation, then the overall diarization error would be very low.

For the test files the "Known/Unknown Pre-Seg Error" (third line) scores the pre-segmentation phase in known/unknown speakers. The pre-segmentation output is obtained using the threshold from equation 5 tuned on the dev files.

Finally for the test files the "Final Diarization error" (fourth line) is the error obtained by the overall segmentation system that uses a known/unknown speaker pre-segmentation cascaded with the diarization system applied on the speech part labelled as unknown speaker (Figure 4).

Table 3: Training data available for only one speaker: Experimental results for the *dev* files (pre-seg. Approach)

File	FrInt 1	FrInt 2	RFI0930 1	RFI1130 1
Reference Diarization Err	7.31	8.49	14.55	23.47
Known speaker Err	7.05	7.56	1.31	4.27

Table 4: Training data available for only one speaker: Experimental results for the *test* files (pre-seg. Approach)

File	FrInt 3	FrInt 4	RFI0930 2	RFI1130 2
Reference Diarization Err	13.29	6.89	13.67	25.29
Known speaker Err	6.83	5.48	2.24	6.14
Known / unkown Pre-Seg Err	3.84	3.63	5.16	1.67
Final Diarization Err	12.35	5.64	17.42	21.26

The results obtained on the dev files (Table 3) indicates us the potential of using or not the known/unknown pre-segmentation. It shows that for 3 out of 4 files there is a possible final gain.

Comparing the first line (Reference) and the fourth line (Final) of Table 4 show an improvement of the overall diarization performance on 3 out of 4 files. For the file which did not lead to an improvement (RFI0930 2), the "known speaker error" was very low, which means that the diarization error was mainly due to the unknown speaker signal part. In this case, a known/unknown pre-segmentation cascaded with the diarization system just added errors to the system.

4.3.2 Post-Segmentation speaker labeling

For this experiment we also assumed data was available only for the news host speaker (about 15% of the RT Evaluation database). We used data available from the RT Dry Run database for training since it is anterior to the data available in the RT Evaluation database which was used for testing. For the experiment we started with the reference segmentation output and tried to label the detected speakers as the known speaker. We usually had about 2 minutes data available per known speaker that we used to train a 32 diagonal GMM. For the unknown model we used the rest of the corresponding Dry Run recording. The unknown model was also a diagonal 32 GMM.

The reference segmentation was then re-labelled according to the rule presented in section 3.3.2.

The results are presented bellow:

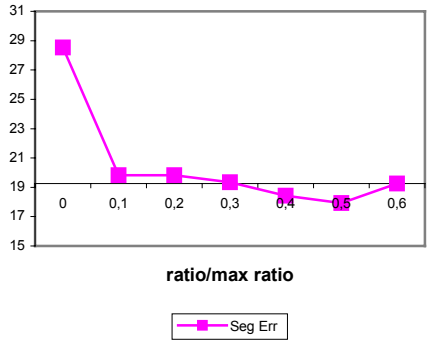
Table 5: Training data available for only one speaker: Experimental results on RT data (post-seg. Approach)

Corpus	**Reference**	**1 known speaker**
Evaluation	19.25	17.92

These preliminary results show the potential of using *a priori* information on one speaker. However these results need more experiments in order to have a reliable conclusion. There is one recording for which the gain is 3.4% in absolute decreasing from 10.14% which actually gives an almost perfectly segmented file.

To show how precarious this method is, we need to show Figure 7 which presents the segmentation error with respect to the threshold equals to $llr(i)$ ratio/maximum $llr(i)$ ratio.

Figure 8: Diarization error when data is available for one speaker only, according to the re-labeling threshold



For comparison if all speakers that have a positive likelihood ratio are labeled as known speakers, meaning that we are not using the condition that the ratio should be greater than 0.5 of the maximum, the segmentation error is 28.53% (it is the case ratio/max ratio 0). This proves that most of the positive ratios are concentrated between 0 and 0.1 of the maximum ratio. The ratios of the known speakers are much greater than the ratio obtained by the speakers not labeled as the known speaker. This way of doing things is equivalent to a constrained clustering.

However the relative small gain, only 1.33% in absolute is due to the fact that the segmentation system already detects well the known speaker (usually the error is about 30 seconds missed speaker time). This means that experiments should be done probably on another database where the possible gain would be greater (more then 30 seconds on a 30 minutes file).

5. Conclusions

In this paper we have investigated the use of *a priori* information for speaker diarization (segmentation).

We started by presenting the stand alone CLIPS diarization system that was used during the NIST Rich Transcription 2003 evaluation.

Then we investigated the possibility of using different *a priori* information in order to improve the stand alone system diarization error. As *a priori* information we were interested in: knowing or not the real number of speakers, using data available for all speakers involved in conversation and using data available for only some of the speakers involved in conversation.

It was difficult to conclude if it is useful or not to know *a priori* the real number of speakers since the results on one of the databases show that it is better to estimate the number of speakers and the results on the other database show that is better to know it *a priori*.

For the case were training data is available for all speakers, experiments have proven as expected that when there is *sufficient data* available for all speakers this data should be used for segmentation. However the amount of *sufficient data* can vary with respect to the number of speakers involved in conversation.

Finally when data was available for only one speaker it was quite difficult to use it and we could only obtain a small diarization error gain. We presented two different methods: one by pre-segmenting the signal in known / unknown speaker and one by post-segmentation speaker labeling.

As a perspective, further experiments should be done on other databases. Experiments on meeting or telephone conversation databases for the case when data is available for all speakers.

6. References

[1] S. Meignier, J.-F. Bonastre, and S. Igounet, “E-HMM approach for learning and adapting sound models for speaker indexing,” In *A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[2] P. Nguyen, J.-C. Junqua, "PSTL's Speaker Diarization system", *DARPA/NIST Rich Transcription Workshop*, Boston, Massachusetts, May 2003

[3] A. G. Adami, S. Kajarekar, H. Hermansky, "A New Speaker Change Detection Method For Two-Speaker Segmentation", *Proc of ICASSP 2002*, Orlando, Florida May 2002

[4] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, "The Elisa Consortium Approaches in Broadcast News Speaker Segmentation During The Nist 2003 Rich Transcription Evaluation", Accepted to *Proc of ICASSP 2004*, Montreal, Canada, May 2004

[5] S.S. Chen, P.S. Gopalakrishnan., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, 127--132, February 1998.

[6] H. Gish, H-H Siu, R. Rohlicek. "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proc of ICASSP 1991*, Toronto, Canada, May 1991

[7] P. Delacourt and C. Wellekens, "DISTBIC: a Speaker-Based Segmentation for Audio Data Indexing", *Speech Communication,* Vol. 32, No. 1-2, September 2000

[8] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA Consortium Research Activities", In *A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[9] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a Block-Segmental Multi-Recognizer Architecture for Automatic Speaker Recognition", *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[10] S. Meignier, D. Moraru, , C. Fredouille, L. Besacier, J.-F. Bonastre, "Benefits of Prior Acoustic Segmentation for Automatic Speaker Segmentation", Accepted to *Proc of ICASSP 2004*, Montreal, Canada, May 2004

[11] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification Using Adaptation Mixture Models”. *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[12] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, K. Choukri, “The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News”, Accepted to *LREC* 2004, Lisbon, Portugal, May 2004

ANNEXE D

Article segmentation en locuteurs conférence A Speaker Odyssey 2004

ELISA Nist RT03 Broadcast News Speaker Diarization Experiments

Daniel Moraru (1), Sylvain Meignier (2),
Corinne Fredouille (2), Laurent Besacier (1), Jean-François Bonastre (2)

[1] CLIPS-IMAG (UJF & CNRS) - BP 53 - 38041 Grenoble Cedex 9 - France
[2] LIA-Avignon - BP1228 - 84911 Avignon Cedex 9 – France

(daniel.moraru,laurent.besacier)@imag.fr
(sylvain.meignier,corinne.fredouille,jean-francois.bonastre)@lia.univ-avignon.fr

ABSTRACT

This paper presents the ELISA consortium activities in automatic speaker diarization (also known as speaker segmentation) during the NIST Rich Transcription (RT) 2003 evaluation. The experiments were achieved on real broadcast news data (HUB4), in the framework of the ELISA consortium. The paper firstly shows the interest of segmentation in acoustic macro classes (like gender or bandwidth) as a front-end processing for segmentation/diarization task. The impact of this prior acoustic segmentation is evaluated in terms of speaker diarization performance. Secondly, two different approaches from CLIPS and LIA laboratories are presented and different possibilities of combining them are investigated. The system submitted as ELISA primary obtained the second lower diarization error rate compared to the other RT03-participant primary systems. Another ELISA system submitted as secondary outperformed the best primary system (i.e. it obtained the lowest speaker diarization error rate).

1. INTRODUCTION

Speaker diarization (or segmentation) is a new speech processing task resulting from the increase in the number of multimedia documents that need to be properly archived and accessed. One key of indexing can be speaker identity. The goal of speaker diarization is to segment a N-speaker audio document in homogeneous parts containing the voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to a same speaker (clustering process). Generally, no *a priori* information is available on the number of speakers involved in the conversation as well as on the identity of the speakers.

The NIST Rich Transcription (RT) Evaluation[1] is sponsored in part of the DARPA Effective Affordable Reusable Speech To Text (EARS) Program. The EARS research effort is dedicated to developing powerful speech transcription technology that provides rich and accurate transcripts. It includes speech transcription but also acoustic segmentation, speaker indexing, disfluency detection induced by spontaneous speech (hesitations, self repairs, word fragments…), etc. Making available this rich transcription will authorize a better job when a machine is detecting, extracting, summarizing, and translating important information. EARS is focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages.

This paper presents the ELISA Consortium [1] activities in automatic speaker segmentation during the NIST RT evaluation campaign organized in 2003.

The first part of this paper focuses on acoustic segmentation, which was mainly introduced to help automatic speech recognition (ASR) systems within the special context of broadcast news transcription. Indeed, at the beginning, one of the main objective of acoustic segmentation was to provide ASR system with an acoustic event classification allowing to discard non speech signal (silence, music, …), to adapt ASR acoustic models to some particular acoustic environments (like speech over music, telephone speech) or simply to speaker gender [2][3][4]. Many papers were dedicated to this particular issue and to the evaluation of acoustic segmentation in the context of ASR task. Nevertheless, rarely discussed in the literature, acoustic segmentation may be useful for other tasks linked to broadcast news corpora. In this sense, the aim of this paper is to investigate the impact of acoustic segmentation when applied as a prior segmentation for the particular task of speaker diarization.

The second part of the paper presents two systems – from CLIPS and LIA laboratories – which exhibit two different segmentation strategies. Various combination schemes of both systems are also investigated (the LIA macro class acoustic segmentation process is combined with the two speaker segmentation systems).

[1] See http://www.nist.gov/speech/tests/rt/rt2003/index.htm for more details

Section 2 presents the LIA macro class acoustic segmentation system, based on a hierarchical strategy. *Section 3* is dedicated to the presentation of the two speaker segmentation approaches involved in this work. Both begin by an acoustic pre-segmentation, also presented in this section. *Section 4* describes the combining strategies. The performance of the various propositions are shown and discussed in *Section 5* (the data are issued from RT 2003 evaluation campaign). Finally, *Section 6* concludes this work and gives some perspectives.

2. ACOUSTIC MACRO CLASS SEGMENTATION

Acoustic macro class segmentation is necessary to discard non speech signal (like music, silence, …) or to adapt acoustic models to specific acoustic environments (telephone speech, speech over music, …). This is especially true when handling broadcast news data with the aim of automatically transcribing speech for instance. Basic speech/non speech detection is also useful for speaker segmentation task in order to avoid music portions to be automatically labeled as a new speaker. However, acoustic segmentation system may be designed to provide finer classification. For example, gender classification could help the segmentation process, by selecting the appropriate a priori knowledge. In this paper, the prior acoustic segmentation is done at three different levels: Speech/Non speech, Clean/Over music/Telephone[2] speech and Male/Female speech.

2.1 Front end processing

The signal is characterized by 39 acoustic features computed every 10 ms on 25 ms Hamming-windowed frames: 12 MFCC augmented by the normalized log-energy, followed by the delta and delta-delta coefficients.

2.2. Hierarchical approach

The system relies on a hierarchical segmentation performed in three successive steps as illustrated in figure 1:

- during the first step, a speech / non speech segmentation of signal (representing a show) is performed using "MixS" and "NS" models. The first model represents all the speech conditions while the second one represents the non speech conditions. Basically, the segmentation process relies on a frame-by-frame best model search. A set of morphological rules are then applied to aggregate frames and label segments.
- during the second step, a segmentation based on 3 classes - clean speech ("S" model), speech over music ("SM" model) and telephone speech ("T" model) is performed only on the speech segments detected by the previous segmentation step. All the models involved during this step are gender-independent. The segmentation process is a Viterbi decoding applied on an ergodic HMM, composed, here, of three states (“S”, “T”, and “SM” models). The transition probabilities of this ergodic HMM are learnt on 1996 HUB 4 broadcast news corpus.
- the last step is devoted to gender detection. According to the label given during the previous step, each segment will be identified as female or male speech by the use of models dependent on both gender and acoustic class. “GT-Fe” and “GT-Ma” models represent female and male telephone speech respectively, “GS-Fe” and “GS-Ma” represent female and male clean speech, while “GSM-Fe” and “GSM-Ma” represent female and male speech over music. Two other models, “GDS-Fe” and “GDS-Ma”, representing female and male speech recorded over degraded conditions are also used, at this stage, to refine the final segmentation. The segmentation process, described in the previous step, is applied in the same way here.

All the state models mentioned above are diagonal GMMs [5] except NS and MixS models which are characterized by 1 and 512 Gaussian components respectively, all the other models are characterized by 1024 Gaussian components. They were trained on the 1996 HUB 4 broadcast news corpus.

•

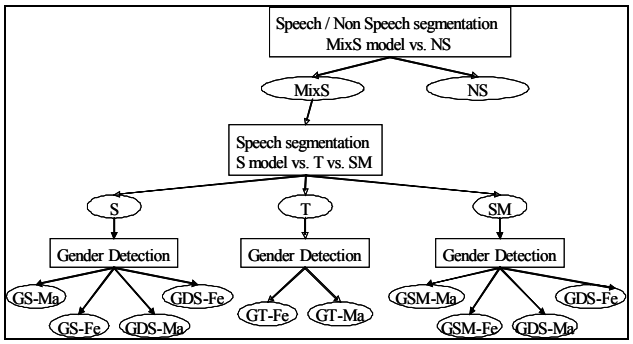


Figure 1: Hierarchical acoustic segmentation

3. SPEAKER SEGMENTATION SYSTEMS

All the speaker segmentation systems were developed in the framework of the ELISA consortium using AMIRAL, the LIA Speaker Recognition System [6].

Two different speaker segmentation systems are presented in this section. They have been developed individually by the CLIPS and LIA laboratories. Basically, the CLIPS system relies on a BIC-detector-based strategy followed by an hierarchical clustering [7]. The LIA system shows a different strategy, based on a HMM modeling of the conversation and an iterative process which adds the speakers one by one. Both of them use the acoustic pre-segmentation - described in section 2 - as a preliminary phase.

3.1. The LIA System

The LIA system is based on Hidden Markov Modeling (HMM) of the conversation [8][9]. Each state of the HMM characterizes a speaker and the transitions model the changes between speakers.

The speaker segmentation system is applied separately on each of the four acoustic classes detected by the acoustic segmentation described in section 2. Finally, the four segmentations are merged and a re-segmentation process is applied.

During the segmentation, the HMM is generated using an iterative process, which detects and adds a new state (i.e. a new speaker) at each iteration. The speaker detection process is composed of four steps:

- *Step 1-Initialization*. A first “speaker” model is trained on the whole test utterance (it is more a generic acoustic model than a given speaker model). The conversation is modeled by a one-state HMM and the whole signal is set to the initial “speaker”.
- *Step 2-Adding a new speaker*. A new speaker model is trained using 3 seconds of test speech that maximize the likelihood ratio computed using the first model and a world model (learned using development data). A corresponding state is added to the previous HMM.

[2] This segmentation may also be referred to as a narrow/wide band speech classification if “speech over music” label is not used and simply considered as “speech” label.

- *Step 3-Adapting speaker models*. First, all the speaker models are adapted, using a MAP approach, according to the current segmentation. Then, a Viterbi decoding is done and produces a new segmentation. The adaptation and decoding steps are performed while the segmentation differs between two successive "adaptation/decoding" phases.
- *Step 4-Assessing the stop criterion*. The likelihood of the previous solution and the likelihood of the last solution are computed using the last HMM model (for example, the solution with two speakers detected and the solution with three speakers detected). The stop criterion is reached when no gain in terms of likelihood is observed [8] or when no more speech is left to initialize a new speaker. A heuristic criterion is added to the likelihood-based criterion: if the last added speaker is tied to only one segment (<4sec), the previous segmentation is kept and a new speaker is added using the second best segment from *Step2*.

When the four (sub) segmentations are obtained independently using the previously described algorithm, they are merged and a re-segmentation phase starts.

The re-segmentation is similar to the adaptation and decoding step *(Step 3*). The main difference between the two phases is the GMM adaptation algorithm. The both adaptation algorithms are variants of MAP Bayesian adaptation. A classical MIT MAP adaptation [5] is used for the re-segmentation phase while a LIA variant optimized for an adaptation on a very short segment [6] is performed during the (sub)segmentation phase. In both adaptation phases only means are adapted.

The signal is characterized by 20 linear Cepstral features (LFCC) computed every 10 ms using a 20ms window. The Cepstral features are augmented by the energy (E). No frame removal or any coefficient normalization is applied. GMM with 128 components (diagonal covariance matrix) are used for the speakers and world/background models. The background models are trained on a subset of Broadcast News 96 data (F0, F1 and F2 acoustic conditions).

The LIA also presented a secondary system[3] closed to the previous one but using another variant of MAP for the speaker model adaptation. This algorithm is based on a linear combination of the estimated data and the a priori information. This adaptation method was employed by the LIA during NIST 2002 speaker recognition evaluation [10].

3.2. The CLIPS System

The CLIPS system [10] is based on a BIC (Bayesian Information Criterion) speaker change detector followed by an hierarchical clustering. The clustering stop condition is the estimation of the number of speakers using a penalized BIC criterion.

The speaker segmentation system is applied separately on each of the four acoustic classes detected by the acoustic segmentation described in section 2. A BIC [7] approach is then used to define first potential speaker changes. A BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-Gaussian models with diagonal covariance matrices are used to model the two windows. A threshold is then applied on the BIC curve to find the most likely speaker change points which correspond to the local maximums of the curve.

[3] This system is used in section 3.2 ("fusion" system).

Clustering starts by first training a 32 components GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion thanks to a classical EM algorithm. Segments models are then trained using MAP adaptation of the background model (means only). Next, BIC distances are computed between segment models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to the N speakers in the conversation).

The number of speakers in the conversation (NSp) is estimated using a penalized BIC (Bayesian Information Criterion), in contrast with the last year CLIPS segmentation system which used a fixed number of speakers [10].

The number of speakers is constrained between 1 (if we are working on an isolated acoustic pre-segmentation class) or 2 (if we are working on the entire audio file) and 25. The upper limit is related to the recording duration. The number of speakers (NSp) is selected to maximize:

$$BIC(M)=\log L(X;M)-\lambda\frac{m}{2}NSp\log NX$$

where M is the model composed of the NSp speaker models, NX is the total number of speech frames involved, m is a parameter that depends on the complexity of the speaker models and λ is a tuning parameter equal to 0.6.

The signal is characterized by 16 mel Cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then the Cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied.

4. COMBINING STRATEGIES

In this section we investigate two possibilities for combining the systems, firstly using an hybridization strategy and secondly by merging the proposed segmentations. The merging strategy is a new way of combining results coming from multiple segmentation systems.

4.1 Hybridization ("piped" system)

The idea of this hybridization strategy is to use the results of the CLIPS system to initialize the LIA re-segmentation system (figure 2). The speakers detected by the CLIPS system (number of speakers and associated audio segments) are inserted in the re-segmentation HMM model (the models are trained using the information issued by the clustering phase). This solution associates the advantages of longer and (quite) pure segments with the HMM modeling and decoding power.

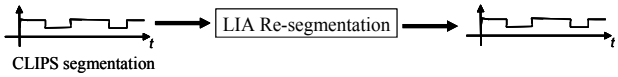


Figure 2: ELISA piped system

4.2 Merging Strategy ("fusion" system)

The idea of "fusion" is to use the segmentations issued from as many as possible experts, four in this paper (figure 3): CLIPS primary system, LIA primary system, LIA secondary system, ELISA piped system.

The merging strategy relies on a frame based decision which consists in grouping the labels proposed by each of the four systems at the frame level. An example (for four systems) is illustrated below:

- Frame *i*: Sys1="S1", Sys2="T4", Sys3="S1", Sys4="F1"
  → label result "S1T4S1F1",
- Frame *i+1* : Sys1="S2", Sys2="T4", Sys3="S1", Sys4="F1"
  → label result "S2T4S1F1".

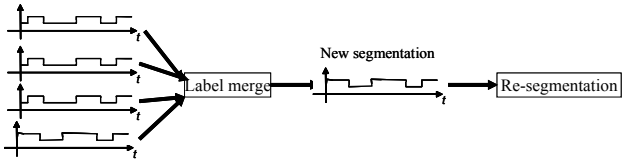


Figure 3: ELISA merge system

This label merging method generates (before re-segmentation) a large set of virtual speakers (~150 virtual speakers per show) composed of:

- Virtual speakers that have a large amount of data assigned. These speakers could be considered as correct hypothesis speakers;
- Virtual speakers generated by few systems, for example the speakers associated with only one short segment (~3s up to 10s). These hypothesis speakers could be suppressed (the weight of these speakers on the final scoring is marginal);
- Virtual speakers that have a smaller amount of data scattered between multiple small segments and that could be considered as zones of indecision.

Based on these considerations, the LIA re-segmentation is then applied on the merged segmentation. Between each adaptation / decoding phase, the virtual speakers for whom total time is shorter than 3s are deleted. The data of these deleted speakers will further be dispatched between the remaining speakers during the next adaptation / decoding phase.

After the first iteration the number of speakers is already drastically reduced (from 150 to about 50) since speakers associated with indecision zones do not catch any data during the Viterbi decoding and are automatically removed.

However, the merging strategy cannot generally solve the wrong behavior of initial systems that could split a “true” speaker in two hypothesis speakers, each tied to a long segment. Suppose all systems agreed on a long segment except one which splits it in two parts. This would produce two virtual speakers (associated with long duration segments) after the merging phase and since we are not doing any clustering before re-segmentation, we would have a "true" speaker spitted in two virtual speakers.

5. EXPERIMENTS AND RESULTS

The experiments described in this paper were conducted in the framework of NIST/RT’03[4] evaluation campaign. In this context, two separate corpora were available:

- *Dev* corpus composed of 6 broadcast news shows of 10mn each, recorded in 1998, available to tune the speaker segmentation systems;
- *Eva* corpus composed of 3 broadcast news shows of 30mn each, recorded in 2001, each containing between 10 and 27 speakers and available for evaluation only.

5.1 Speaker Diarization performance measure

In order to measure the accuracy of the segmentation, we used the scoring system proposed during RT’03 evaluation, a scoring based on the percentage of frame well classified.

Firstly, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed. The measure of optimality is the aggregation, over all reference speakers, of time that is jointly attributed to both the reference speaker and the (corresponding) system output speaker to which that reference speaker is mapped. The resulting time-based speaker diarization error score corresponds to the fraction of speaker time that is not attributed correctly to a speaker. This scoring takes into account both speaker segmentation error (speech segments attributed to a wrong speaker) and missed/false alarm speaker errors, directly linked to speech/non speech classification errors.

5.2. Raw performance of the prior acoustic segmentation

Table 1 gives the performance of the prior acoustic segmentation process, taken alone. The speech/non speech segmentation error is around 4.5% (in terms of duration) compared to 4.4% for the best system during NIST RT’03 evaluation campaign. The gender detection error goes from 1.5% for the *Dev* set to 5.5% for the *Eva* set. Thanks to the manual segmentation Hand S/NS-Gender-T/NT, the accuracy of acoustic segmentation system at the level of telephone and non telephone classification is evaluated: less than 0.1% for *Dev* corpus and 3% for *Eva*.

Corpus	Missed Speech Error	False Alarm Speech Error	Gender Error	Telephone / Non telephone Speech error
Dev	**2.3%**	2.2%	1.5%	0.09 %
Eva	1.8%	2.7%	5.5%	3 %

Table 1: Acoustic segmentation errors on Dev and Eva sets.

5.3. Impact of acoustic macro class segmentation on speaker diarization performance

For evaluating the impact of acoustic macro class segmentation on speaker diarization performance, different levels of acoustic segmentation granularity are evaluated on both speaker segmentation systems:

- Speech/non speech classification only (S/NS);
- segmentation based on speech/non speech and gender detection (S/NS-Gender);
- segmentation based on speech/non speech, gender and telephone/non telephone speech detection (S/NS-Gender-T/NT);
- segmentation based on speech/non speech, gender and telephone/clean speech/speech over music/degraded speech (S/NS-Gender-T/S/MS/DS).

For comparison, some speaker segmentation results will be also presented based on an acoustic segmentation, obtained manually and based on speech/non speech, gender and telephone / non telephone speech detection (Hand S/NS-Gender-T/NT).

For both corpora, the advertisement portions were manually discarded before any treatment which explains that results presented in this section do not correspond exactly to official RT’03 results and to ones presented in the next section.

Table 2 and table 3 present the experimental results obtained when combining speaker segmentation systems (speaker turn detection based and ascending/HMM based approaches) with acoustic segmentation before and after the re-segmentation step respectively. These results, expressed in terms of diarization error rates, show :

- a large variation in terms of performance between *Dev* and *Eva* corpora, especially for the ascending/HMM approach;
- a gain in performance (the best one) for both speaker segmentation systems when they are combined with manual acoustic segmentation;

[4] More information about RT03 Evaluation could be find at: http://www.nist.gov/speech/tests/rt/index.htm

- a significant improvement of ascending/HMM approach results on *Eva* corpus due to speech/non speech, gender and telephone/non telephone segmentations (S/NS-Gender-T/NT) without (from 26.9% to 18.1%) and with (from 26.6% to 14.1%) re-segmentation phase. On *Dev* corpus, this improvement is visible only after re-segmentation phase (from 15.5% to 12.8%);
- no real effect on speaker-turn-based approach regarding the different levels of acoustic segmentation granularity, except on *Eva* corpus after re-segmentation step for which result improvement can be observed (from 15.7 to 13.9%);
- no improvement or even performance loss when the finest acoustic segmentation (S/NS-Gender-T/S/MS/DS) was involved;
- the interest of the re-segmentation step for both speaker segmentation strategies since best performance may be observed in most of the cases.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*13.1%*	*15.4%*	*14.1%*	*10.6%*
S/NS	15.4%	26.9%	19.7%	17.1%
S/NS-Gender	14.8%	27.4%	18.7%	18.6%
S/NS-Gender-T/NT	15.1%	18.1%	19.0%	18.2%
S/NS-Gender-T/S/MS/DS	25.9%	30.5%	19.1%	27.6%

Table 2: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets before re-segmentation step.

	Ascend./HMM approach		Speaker turn approach	
Acoustic segmentation	***Dev***	***Eva***	***Dev***	***Eva***
Hand S/NS-Gender-T/NT	*10.8%*	*13.2%*	*15.7%*	*9.4%*
S/NS	15.5%	26.6%	19.3%	15.7%
S/NS-Gender	13%	24.9%	18.2%	15.3%
S/NS-Gender-T/NT	12.8%	14.1%	19.0%	13.9%
S/NS-Gender-T/S/MS/DS	15.6%	14.3%	18.7%	15.1%

Table 3: Diarization error rates for different levels of acoustic granularity on Dev and Eva sets after re-segmentation step.

	Miss Speech	FA Speech	SPK ERR	**ERR**
CLIPS primary	2.0%	2.9%	14.3%	**19.25%**
LIA primary	1.1%	3.8%	12.0%	**16.90%**
LIA second	1.1%	3.8%	19.8%	**24.71%**
ELISA "merged"	1.1%	3.8%	9.3%	**14.24%**
ELISA "piped"	1.1%	3.8%	8.0%	**12.88%**

Table 4 Experimental results on RT 2003 data

5.4. Performance of ELISA systems during RT03 evaluation

The fusion system submitted as ELISA primary system obtained the second lower segmentation error rate compared to the other RT03-participant primary systems. The ELISA pipe system submitted as secondary system outperformed the best primary system and obtained the lowest speaker segmentation error rate.

The table 4 summarizes the performance achieved by the different proposed systems during RT03. It shows that:

- Even if the five systems are based on the same acoustic segmentation, the *Miss Speech* and *False Alarm Speech* errors are different. This is due to the LIA and ELISA system behavior which work at 0.2s block level (all the segments boundaries are aligned on a 0.2s scale) whereas CLIPS system works at a frame level. It gives small differences in the border positions of the segments but the sum (Miss+FA) remains the same.
- The LIA and CLIPS systems obtained satisfactory results, compared to the other RT03 participants. The LIA HMM based primary system outperforms slightly the CLIPS classical approach (16,9% of total error compared to 19,25%). But the second LIA system - with a different model adaptation strategy - obtained only 24,71% of total error This result illustrates the difficulty of adapting a large statistical model in borderline conditions (only few seconds of adaptation data).
- The "piped" technique improves the performance. Giving good segment boundaries to the HMM based method increases drastically the performance (from 16,9% to 12,88% of total error). Indeed the re-segmentation phase improves the accuracy of the CLIPS segmentation and allows to reduce the segmentation error by 33% (relative).
- The "merged" strategy performs better than the "piped" strategy over two recordings (8% relative gain). Unfortunately a drastic loss is observed on the last recording. The lose on that particular recording is a good example of the limitation of the merging technique explained in 4.2: one of the systems disagreed with the others. This resulted in too many speakers detected and, most important, in a long speaker split in two that generated an important error[5].

For the CLIPS system, complementary experiments showed that estimating automatically the number of speakers during the clustering process generates only about 3% more of absolute segmentation error than the optimal number of speakers[6]. The CLIPS algorithm missed only 7% of the real speakers involved in the files (4 speakers out of 57 total speakers).

6. CONCLUSIONS

This paper summarizes the ELISA Consortium strategies for the speaker segmentation task. The ELISA effort was focused in the framework of NIST 2003 speaker diarization evaluation campaign.

This paper firstly investigated the impact of *prior* acoustic macro class segmentation when it is combined with speaker segmentation. This impact was evaluated in terms of speaker diarization error rate, according to NIST/RT'03 scoring. The *prior* acoustic macro class segmentation, presented in this paper, was developed by the LIA lab. It relies on a GMM model based hierarchical segmentation, designed to provide different levels of segmentation granularity (from simple speech/non speech detection to

[5] The problem could also come from the nature of the test file: it is the only one narrow band file.

[6] The optimal number of speakers is the number of speakers that minimizes the segmentation error and not the real number of speakers involved in dialogue. Usually the optimal number is smaller than the real number. This is due to the fact that in the conversations some speakers pronounced only very short utterances and missing them does not have a significant effect on the total diarization error rate.

gender dependent acoustic classes such as speech over music, degraded speech or telephone speech). Experiments based on the combination of each speaker segmentation system with the acoustic segmentation were conducted according to the different levels of acoustic segmentation granularity. The results presented in this paper show the benefit of acoustic segmentation for speaker segmentation performance, especially for the ascending/HMM approach, while the speaker turn detection based system seems to be less dependent of an acoustic segmentation.

Then, we described two approaches for speaker diarization, the LIA system, based on a HMM modeling of each conversation (where all the information is reevaluated at each detection of a new speaker or a new segment), and the CLIPS system, which uses a standard approach based on speaker turn detection and clustering. Despite the differences between the approaches, the results obtained during the NIST RT03 evaluation showed the interest of each technique.

Several ways of combining the two systems were also proposed. The "piped" system improved significantly the performance, up to 33% of relative error reduction (from 19.25 % to 12.88%) and achieved the best performance during RT03 evaluation. A complete analysis of the results is necessary, to understand which part of the gain comes from the various ways of processing the information and which part comes from the correction of each system intrinsic errors.

One of the main drawback of both systems is the difficulty to detect the minority speakers that do not speak very much. Depending of the nature of the audio files, they could generate a large part of the segmentation errors.

As a perspective, we are currently working on adding to the conversation model a priori information for the segmentation system [11], for both ELISA approaches. Further work should also study the way of taking benefit of finer acoustic classes such as speech over music or degraded speech for the speaker segmentation task.

REFERENCES

[1] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA consortium research activities," *A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[2] P.C. Woodland, "The development of the HTK Broadcast News transcription system: An overview", *Speech Communication*, Vol. 37, pp. 291-299, 2002.

[3] T. Hain, and P.C. Woodland, "Segmentation and Classification of Broadcast News audio", *ICSLP'98*, Sydney, Australia.

[4] J.L. Gauvain, L. Lamel, and G. Adda. "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[5] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adaptation Mixture Models". *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[6] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a block-segmental multi-recognizer architecture for automatic speaker recognition," Digital Signal Processing, Vol. 10, No. 1-3, January/April/July 2000.

[7] P. Delacourt and C. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication,* Vol. 32, No. 1-2, September 2000.

[8] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," *A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[9] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for Multi-Speaker Tracking System". *ICASSP'00*, 5-9 June 2000, Istanbul, Turkey.

[10] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation". *ICASSP'03*, Hong Kong.

[11] D. Moraru, L. Besacier, E. Castelli, « Using a priori information for speaker diarization », Other Paper Submitted to Odyssey 2004.

Liste des abréviations

BBN	Société américaine Bolt Beranek and Newman
BIC	Bayesian Information Criterion, critère d'information Bayesien
CLR	Cross Likelihood Ratio, rapport de vraisemblance croisé
CMS	Cepstral Mean Substraction, soustraction de la moyenne cepstrale
DARPA	Defense Advanced Research Projects Agency, agence de recherche du ministère de la défense américain
DCT	Discrete Cosine Transform, transformée discrète en cosinus
EM	Espectation Maximization, algorithme d'apprentissage d'un modèle statistique
ESTER	Evaluation des Systèmes de Transcription Enrichie des émissions Radiophoniques
F0	Fréquence fondamentale ou fréquence d'excitation laryngienne
FFT	Fast Fourier Transform, transformée de Fourier rapide
GLR	Generalized Likelihood Ratio, rapport de vraisemblance généralisé
GMM	Gaussian Mixture Model, modèle de mélange de gaussiennes
HMM	Hidden Markov Model, modèle de Markov caché
LFCC	Linear Frequency Cepstral Coefficients, coefficients cepstraux utilisant un banc de filtres repartis sur une échelle linéaire
LIA	Laboratoire Informatique d'Avignon
LPCC	Linear Prediction Cepstral Coefficients, coefficients cepstraux calculés à partir d'une analyse par prédiction linéaire
MAP	Maximum A Posteriori, adaptation d'un modèle par maximum *a posteriori*
MFCC	Mel Frequency Cepstral Coefficients, coefficients cepstraux utilisant un banc de filtres repartis sur une échelle Mel
MIT	Massachusetts Institute of Technology
MLLR	Maximum Likelihood Linear Regression, adaptation d'un modèle par régression linéaire
NIST	National Institut of Standards and Technologies
P	Précision, en recherche d'information, la quantité d'information pertinente trouvée divisée par la quantité d'information trouvée
PM	Précision Moyenne, en recherche d'information, la moyenne des précisions calculées après chaque document trouvé, l'information doit être ordonnée selon une mesure de confiance

R	Rappel en recherche d'information, la quantité d'information pertinente trouvée divisée par la quantité d'information pertinente à trouver
RT	Rich Transcription, transcription riche
TDM	Taux de Détections Manquées
TFA	Taux de Fausses Alarmes
TREC	Text REtrieval Conference

Bibliographie

[Adami 2002] Adami A., Kajarekar S., Hermansky H., "A New Speaker Change Detection Method for Two-Speaker Segmentation", dans Proc of ICASSP 2002, Orlando, Mai 2002

[Ajmera 2003] Ajmera J., Wooters C., Peskin B., Oei C., "Speaker Segmentation and Clustering", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Amir 2003] Amir A., Berg M., Chang S.-F., Hsu W., Iyengar G., et al. "IBM Research TRECVID 2003 Video Retrieval System", 12th Text Retrieval Conference, Gaithersburg, MD, USA, 17-18 Novembre, 2003

[André-Obrecht 1988] André-Obrecht R., "A New Statistical Approach for The Automatic Segmentation of Continuous Speech Signals", IEEE Transactions ASSP, vol. 36, no. 1, pages 29-40, Janvier 1988

[Atal 1976] Atal B., "Automatic Recognition of Speakers from Their Voices", IEEE Transactions, vol 64(4), pages 460-475, 1976

[Baker 1975] Baker J., "The Dragon System An Overview", dans IEEE Transactions on Acoustic Speech and Signal Processing, vol. 23, pages 24-29, Février 1975

[Baum 1970] Baum L., Petrie T., Soules G., Weiss N., "A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Functions of Markov Chains", Ann. Math Statistics, vol. 41,no. 1, pages 164-171, 1971

[Beal 2002] Beal M., Nebojsa J., Attias H., "A Self-Calibrating Algorithm for Speaker Tracking Based On Audio-Visual Statistical Models", Proc of ICASSP 2002, Orlando, Mai 2002

[Benois-Pineau 2002] Benois-Pineau J., Desainte-Catherine M., Louis N., "A method for extraction of audio-visual leitmotif in movies by cross media analysis", Proc of EUSIPCO 2002, Toulouse, France, Septembre 2002

[Berrut 1997] Berrut C., *"Indexation de données multimédia, utilisation dans le cadre d'un système de recherche d'information"*, Habilitation à diriger des recherches, Université Joseph Fourier 1997

[Besacier 1998] Besacier L., Bonastre J. F., "Frame Pruning for Speaker Recognition", Proc of ICASSP 1998, Seattle, Mai 1998

[Besacier 2004] Besacier L., Quénot G., Ayache S., Moraru D., "Video Story Segmentation With Multi-Modal Features: Experiments on TRECVID 2003", Proc of ACM Multimedia Information Retrieval, New York, Etats-Unis, Octobre 2004

[Bilmes 1998] Bilmes J., "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", Rapport Technique, U.C Berkley, April 1998

[Bimbot 1993] Bimbot F., Mathan L., "Text-Free Speaker Recognition Using an Arithmetic Harmonic Sphericity Measure", Proc Eurospeech 1993, Vol 1., pages 169-172, Berlin, Allemagne, Septembre 1993

[Bonastre 2000] Bonastre J.-F., Delacourt P., Fredouille C., Meignier S., Merlin T., Wellekens C. J., "Differentes strategies pour le suivi du locuteur", Reconnaissance de Formes et Intelligence Artificielles (RFIA), pages 123-129, 2000

[Brehélin 2000] Bréhélin L., Gascuel O., "Modèles de Markov cachés et apprentissage de séquences", dans Le temps, l'espace et l'évolutif en sciences du traitement de l'information, pages 407-421, Eds. Cépaduès, 2000.

[Calliope 1989] Calliope, *"La parole et son traitement automatique"*, Masson et CENT-ENST, Paris, 1989, ISBN 2-225-81516-X

[Carbonell 1999] Carbonell J., Yang Y., Lafferty J., Brown R., Pierce T., Liu X, "CMU Report on TDT-2: Segmentation, Detection and Tracking", Proc of DARPA Broadcast News Workshop, Herndon, Etats-Unis, Mars 1999

[Carey 1999] Carey M. J., Parris E. S., Lloyd-Thomas H., "A Comparison of Features for Speech, Music Discrimination", Proc of ICASSP 1999, Phoenix, Mars 1999

[Cettolo 2001] Cettolo M., "Speaker Tracking in A Broadcast News Corpus", A Speaker Odyssey 2001, Crete, Juin 2001

[Chaisorn 2003] Chaisorn L., Koh C., Zhao Y., Xu H., Chua T.-S., Qi T., "Two-Level Multi-modal Framework for News Story Segmentation of Large Video Corpus", 12th Text Retrieval Conference, Gaithersburg, MD, USA, 2003.

[Chen 1998] Chen S., Gopalakrishnan P., "Speaker, Environment and Channel Change Detection via The Bayesian Information Criterion", DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Février 1998

[Cooper 2003] Cooper M., Foote J., Adcock J., Casi S., "Shot Boundary Detection via Similarity Analysis", Proc of TREC 2003 Workshop, Gaithersburg, MD, USA, Novembre 2003

[Delacourt 2000a] Delacourt P., *"Indexation de données audio: segmentation et regroupement par locuteurs"*, Thèse de doctorat, 182 pages, ENST-Eurecom 2000

[Delacourt 2000b] Delacourt P., Wellekens C.J., "DISTBIC: A Speaker Based Segmentation for Audio Data Indexing", Speech Communication 32, pages 111-126, 2000

[Dempster 1977] Dempster A.P., Laird N. M., Rubin D.B., "Maximum Likelihood from Incomplete Data via The EM Algorithm", Journal of Royal Statistical Society, B(39), no1, pages 1-38, 1977

[Duda 2001] Duda R., Hart P., Stork D., *"Pattern Classification"*, John Wiley & Sons, Second Edition, 2001, ISBN 0-417-05699-3

[Dugad 1996] Dugad R., Desai U., "A Tutorial on Hidden Markov Models", Rapport Technique, Indian Institute of Technology Bombay, May 1996

[Everitt 1993] Everitt B., *"Cluster Analysis"*, Oxford University Press Inc., New York; Troisième edition, 1993, ISBN 0-340-76119-9

[Fredouille 2000] Fredouille C., Bonastre J.F., Merlin T., "AMIRAL: A Block Segmental Multi-Recognizer Architecture for Automatic Speaker Recognition", Digital Signal Processing, Special Issue on NIST 1999 Speaker Recognition Workshop 10(1-3), 2000

[Fredouille 2004] Fredouille C., Moraru D., Meignier S., Besacier L., Bonastre J. F., "The NIST 2004 Spring Rich Transcription Evaluation: Two-Axis Merging Strategy in The Context of Multiple Distant Microphone Based Meeting Speaker Segmentation", Proc of RT 04 Meeting Workshop, Montreal, Canada, Mai 2004

[Freitag 1999] Freitag D., McCallum A., "Information Extraction with HMM and Shrinkage", Proc of AAAI 99 Workshop on ML for Information Extraction, pages 31-36, 1999

[Fukunaga 1990] Fukunaga K., *"Introduction to Statistical Pattern Recognition"*, Academic Press Inc., 1990, ISBN 0-122-69851-7

[Furui 1981] Furui S., "Cepstral Analysis Technique for Speaker Verification", IEEE Transactions on Acoustic Speech and Signal Processing, vol 29(2), pages 254-272, 1981

[Furui 1994] Furui S., "An Overview of Speaker Recognition Technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994

[Gauvain 1994] Gauvain J.-L., Lee C.-H., "Maximum *A Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Transactions on Speech Audio Processing, vol 2(2), pages 291-298, avril 1994

[Gauvain 2002] Gauvain J.L., Lamel L., Adda. G., "The LIMSI Broadcast News Transcription System", Speech Communication, 37(1-2) 89-108, 2002. ftp://tlp.limsi.fr/public/spcH4_limsi.ps.Z

[Gauvain 2003] Gauvain J.L., Barras C., "Speaker Diarization", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Gish 1991] Gish H, Siu M.-H., Rohlicek R.., "Segregation of Speakers for Speech Recognition and Speaker Identification", Proc of ICASSP 1991, Toronto, Mai 1991

[Godfrey 1992] Godfrey J., "SWITCHBOARD: Telephone Speech Corpus for Research and Development", Proc of ICASSP 1992, pages 517-520, San Francisco, Mars 1992

[Gravier 2004] Gravier G., Bonastre J.-F., Galliano S., Geoffrois E., Mc Tait K. and Choukri K., "The ESTER Evaluation Campaign of Rich Transcription of French Broadcast News", Proc. Language Resources and Evaluation Conference 2004, Lisbone, Portugal, Mai 2004.

[Hain 1998] Hain T., Woodland P., "Segmentation and Classification of Broadcast News Audio", Proc of ICSLP 98, Sydney, Decembre 1998

[Hjelsvold 1994] Hjelsvold R., Midstraum R., "Modeling and Querying Video Data" Proc of 12th Very Large Database Conference, Santiago, Chili, Septembre 1994

[Hsu 2004] Hsu W., Kennedy L., Huang C.-W., Chang S.-F., Lin C.-Y., Iyengar G., "News Video Story Segmentation Using Fusion of Multi-level Multi-modal Features in TRECVID 2003", Proc of ICASSP 2004, Montreal, Canada, Mai 2004

[Huang 2003] Huang X., Wei G., Petrushin V., "Shot Boundary Detection and High-Level Features Extraction for The TREC Video Evaluation 2003", Proc of TREC 2003 Workshop, Gaithersburg, MD, USA, Novembre 2003

[Jelinek 1976] Jelinek F., "Continuous Speech Recognition by Statistical Methods", Proc of IEEE, vol. 64, no. 4, pp. 532-556, Avril 1976

[Jin 2004] Jin Q., Laskowski K., Schultz T., Waibel A., "Speaker Segmentation and Clustering in Meetings", Proc of RT 04 Meeting Workshop, Montreal, Canada, Mai 2004

[Johnson 1997] Johnson S., *"Speaker Tracking"*, Master thesis, University of Cambridge, 1997

[Kemp 2000] Kemp T., Schmidt M., Westphal M., Waibel A., "Strategies for Automatic Segmentation of Audio Data", Proc of ICASSP 2000, Istambul, Juin 2000

[Kieseppa 1997] Kieseppa I., "Akaike Information Criterion , Curve-Fitting and the Philosophical Problem of Simplicity", British Journal for The Philosophy of Science vol 48(1), pages 21-48, Mars 1997

[Lamy 2003] Lamy R., Moraru D., Bigi B., Besacier L., "Premiers pas du CLIPS sur les données d'évaluation ESTER", Journées d'Etude sur la Parole 2004, Fès, Maroc, Avril 2004

[LDC 1998] LDC, Linguistic Data Consortium, http://www.ldc.upenn.edu 1998

[Leggetter 1995] Leggetter C., Woodland P. C., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, vol. 9, pages 171-185, 1995

[Linde 1980] Linde Y., Buzo A., Gray R.M., "An Algorithm for Vector Quantizer Design", IEEE Transaction on Communication, vol 28., Janvier 1980

[Liu 2002] Liu M., Chen T., Ma C., Li X., Chang E., "MSRA NIST 2002 SRE System", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[Liu 2004] Liu D., Kubala F., "Online Speaker Clustering", Proc of ICASSP 2004, Montreal, Mai 2004

[Magrin-Chagnolleau 2001] Magrin-Chagnolleau I., Gravier G., Blouet R., "Overview of ELISA Consortium Research Activities", A Speaker Odyssey 2001, Crete, Juin 2001

[Martin 2000] Martin A., Przybocki M., "The NIST 1999 Speaker Recognition Evaluation – An Overview", Digital Signal Processing, Special Issue on NIST 1999 Speaker Recognition Workshop 10(1-3), 2000

[Martin 2001] Martin A., Przybocki M., "The NIST Speaker Recognition Evaluations: 1996-2001", A Speaker Odyssey 2001, Crete, Juin 2001

[Meignier 2000] Meignier S., Bonastre J.-F., Fredouille C., Merlin T., "Evolutive HMM for Speaker Tracking System", Proc of ICASSP 2000, Istambul, Juin 2000

[Meignier 2001] Meignier S., Bonastre J. F., Igounet S., "E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing", A Speaker Odyssey 2001, Crete, Juin 2001

[Meignier 2002] Meignier S., Bonastre J.-F., Magrin-Chagnolleau I., "Speaker Utterances Tying Among Speaker Segmented Audio Documents Using Hierarchical Classification: Towards Speaker Indexing of Audio Databases", Proc of ICSLP 2002, Denver, Etats-Unis Septembre 2002

[Meignier 2004] Meignier S., Moraru D., Fredouille C., Besacier L., Bonastre J.-F., "Benefits of Prior Acoustic Segmetation for Automatic Speaker Segmentation", Proc of ICASSP 2004, Montreal, Canada, Mai 2004

[Montacie 1997] Montacie C., Caraty M.-J., "Sound Channel Video Indexing", Proc of Eurospeech 1997, pages 2359-2362, Rhodes, Grece, Septembre 1997

[Moon 1996] Moon T. K., "The Expectation-Maximization Algorithm", IEEE Signal Processing, vol. 13(6), pages 47-60, Novembre 1996

[Moraru 2003a] Moraru D., Besacier L., "Towards Conversational Model for Speaker Segmentation", Proc of Sped 2003, Bucarest, Roumanie, Mai 2003

[Moraru 2003b] Moraru D., Meignier S., Besacier L., Bonastre J.-F., Magrin-Chagnolleau I., "The ELISA Consortium Approaches In Speaker Segmentation During The NIST 2002 Speaker Recognition Evaluation", Proc of ICASSP 2003, Hong-Kong, China, Mai 2003

[Moraru 2003c] Moraru D., Besacier L., Meignier S., Fredouille C., Bonastre J.-F., "ELISA, CLIPS and LIA NIST 2003 Segmentation", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Moraru 2004a] Moraru D., Besacier L., Castelli E., "Using A Priori Information For Speaker Diarization", A Speaker Odyssey 2004, Toledo, Spain, Mai 2004

[Moraru 2004b] Moraru D., Meignier S., Fredouille C., Besacier L., Bonastre J.-F., "ELISA NIST RT 03 Broadcast News Speaker Diarization Experiments", A Speaker Odyssey 2004, Toledo, Spain, Mai 2004

[Moraru 2004c] Moraru D., Meignier S., Fredouille C., Besacier L., Bonastre J.-F., "The ELISA Consortium Approaches in Broadcast News Speaker Segmentation During The Nist 2003 Rich Transcription Evaluation", Proc of ICASSP 2004, Montreal, Canada, Mai 2004

[Moraru 2004d] Moraru D., Meignier S., Fredouille C., Besacier L., Bonastre J.-F., "Segmentation selon le locuteur: les activités du Consortium ELISA dans le cadre de Nist RT03", JEP 2004, Fès, Maroc, Avril 2004

[Navratil 2002] Navratil J., Ramaswamy G., Chaudhari U., Zilca R., "IBM 1-Sp Detection Systems", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[Nguyen 2003a] Nguyen P., Junqua J.-C., "PSTL's Speaker Diarization", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Nguyen 2003b] Nguyen P., "SWAMP: An Isometric Frontend for Speaker Clustering", Proc of NIST RT 2003 Workshop, Boston, Mai 2003, Submited to Eurospeech 2003

[NIST 2001] NIST, The NIST 2001 Speaker Recognition Evaluation http://www.nist.gov/speech/tests/spk/2001/doc/index.htm

[NIST 2002] NIST, The NIST 2002 Speaker Recognition Evaluation http://www.nist.gov/speech/tests/spk/2002/doc/index.htm

[NIST 2003] NIST, The NIST 2003 Rich Transcription Evaluation http://www.nist.gov/speech/tests/rt/rt2003/spring/

[Olsen 1995] Olsen J., "Separation of Speakers In Audio Data", Proc of Eurospeech 1995, pages 355-358, Madrid, Septembre 1995

[Parlangeau 2003] Parlangeau-Vallès N., Farinas J., Fohr D., Illina I., Magrin-Chagnolleau I., Mella O., Pinquier J., Rouas J.L., Sénac C., "Audio Indexing on the Web: A Preliminary Study of Some Audio Descriptors", 7th World Multiconference on Systematics, Cybernetics and Informatics 2003, Orlando, Juillet 2003

[Pinquier 2001] Pinquier J., Chambert N., "La première étape d'un système d'Indexation audio (Parole/Musique/Bruit)", Quatrièmes Rencontres Jeunes Chercheurs en Parole, Mons, Septembre 2001

[Pinquier 2003a] Pinquier J., Rouas J.L., André-Obrecht R., "Fusion de paramètres pour une classification automatique parole/musique robuste", Technique et science informatiques (TSI) : Fusion numérique/symbolique, Hermès, 8, quai du marche neuf, F-75004 Paris, V. 22, p. 831-852, 2003

[Pinquier 2003b] Pinquier J., André-Obrecht R., "Audio Indexing: Primary Components Retrieval", Proc of CBMI 2003, INRIA Rennes, Septembre 2003, p. 293-296, Vol. I.

[Quenot 2001] Quenot G., "TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation", 10th Text Retrieval Conference, Gaithersburg, MD, USA, 13-16 Novembre, 2001

[Quénot 2002] Quénot G. M., Moraru D., Besacier L., Mulhem P., "CLIPS-IMAG at TREC-11 : Experiments in Video Retrieval", 11th Text Retrieval Conference, Gaithersburg, MD, USA, Novembre 2002

[Rabiner 1989] Rabiner L. R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", IEEE Transactions on Speech Audio Processing, tome 77(2), pages 257-285, 1989

[Rabiner 1993] Rabiner L. R., Juang B.H., *"Fundamentals of Speech Recognition"*, Prentice Hall, 1993, ISBN 0-13-015157-2

[Rapp 1998] Rapp S., Dogil G., "Same News Is Good News: Automatically Collecting Reoccurring Radio News Stories", Proc of ICSLP 1998, pages 1587-1590, Sydney, Australia, 1998

[Reynolds 1992] Reynolds D., *"A Gaussian Mixture Modeling Approach to Text-Independent Speaker Verification"*, PhD Thesis, Georgia Institute of Technology, 1992

[Reynolds 1995] Reynolds D., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, tome 17(1-2), pages 91-108, 1995

[Reynolds 1996] Reynolds D., "The Effects of Handset Variability on speaker Recognition Performance: Experiments on The Switchboard Corpus", Proc of ICASSP 1996, Atlanta, Mai 1996

[Reynolds 1997] Reynolds D., "Comparation of Background Normalization Methods for Text Independent Speaker Verification", Proc of EuroSpeech 1997, Rhodos, Greece, Septembre 1997

[Reynolds 1998] Reynolds D., Singer E., Carlson B., O'Leary J., Zissman M., "Blind Clustering of Speech Utterances Based on Speaker and Language Characteristics" Proc of ICSLP 1998, Sydney, Decembre 1998

[Reynolds 2000] Reynolds D., Quatieri T., Dunn R., "Speaker Verification Using Adaptation Mixture Models", Digital Signal Processing, Vol. 10, No. 1-3, Janvier/Avril/Juillet 2000

[Reynolds 2002] Reynolds D., Campbell J., Dunn B., Jones D., Sturim D., Quatieri T., "MIT Lincoln Laboratory System: 1sp, 2sp and Segmentation", Proc of NIST SpRec 2002 Workshop, Vienna, VA, Mai 2002

[Reynolds 2003] Reynolds D., Torres P., Roy. R, "EARS RT03 Diarization", Proc of NIST RT 2003 Workshop, Boston, MA, Mai 2003

[Rissanen 1989] Rissanen J., "Stochastic Complexity in Statistical Inquiry", World Scientific, 1989, Series in Computer Science, Vol 15

[Rowley 1997] Rowley H., Baluja S., Kanade T. "Neural Network-Based Face Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 20, no. 1, pages 23-38, Janvier 1998

[Saunders 1996] Saunders J., "Real-Time Discrimination of Broadcast Speech/Music", Proc of ICASSP 1996, Atlanta, Mai 1996

[Scheirer 1997] Scheirer E., Slaney M., "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator", Proc of ICASSP 1997, Munich, Avril 1997

[Schwarz 1978] Schwarz G., "Estimating The Dimension of A Model", The Annals of Statistics Vol 6(2), pages 461-464, 1978

[Seck 2001] Seck M. *"Détection de ruptures et suivi de classes de sons pour l'indexation sonore"*, Thèse de doctorat, Université de Rennes 1, 169 pages, IRISA 2001

[Siegler 1997] Siegler M., Jain U., Raj B., Stern R., "Automatic Segmentation and Clustering of Broadcast News Audio", DARPA Speech Recognition Workshop, Westfields, Virginia, 1997

[Smeaton 2001] Smeaton A., "Indexing, Browsing and Searching of Digital Video and Digital Audio Information", Lecture Notes on Information Retrieval, pages 93-110, Springer-Verlag New York, 2001

[Smeaton 2003] Smeaton A., Kraaij W., Over P., "TRECVID 2003 – An introduction", 12th Text Retrieval Conference, Gaithersburg, MD, USA, 2003

[Solomonoff 1998] Solomonoff A., Mielke A., Schmidt M., Gish H., "Clustering Speakers by Their Voices", Proc of ICASSP 1998, Seattle, Mai 1998

[Sonmez 1999] Somnez K., Heck L., Weintraub N., "Speaker Tracking and Detection with Multiple Speakers", Proc of Eurospeech 1999, Budapest, Hungary, Septembre 1999

[Soong 1988] Soong F. K., Rosenberg A. E., "On The Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", IEEE Transactions on Acoustic, Speech and Signal Processing, tome 36(6), pages 871-879, 1988

[Soong 1992] Soong F. K., Rosenberg A. E., Rabiner L. R., Juang B. H., "A Vector Quantization Approach to Speaker Recognition", Proc of ICASSP 1992, pages 387-390, San Francisco, Mars 1992

[Standford 2003] Standford V., Garofolo J., Galibert O., Michel M., Laprun C., "The NIST Smart Space and Meeting Room Projects: Signal, Acquisition, Annotation and Metrics", Proc of ICASSP 2003, Hong-Kong, China, Mai 2003

[Thong 2000] Thong J.-M., Goddeau D., Litvinova A., Logan B., Moreno P., Swain M., "SpeechBot: A Speech Recognition Based Audio Indexing System for the Web", Conference on Content-Based Multimedia Information Access, RIAO 2000, 2000

[Thyes 2000] Thyes O., Kuhn R., Nguyen P., Junqua J.-C., "Speaker Identification and Verification Using Eigenvoices", Proc of ICSLP 2000, Beijing, China, Octobre 2000

[Tranter 2003a] Tranter S., Yu K., "Diarization for RT-03s at Cambridge University", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Tranter 2003b] Tranter S., Yu K., Reynolds D., Kim D.Y., Evermann G., Woodland P., "Interactions Between Diarization and STT", Proc of NIST RT 2003 Workshop, Boston, Mai 2003

[Tranter 2004] Tranter S., Reynolds D., "Speaker Diarization for Broadcast News", A Speaker Odyssey 2004, Toledo, Spain, Mai 2004

[Tritschler 99] Tritschler A., Gopinath R., "Improved Speaker Segmentation and Segments Clustering Using The Bayesian Information Criterion", Proc of Eurospeech 1999, Budapest, Hungary, Septembre 1999

[Viterbi 1967] Viterbi A., "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm", Proc of IEEE Transactions on Information Theory, vol. 13, pages 260-269, Avril 1967

[Wilcox 1994a] Wilcox L., Kimber D., Chen F., "Audio Indexing Using Speaker Identification" Proc of SPIE Conference on Automatic Systems for the Inspection and Identification of Humans, pp 149-157, San Diego, CA, Juillet 1994

[Wilcox 1994b] Wilcox L., Chen F., Kimber D., Balasubramaian V., "Segmentation of Speech Using Speaker Identification", Proc of ICASSP 1994, Adelaide, Australia, Avril 1994

[Wu 2003] Wu L., Guo Y., Qiu X., Feng Z., Rong J., Jin W., Zhou D., Wang R., Jin M., "Fundan University at TRECVID2003", Proc of TREC 2003 Workshop, Gaithersburg, MD, USA, Novembre 2003

[Zhang 1993] Zhang H.-J., Kankanhalli A., Smoliar S.W., "Automatic Partitioning of Full-Motion Video", Multimedia Systems, 1(1), pp 10-28, juin 1993